applied sciences

MDPI

Editorial

Special Issue on Big Data for eHealth Applications

Stefano Silvestri * and Francesco Gargiulo *

Institute for High Performance Computing and Networking, National Research Council of Italy, Via Pietro Castellino, 111, 80131 Naples, Italy

* Correspondence: stefano.silvestri@icar.cnr.it (S.S.); francesco.gargiulo@icar.cnr.it (F.G.)

1. Introduction

In the last few years, the rapid growth in available digitised medical data has opened new challenges for the scientific research community in the healthcare informatics field. In this scenario, the constantly increasing volume of medical data, as well as the complexity and heterogeneity of this kind of data require innovative approaches based on Big Data Analytics (BDA) and Artificial Intelligence (AI) methods for extracting valuable insights [1–5], and at the same time, these new approaches must also guarantee the required levels of privacy and security [6]. These solutions must also provide effective and efficient tools for supporting the daily routine of physicians, medical professionals, and policy makers, improving the quality of healthcare systems. Finally, they should leverage the huge amount of information buried under these Big Data [7], exploiting, in this way, their full potential.

Furthermore, new heterogeneous and extensive COVID-related datasets have been collected during the recent pandemic and have often been made available to the scientific community. In this case, the need for new and specific Big Data approaches for processing such data makes exploiting these data and providing new and innovative approaches for facing the COVID-19 pandemic more urgent [8–10].

In this Special Issue, some innovative applications, tools, and techniques specifically tailored to address issues related to the eHealth domain by leveraging BDA methodologies are presented. Moreover, these techniques are also presented in this Special Issue, given the definition of complex systems and architectures for the eHealth domain fundamentally based on the combination of Internet of Things (IoT) devices and Artificial Intelligence (AI) methods. Finally, the Cyber Security (CS) for eHealth topic is also addressed given the significant increase in cyber threats in the healthcare sector during the last few years.

2. Big Data for eHealth Applications

In light of the above, this Special Issue was introduced to collect the latest research on relevant topics and, more importantly, to address present challenges with using Big Data for eHealth applications. Moreover, it considered AI and/or IoT-based technologies, the combined use of which can lead to the definition and implementation of effective and innovative solutions [11]. Finally, CS techniques for the eHealth domain were also taken into account.

There are 10 contributions selected for this Special Issue, representing innovative applications in the areas mentioned above from original contributions of researchers with broad expertise in various and multidisciplinary fields, considering the medical, informatics, and engineering fields. The Special Issue includes the following papers:

- Iterative Annotation of Biomedical NER Corpora with Deep Neural Networks and Knowledge Bases [12]
- Cyberattack Path Generation and Prioritisation for Securing Healthcare Systems [13]
- The Assessment of COVID-19 Vulnerability Risk for Crisis Management [14]
- Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study [15]
- Design of a Wearable Healthcare Emergency Detection Device for Elder Persons [16]

check for updates

Citation: Silvestri, S.; Gargiulo, F. Special Issue on Big Data for eHealth Applications. *Appl. Sci.* **2022**, *12*, 7578. https://doi.org/10.3390/app12157578

Received: 22 July 2022
Accepted: 26 July 2022
Published: 28 July 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

- Reducing the Heart Failure Burden in Romania by Predicting Congestive Heart Failure Using Artificial Intelligence: Proof of Concept [17]
- Nonlinear Random Forest Classification, a Copula-Based Approach [18]
- A Novel Unsupervised Computational Method for Ventricular and Supraventricular Origin Beats Classification [19]
- A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications [20]
- On Combining Feature Selection and Over-Sampling Techniques for Breast Cancer Prediction [21]

The aforementioned papers refer to the following main topics within the healthcare scenario: (i) COVID-19 datasets and models [14,15], (ii) large dataset annotation [12], (iii) cyber security [13], (iv) federated learning [20], (v) smart biomedical systems and devices [16], and (vi) artificial intelligence approaches [17–19,21].

More in detail, in [12], a methodology for reducing the manual effort needed to annotate a biomedical-named entity recognition (B-NER) corpus was presented, exploiting both active learning and distant supervision, respectively, based on Deep Learning models (e.g., Bi-LSTM, word2vec FastText, ELMo, and BERT) and biomedical knowledge bases to speed up the annotation task. The proposed approach is also able to limit class imbalance issues. The results showed that this method allows us to annotate an effective and large B-NER corpus with a fraction of the time required by a fully manual annotation, addressing the lack of annotated corpora in the biomedical domain [22]. The authors also analysed the most effective embedding model to represent the input words [23] and the applicability of this approach to other domains.

In [13], a novel methodology for the cyberattack path discovery to ensure security within the healthcare ecosystem is presented. This approach is based on the Common Vulnerability Scoring System (CVSS), so that base metrics and exploitability features can be used to determine and prioritise the possible attack paths based on the threat actor capability, asset dependency, and target user profile and evidence of indicator of compromise. The work includes a real example from the healthcare use case to demonstrate the methodology used for attack path generation. The result from the studied context, which processes Big Data from healthcare applications, shows that the uses of various parameters such as CVSS metrics, threat actor profile, and the indicator of compromise are able to generate realistic attack paths. In this way, healthcare practitioners can be supported in identifying the controls that are required to secure the overall healthcare ecosystem.

The authors of [14] presented a methodology that is used determine COVID-19 vulnerability risk and its change over time in association with the state health care system, turnover, and transport to support the crisis management decision-making process. In detail, this method aims to determine the COVID-19 Vulnerability Index (CVI) based on the selected criteria. The risk assessment was carried out with methodology that includes the application of a multi-criteria analysis and spatio-temporal aspects of available data. Particularly, the Spatial Multicriteria Analysis (SMCA) compliant with the Analytical Hierarchy Process (AHP), which incorporated selected population and environmental criteria were used to analyse the ongoing pandemic. The influence of combining several factors in an analysis of the pandemic was illustrated, and the static and dynamic factors to COVID-19 vulnerability risk were determined to prevent and control the spread of COVID-19 at the early stages of the pandemic. As a result, areas with a certain level of risk in different periods of time were determined. Furthermore, the number of people exposed to a COVID-19 vulnerability risk was presented with time. The results obtained proved that the this approach can support the decision-making process by showing the area where preventive actions should be considered.

In [15], a new pre-trained neural language model based on the BERT model [24] was introduced. This model was named CovBERT, and it was specifically designed to improve the overall review task performances on the COVID-19 literature with respect to the classic BERT model. CovBERT was pretrained on a very large corpus formed by

scientific publications in the biomedical domain related to COVID-19. The CovBERT was tested on the classification task of short texts of biomedical articles. The obtained results demonstrated significant improvements. In addition, the authors also made a COVID-19 corpus available, entitled *COV-Dat-20*.

The authors of [16] proposed a wearable system that takes advantage of sensors embedded in a smart device to collect data for movement identification (running, walking, falling, and daily activities) of an older adult user in real-time. To provide high efficiency in fall detection, the sensor readings were analysed using a neural network. If a fall is detected, an alert is sent though a smartphone connected via Bluetooth. The proposed system was tested in both inside and outside environments, and the results of the experiments showed that it is extremely portable and is able to provide high success rates in fall detection in terms of accuracy and loss.

In [17], a noncontact system that can predict heart failure exacerbation through vocal analysis was studied and implemented. The system was designed to evaluate the voice characteristics of every patient, used to identify variations using a Machine Learning-based approach. The authors collected voice data from real hospitalised patients since their admission to a hospital, when their general status was critical, until the day of discharge, when they were clinically stable. Each patient was classified adopting the New York Heart Association Functional Classification (NYHA) classification system for heart failure in order to include them in different stages based on their clinical evolution. Different ML algorithms were tested, namely Artificial Neural Networks (ANN), Support Vector Machine (SVM), and K-Nearest Neighbours (KNN), trained on voice data. The experiments demonstrated that the KNN obtained the best results and was able to correctly classify the NYHA stages of the patients exploiting only their voice recording, with an accuracy of 0.945.

In [18], a study on the copula-based approach to selecting the most important features for a Random Forest classification was used to classify a label-valued outcome. The methodology was simulated on a real dataset of COVID-19 and diabetes. In detail, based on associated copulas between these features, the authors carried out this feature selection and then embedded the selected features into a Random Forest algorithm to classify a label-valued outcome. This algorithm allowed us to select the most relevant features when the features are not necessarily connected by a linear function, and it can stop the classification when the desired level of accuracy is reached. The experimental assessment successfully applied the proposed method on a simulation study as well as a real dataset of COVID-19 and for a diabetes dataset.

The study presented in [19] focused on a new unsupervised algorithm that adapts to every patient using the heart rate and morphological features of the ECG beats to classify beats between supraventricular origin and ventricular origin in order to predict arrhythmia. The results of the experiments performed obtained F-scores equal to 0.88, 0.89, and 0.93 for the ventricular origin beats for three popular ECG databases and around 0.99 for the supraventricular origin for the same databases, comparable with supervised approaches presented in other works, opening a new path to making use of ECG data to classify heartbeats without the assistance of a physician.

The work presented in [20] is a review paper, where a comprehensive and up-to-date review of research employing Federated Learning in healthcare applications was provided. Moreover, the paper highlighted a set of recent challenges from a data-centric perspective in Federated Learning, such as data partitioning characteristics, data distributions, data protection mechanisms, and benchmark datasets, was evaluated. Finally, several potential challenges and future research directions in healthcare applications were pointed out.

In [21], the imbalanced class problem was addressed, in particular for breast cancer prediction datasets. The authors presented a methodology that used a combination of the Information Gain (IG) and Genetic Algorithm (GA) feature selection methods and the Synthetic Minority Over-sampling TEchnique (SMOTE) to overcome this issue. The experimental results based on two breast cancer datasets showed that the combination of feature

selection and over-sampling outperformed the single usage of either feature selection and over-sampling for the highly class imbalanced datasets. In particular, performing IG first and SMOTE second is the better choice. For other datasets with a small class imbalance ratio and a smaller number of features, performing SMOTE is enough to construct an effective prediction model.

3. Future in Big Data for eHealth Applications

Although this Special Issue is now closed, more in-depth studies in Big Data Analytics applications developed explicitly for eHealth are expected. The outcomes of the research published in this Special Issue provided some new solutions in this area but also highlighted some of the still open issues that must be addressed to fully exploit Big Data in the healthcare domain in the future.

In detail, the presented papers underlined the need for extensive collections of biomedical annotated data, allowing for the training of high-performance ML and DL models to support physicians in their daily work. ML and AI approaches will support the daily routine of physicians and medical practitioners, but their extensive use will also raise privacy and security issues. It is also clear that the integration among IoT devices and sensors, AI and ML models, and Big Data approaches will be more pervasive for developing eHealth complex systems in the future. Adopting specifically pretrained neural language models will enable the researchers to define more intelligent systems for analysing large natural language clinical documents, fully exploiting their informative content. Finally, the large and heterogeneous data analyses related to COVID-19 can provide innovative pathways, more profound knowledge, and innovative approaches to address the risks of the current pandemic.

Author Contributions: Conceptualization: S.S. and F.G.; writing: S.S. and F.G.; reviewing and editing: S.S. and F.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This Special Issue would not have been possible without the contributions of the authors, the reviewers, and the dedicated editorial team of *Applied Sciences*. We congratulate all authors on their research. Moreover, we take this opportunity to express our sincere gratefulness to all reviewers. Finally, we express our gratitude to the editorial team of *Applied Sciences* and give a special thanks to Assistant Editor, for her continuous support.

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

References

1. Lv, Z.; Qiao, L. Analysis of healthcare big data. *Future Gener. Comput. Syst.* **2020**, *109*, 103–110. [CrossRef]
2. Cozzoli, N.; Salvatore, F.P.; Faccilongo, N.; Milone, M. How can big data analytics be used for healthcare organization management? Literary framework and future research from a systematic review. *BMC Health Serv. Res.* **2022**, *22*, 1–14. [CrossRef]
3. Karatas, M.; Eriskin, L.; Deveci, M.; Pamucar, D.; Garg, H. Big Data for Healthcare Industry 4.0: Applications, challenges and future perspectives. *Expert Syst. Appl.* **2022**, *200*, 116912. [CrossRef]
4. Luchini, C.; Pea, A.; Scarpa, A. Artificial intelligence in oncology: Current applications and future perspectives. *Br. J. Cancer* **2022**, *126*, 4–9. [CrossRef] [PubMed]
5. Busnatu, S.; Niculescu, A.G.; Bolocan, A.; Petrescu, G.E.D.; Păduraru, D.N.; Năstasă, I.; Lupușoru, M.; Geantă, M.; Andronic, O.; Grumezescu, A.M.; et al. Clinical Applications of Artificial Intelligence—An Updated Overview. *J. Clin. Med.* **2022**, *11*, 2265. [CrossRef] [PubMed]
6. Ciampi, M.; Sicuranza, M.; Silvestri, S. A Privacy-Preserving and Standard-Based Architecture for Secondary Use of Clinical Data. *Information* **2022**, *13*, 87. [CrossRef]
7. Silvestri, S.; Esposito, A.; Gargiulo, F.; Sicuranza, M.; Ciampi, M.; De Pietro, G. A Big Data Architecture for the Extraction and Analysis of EHR Data. In Proceedings of the 2019 IEEE World Congress on Services (SERVICES), Milan, Italy, 8–13 July 2019; IEEE: Piscatway, NJ, USA, 2019; Volume 2642-939X, pp. 283–288. [CrossRef]
8. Alsunaidi, S.J.; Almuhaideb, A.M.; Ibrahim, N.M.; Shaikh, F.S.; Alqudaihi, K.S.; Alhaidari, F.A.; Khan, I.U.; Aslam, N.; Alshahrani, M.S. Applications of Big Data Analytics to Control COVID-19 Pandemic. *Sensors* **2021**, *21*, 2282. [CrossRef] [PubMed]
9. Lin, L.; Hou, Z. Combat COVID-19 with artificial intelligence and big data. *J. Travel Med.* **2020**, *27*, taaa080. [CrossRef] [PubMed]

10. Catelli, R.; Gargiulo, F.; Casola, V.; De Pietro, G.; Fujita, H.; Esposito, M. Crosslingual named entity recognition for clinical de-identification applied to a COVID-19 Italian data set. *Appl. Soft Comput.* **2020**, *97*, 106779. [CrossRef] [PubMed]
11. Ciampi, M.; Coronato, A.; Naeem, M.; Silvestri, S. An intelligent environment for preventing medication errors in home treatment. *Expert Syst. Appl.* **2022**, *193*, 116434. [CrossRef]
12. Silvestri, S.; Gargiulo, F.; Ciampi, M. Iterative Annotation of Biomedical NER Corpora with Deep Neural Networks and Knowledge Bases. *Appl. Sci.* **2022**, *12*, 5775. [CrossRef]
13. Islam, S.; Papastergiou, S.; Kalogeraki, E.M.; Kioskli, K. Cyberattack Path Generation and Prioritisation for Securing Healthcare Systems. *Appl. Sci.* **2022**, *12*, 4443. [CrossRef]
14. Wyszyński, M.; Grudziński, M.; Pokonieczny, K.; Kaszubowski, M. The Assessment of COVID-19 Vulnerability Risk for Crisis Management. *Appl. Sci.* **2022**, *12*, 4090. [CrossRef]
15. Khadhraoui, M.; Bellaaj, H.; Ammar, M.B.; Hamam, H.; Jmaiel, M. Survey of BERT-Base Models for Scientific Text Classification: COVID-19 Case Study. *Appl. Sci.* **2022**, *12*, 2891. [CrossRef]
16. Amato, F.; Balzano, W.; Cozzolino, G. Design of a Wearable Healthcare Emergency Detection Device for Elder Persons. *Appl. Sci.* **2022**, *12*, 2345. [CrossRef]
17. Pană, M.A.; Busnatu, S.S.; Serbanoiu, L.I.; Vasilescu, E.; Popescu, N.; Andrei, C.; Sinescu, C.J. Reducing the Heart Failure Burden in Romania by Predicting Congestive Heart Failure Using Artificial Intelligence: Proof of Concept. *Appl. Sci.* **2021**, *11*, 11728. [CrossRef]
18. Mesiar, R.; Sheikhi, A. Nonlinear Random Forest Classification, a Copula-Based Approach. *Appl. Sci.* **2021**, *11*, 7140. [CrossRef]
19. Casas, M.M.; Avitia, R.L.; Cardenas-Haro, J.A.; Kalita, J.; Torres-Reyes, F.J.; Reyna, M.A.; Bravo-Zanoguera, M.E. A Novel Unsupervised Computational Method for Ventricular and Supraventricular Origin Beats Classification. *Appl. Sci.* **2021**, *11*, 6711. [CrossRef]
20. Prayitno; Shyu, C.R.; Putra, K.T.; Chen, H.C.; Tsai, Y.Y.; Hossain, K.S.M.T.; Jiang, W.; Shae, Z.Y. A Systematic Review of Federated Learning in the Healthcare Area: From the Perspective of Data Properties and Applications. *Appl. Sci.* **2021**, *11*, 11191. [CrossRef]
21. Huang, M.W.; Chiu, C.H.; Tsai, C.F.; Lin, W.C. On Combining Feature Selection and Over-Sampling Techniques for Breast Cancer Prediction. *Appl. Sci.* **2021**, *11*, 6574. [CrossRef]
22. Silvestri, S.; Gargiulo, F.; Ciampi, M.; De Pietro, G. Exploit Multilingual Language Model at Scale for ICD-10 Clinical Text Classification. In Proceedings of the 2020 IEEE Symposium on Computers and Communications (ISCC), Rennes, France, 7–10 July 2020; IEEE: Piscatway, NJ, USA, 2020; pp. 1–7. [CrossRef]
23. Silvestri, S.; Gargiulo, F.; Ciampi, M. Improving Biomedical Information Extraction with Word Embeddings Trained on Closed-Domain Corpora. In Proceedings of the 2019 IEEE Symposium on Computers and Communications (ISCC), Barcelona, Spain, 29 June–3 July 2019; IEEE: Piscatway, NJ, USA, 2019; pp. 1129–1134. [CrossRef]
24. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; ACL: Minneapolis, MN, USA, 2019; Volume 1, pp. 4171–4186. [CrossRef]