processes

Editorial

Special Issue on "Big Data in Biology, Life Sciences and Healthcare"

Q. Peter He * and Jin Wang *

Department of Chemical Engineering, Auburn University, Auburn, AL 36849, USA
* Correspondence: qhe@auburn.edu (Q.P.H.); wang@auburn.edu (J.W.);
Tel.: +1-334-844-7602 (Q.P.H.); +1-334-844-2020 (J.W.)

In the past few decades, we have witnessed tremendous advancements in biology, life sciences and healthcare. These advancements are due, in no small part, to computer-enabled analyses (i.e., machine learning) of Big Data made available by various high-throughput technologies (e.g., various omics technologies), the digitalization of health/medical records, wearable and portable Internet-of-Things (IoT) sensors and digitized medical images. These advancements would also not have been possible without the ever-increasing computing power and rapid algorithmic advances in machine learning, especially in deep learning, which make the analysis of Big Data possible. In this Special Issue on "Big Data in Biology, Life Sciences and Healthcare," we have included contributions that demonstrate the application of Big Data modeling and analysis to support scientific research, clinical decision making, the assessment of existing healthcare systems and the proposal of new ones, among others. The Special Issue is available online at the following website: https://www.mdpi.com/journal/processes/special_issues/big_data_biology (accessed on 13 December 2021).

Big Data in Healthcare Systems

Healthcare systems in different countries are dealing with similar challenges, such as an aging population and the increasing impact of chronic diseases. Addressing these challenges requires an assessment of the existing systems and recommendations for improvements or the proposal of new systems. For example, Carnicero et al. [1] identified the main challenges of the National Health Service of Spain and proposed its transformation into a Learning Health System (LHS), which consists of Big Data tools and machine learning systems. The authors first performed a comprehensive comparison of Spain's health and healthcare-related indicators with those of European Union averages. They then conducted longitudinal studies of the healthcare system in Spain, which included yearly health expenditure as %GDP, waiting times for elective surgery, etc. Based on these analyses, the authors identify the limitations of the individual clinical information in electronic health records (EHRs). For example, EHRs do not contain information about over-the-counter drugs, lifestyle habits (e.g., sleep, diet or physical activity), environmental risk factors and information from the perspective of the patient or his/her priorities, which are important for clinical decision making. The authors further emphasize the importance of including resources generated by patients on social media, apps, sensors, wearables, direct services to consumers, polls and questionnaires in the proposed LHS. The authors detail the required quality control and transformation process (through standardization and normalization) in order to incorporate data from different sources into LHS, which must satisfy format compatibility, be suitable to needs, have completeness and have integrity and consistency, as well as precision and accuracy, etc. The authors provide a detailed holistic technical architecture for guiding a preliminary implementation of LHS. The architecture is composed of the following three main layers: (1) data gathering and data homogenization; (2) Big Data and machine learning; and (3) presentation and data analysis. The authors

check for updates

Citation: He, Q.P.; Wang, J. Special Issue on "Big Data in Biology, Life Sciences and Healthcare". *Processes* **2022**, *10*, 41. https://doi.org/10.3390/pr10010041

Received: 16 December 2021
Accepted: 22 December 2021
Published: 27 December 2021

also outline the main stages for implementing the system and for addressing potential challenges, which include twelve steps with detailed descriptions. The bioethical and legal aspects of LHS are also lightly touched upon. The authors also provided a list of twelve elements that are considered critical for the success of the project. The simple, flexible and holistic three-step architecture, the detailed twelve implementation steps and the twelve critical components for success are truly valuable in providing thoughtful recommendations for other researchers and policymakers who are interested in improving or transforming existing healthcare systems.

In another study, Yoon et al. [2] examined the current status and utilization of 22 health promotion projects that use the healthcare information system in Korea. The health promotion projects include 13 community-integrated health promotion projects, including smoking prohibition; sobriety; physical activity; nutrition; the preventive management of obesity; oral health; the preventive management of cardiovascular and cerebrovascular diseases; oriental medicine health promotion; the preventive management of atopy and asthma; maternal and child health; dementia management; community-based rehabilitation; and visiting health care. The nine other health promotion projects include mental health, elderly, health examination results counseling, tuberculosis, Hansen's disease, sexually transmitted infection, maternal and child health and vaccination. The comprehensive health promotion projects studied in this paper can be leveraged to other regions or countries. The findings suggest linked applications between projects are more effective in improving local residents' health when compared to isolated applications of individual projects. Based on the research results, the authors suggest the following three policy considerations for future health promotion projects: (1) identify and categorize health promotion projects that need to be linked and strengthen collaboration between project entities; (2) define and adjust health promotion priorities according to local characteristics; and (3) unify project entities to enhance the effectiveness of linkages. The findings from this research add knowledge to the field of health promotion and evidence-based recommendations made by the authors can be helpful for the implementation of linked or collaborative health promotion projects elsewhere.

The third study of healthcare systems focuses on reducing waiting time in hospitals by employing advanced nurses (ANs) in emergency services (ESs) in Turkey [3]. The study was conducted via a discrete-event simulation model with and without ANs. It was found that employing ANs can significantly increase the number of treated patients and reduce the waiting time and length of stay.

The fourth study, which is by Lee and Yoon, aims to examine the determinants of catastrophic health expenditure in households with cancer patients [4]. Transition probability analysis was conducted to determine the occurrence probability for the following year when no catastrophic health expenditure occurred in the current year. This study offers key insights into factors determining the occurrence of catastrophic health expenditure in households with cancer patients. The authors also provide specific policy suggestions that aim at lowering the risk of catastrophic health expenditure in cancer households, especially for poverty and near-poverty groups.

There are two studies that focus on addressing two of the most common technical challenges of analyzing healthcare Big Data: the high dimensionality of healthcare data collected from a variety of sources and the missing value problems that are common in healthcare data. Bodur and Atsa'am [5] focus on addressing the high dimensionality of healthcare data by proposing a filter type feature selection algorithm based on risk ratios (RR). The algorithm is advantageous in several aspects, including its superior performance when compared to six existing feature selection methods, and its independence of a machine learning tool, which allows any modeling or classification strategy to be implemented on the reduced feature space. Sun et al. [6] focused on addressing missing value problems that are common in healthcare data. A novel discriminative restricted Boltzmann machine (DRBM) model was proposed to first extract the latent features in the dataset with various degree of missing values and then to predict missing values via a latent feature

model. The performance of the proposed method is demonstrated using publicly available clinical datasets.

Big Data from IoT Sensors, Wearables and Biosignals

This Special Issue includes several contributions that focus on the application of novel wearable or portable sensors and biosignals for improved care. Huh [7] proposes a solitary senior citizen care system based on Internet-of-Things (IoT) sensors and power line communication (PLC) technologies to complement more commonly utilized sensors such as environment/object sensors (e.g., on-off switches, pressure sensors and infrared door sensors), motion sensors (e.g., various wearable sensors) and camera sensors. The proposed system detects an individual appliance's power-use pattern and compares it with historical patterns to raise alarms of potential health, mental or emotional issues with the resident. The system can also incorporate other sensors that detect gas and water usage or the intensity of illumination in the house. The advantage of the proposed system over other monitoring systems is that it is particularly suitable for poor communities where there is no internet service but only electricity and water supplies. This study also provides quite a comprehensive overview of the field by summarizing existing solutions. In another study, Lee et al. [8] examined the feasibility of utilizing biosignals to detect anxiety caused by driving situations. The study employs multimodal biosignals from wearable sensors, including electroencephalography (EEG), photoplethysmography (PPG), electrodermal activity (EDA) and pupil size to estimate anxiety under various driving situations. The results indicate that an EEG is a better biosignal than others when used in logistic regression (LR) to detect anxiety. Adding other biosignals such as EDA or pupil size to an EEG further enhances the detection performance in some participants.

Big Data in Omics

The Special Issue includes two review papers in the area of bioinformatics and machine learning tools for proteomics, which has emerged as an important dimension of omics. Paul et al. [9] reviewed bioinformatics tools developed for phosphoproteomics. The study is inspired by the profound impact phosphorylation has on biological functions from intrinsic activity and extrinsic executions to cellular localization and the strength of phosphoproteomics research in that it provides an overall picture of the workforce of the cell. This review consolidates existing bioinformatics tools developed for phosphoproteomics and highlights the gap between the development of bioinformatics tools and their implementation in clinical research. In another study, Oh et al. [10] review bioinformatics tools that have been developed or applied for diabetic nephropathy (DN), which has become a rising concern amongst diabetics and diabetologists. The review summarizes bioinformatic tools applied for diabetes mellitus with a focus on the proteomic advances made in DN. The findings from this review indicate that there is an urgent need for bioinformatics tools to identify new biomarkers and new drug targets for DN. In addition to these two review papers, Li et al. [11] studied the antimicrobial resistance of clinical pathogens using historical samples from six countries. The study is conducted in response to the spread of antimicrobial resistance pathogens in humans that is increasingly threatening public health. The similarities among different countries in terms of genes and pathogens are investigated in order to understand the potential avenues for antimicrobial-resistance gene spreading. The novel integration of principal component analysis (PCA) and hierarchical clustering enables the identification of both the important genes and important pathogens in the data. The genes and pathogens that are closely involved in antimicrobial resistance are further studied temporally to evaluate the trend of antimicrobial resistance. The increasing occurrence of antimicrobial resistance is correlated to the global increase in antibiotic consumption in recent years. The increase in antimicrobial resistance genes could also be linked to increased trade and travel between countries. The trend is contributing to the worsening antimicrobial resistance epidemic that is posing a critical health hazard on a global scale.

Big Data in Biomedical Images

The significant advancement of machine learning techniques, especially deep learning techniques such as convolutional neural networks (CNNs), has enabled automated analysis of biomedical images for classification, object detection, segmentation and registration. In this Special Issue, Buiu et al. [12] proposed an automated colposcopy image analysis framework for the classification of precancerous and cancerous lesions of the uterine cervix using an ensemble of MobileNetV2 networks. MobileNetV2 is a family of general-purpose computer vision neural networks developed by Google Research and designed with mobile devices in mind to support mobile visual recognition including classification, object detection and semantic segmentation. The ensemble model takes advantage of different types of images available in a colposcopy procedure. The high classification accuracies achieved using the proposed method, together with its fast execution speed and low memory requirement, demonstrate its potential use as a valuable real-time tool for assisting doctors in a colposcopy.

Technical and Ethical Considerations in Biological Big Data Analytics

This Special Issue also includes articles that provide in-depth discussions on the technical and ethical considerations in biological Big Data analytics. This is due to the fact that the findings from these studies could have significant implications when they influence clinical decisions and/or policy determinations. He and Wang [13] focus on the application of systems engineering principles and techniques in addressing some of the technical challenges in Big Data analytics for biological, biomedical and healthcare applications, including the principle of parsimony in addressing overfitting, the dynamic analysis of biological data and the role of domain knowledge in biological data analytics. Specifically, this work addresses overfitting and model validation, two of the most important issues in developing data-driven machine learning models, especially those of high flexibility such as nonlinear-nets, kernel-nets, or neural-nets-based models. This study provides detailed discussions on various techniques for overfit checking and a full spectrum of strategies for reducing/avoiding the risk of overfitting. In addition, the systems engineering principles advocate the integration of domain knowledge with data-driven machine learning in analyzing biological Big Data. Based on systems theory and the degree of freedom analysis, in particular, the authors promote knowledge-matching-based model validation over the conventional point matching method. This consideration is based on the fact that the extreme complexity of biological systems, in conjunction with often severely incomplete knowledge, results in models with very high degrees of freedom. A good agreement between model predictions and experimental measurements at certain points is not sufficient to guarantee the predictive capability of a model. This is because it is not difficult to change a few parameters and/or constraints among hundreds or even thousands to match the normally small number of experimental measurements for post hoc explanations. As a result, point-matching alone is not a reliable approach for validating a model with a high degree of freedom. This paper also reviews advances made in knowledge-guided learning, including unsupervised and supervised learning, as well as knowledge-guided feature engineering and selection. Knowledge-guided feature engineering is particularly important for clinical data modeling. For these applications, it has been recognized that without using features that are rooted in (patho)physiology for clinical prediction, there is a risk of capturing a factor that may change at any time, invalidating a model that appears to produce computationally reproducible results. He and Wang [13] also stress the importance of dynamic analysis of biological data by reviewing progress made toward the dynamic analysis of gene regulatory networks, metabolic networks and signal transduction networks, as well as the integrated dynamic analysis of these networks simultaneously. As cellular systems are networks of interacting components that change with time in response to external and internal events, analyzing the steady-state behavior alone is insufficient for characterizing responses. The authors envision that studying the dynamic behavior of these networks will become the basis for understanding cellular functions and disease mechanisms.

Ethical considerations are equally important in biological Big Data analytics, again, due to the potentially significant impacts of analysis results on clinical decisions and/or policy determinations. Leon-Sanz [14] provides insightful considerations for the ethical analysis of healthcare Big Data. Their study presents specific ethical issues arising from the use of Big Data in life sciences and healthcare. For example, the author notes that the ethical requirements of Big Data should include the technical precision of data analysis, accuracy and statistical performance. Otherwise, the information obtained may be subject to biases and errors, which could result in negative societal consequences. In addition, care should be taken not to extrapolate the results beyond the scope/extrapolability of the model. These points are often not stressed by many researchers who analyze biological Big Data. The "usual" sense of ethics (e.g., data confidentiality and individual privacy) is only one of many ethical aspects discussed in this study. Therefore, this paper can significantly broaden our view on the key points of ethics when analyzing biological Big Data. Another important point is transparency, which should refer to both the origin of information and the operations carried out on the data because of the complex algorithms applied in most studies. The author also notes the importance of protecting the so-called "group identity" within the scope of the Big Data, which is important to avoid the discrimination of individuals or groups of people. The author identifies some particularly sensitive areas, which include disability, mental illnesses, genetic diseases, sexual orientation, drug addictions, juvenile delinquency and political or religious issues. The author also provides in-depth discussions on emerging issues and provides potential solutions to the challenges in the field. The paper concludes that good practices in the management of Big Data related to life sciences and healthcare depend on respect for the rights of individuals, the improvement of these practices that can introduce assistance to individual patients, the promotion of society's health in general and the advancement of scientific knowledge. It is recommended by the author that spaces for ethical reflection should be built into research to address potential ethical tension and conflicts of interest, as well as to resolve possible contradictions.

Author Contributions: Writing—original draft preparation, Q.P.H.; writing—review and editing, J.W. All authors have read and agreed to the published version of the manuscript.

Funding: National Science Foundation (NSF-CBET #1805950), and U.S. Department of Energy (DE-SC0019181).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Carnicero, R.; Rojas, D.; Elicegui, I.; Carnicero, J. Proposal of a Learning Health System to Transform the National Health System of Spain. *Processes* **2019**, *7*, 613. [CrossRef]
2. Yoon, K.; Park, S.; Choi, S.; Lee, M. A Proposal for Public Health Information System-Based Health Promotion Services. *Processes* **2020**, *8*, 338. [CrossRef]
3. Atalan, A.; Donmez, C.C. Employment of Emergency Advanced Nurses of Turkey: A Discrete-Event Simulation Application. *Processes* **2019**, *7*, 48. [CrossRef]
4. Lee, M.; Yoon, K. Catastrophic Health Expenditures and Its Inequality in Households with Cancer Patients: A Panel Study. *Processes* **2019**, *7*, 39. [CrossRef]
5. Bodur, E.K.; Atsa'Am, D.D. Filter Variable Selection Algorithm Using Risk Ratios for Dimensionality Reduction of Healthcare Data for Classification. *Processes* **2019**, *7*, 222. [CrossRef]
6. Sun, M.; Min, T.; Zang, T.; Wang, Y. CDL4CDRP: A Collaborative Deep Learning Approach for Clinical Decision and Risk Prediction. *Processes* **2019**, *7*, 265. [CrossRef]
7. Huh, J.-H. An Efficient Solitary Senior Citizens Care Algorithm and Application: Considering Emotional Care for Big Data Collection. *Processes* **2018**, *6*, 244. [CrossRef]

8. Lee, S.; Lee, T.; Yang, T.; Yoon, C.; Kim, S.-P. Detection of Drivers' Anxiety Invoked by Driving Situations Using Multimodal Biosignals. *Processes* **2020**, *8*, 155. [CrossRef]
9. Paul, P.; Muthu, M.; Chilukuri, Y.; Haga, S.W.; Chun, S.; Oh, J.-W. In Silico Tools and Phosphoproteomic Software Exclusives. *Processes* **2019**, *7*, 869. [CrossRef]
10. Oh, J.-W.; Muthu, M.; Haga, S.W.; Anthonydhason, V.; Paul, P.; Chun, S. Reckoning the Dearth of Bioinformatics in the Arena of Diabetic Nephropathy (DN)—Need to Improvise. *Processes* **2020**, *8*, 808. [CrossRef]
11. Li, K.; Zheng, J.; Deng, T.; Peng, J.; Daniel, D.; Jia, Q.; Huang, Z. An Analysis of Antimicrobial Resistance of Clinical Pathogens from Historical Samples for Six Countries. *Processes* **2019**, *7*, 964. [CrossRef]
12. Buiu, C.; Dănăilă, V.-R.; Răduţă, C.N. MobileNetV2 Ensemble for Cervical Precancerous Lesions Classification. *Processes* **2020**, *8*, 595. [CrossRef]
13. He, Q.P.; Wang, J. Application of Systems Engineering Principles and Techniques in Biological Big Data Analytics: A Review. *Processes* **2020**, *8*, 951. [CrossRef]
14. Leon-Sanz, P. Key Points for an Ethical Evaluation of Healthcare Big Data. *Processes* **2019**, *7*, 493. [CrossRef]