Report on the Fifth International Workshop on Health Data Management in the Era of AI (HeDAI 2023)

Haridimos Kondylakis
Univ. of Crete & FORTH-ICS
Heraklion, Greece
kondylak@ics.forth.gr

Kostas Stefanidis
Tampere University
Tampere, Finland
kostas.stefanidis@uta.fi

Praveen Rao
Univ. of Missouri-Columbia
Columbia, USA
praveen.rao@missouri.edu

ABSTRACT

Artificial intelligence (AI) technologies have the potential to drastically improve healthcare delivery globally. However, healthcare data are diverse, complex, and very large in size. Better information management techniques are needed to deal with the volume and heterogeneity of healthcare data. Better AI techniques are needed to fully harness healthcare datasets to improve healthcare delivery and quality of care for patients. Our workshop aimed to bring together diverse group of researchers and practitioners interested in developing next-generation solutions for challenging problems in healthcare systems. A wide range of topics ranging from federated approaches for data sharing to machine learning for predictive analytics to medical question answering were discussed during the workshop.

1. INTRODUCTION

As AI technologies are increasingly employed for the management of health data, new challenges occur daily that dictate new solutions. For example, bringing data into centralized repositories has become more difficult, mainly due to regulatory, competing interests, and trust issues. As such federated approaches are emerging allowing models to be sent to the data to be trained instead of accumulating all data central and then training the models on top. This is not only a result of technological advancements in privacy-preserving machine learning but also of the increased requirements to preserve the confidentiality and privacy of individuals. Further various types of data are being used for predictive analytics such as electronic health records, electroencephalography portable devices, and ultra-wideband radar sensors which dictate the effective management of the collected data and their usage for predictive downstream tasks. However, besides predictive tasks, other tasks such as personalized recommendations and intelligent question answering are equally important.

The goal of HeDAI 2023, co-organized with the EDBT conference in Ioannina, Greece, was to bring together researchers with interests cross-cutting the fields of Semantic Web, AI, data science, data management, and health informatics to discuss the challenges in healthcare data management and to propose novel and practical solutions for the next generation of data-driven healthcare systems. Developing optimal frameworks for integrating, curating, and sharing large volumes of clinical data has the potential for a tremendous impact on healthcare, enabling better outcomes at a lower and affordable cost. The ultimate goal is to enable innovations in Semantic Web, knowledge management, and data management for healthcare systems to move the needle to achieve the vision of precision medicine.

2. KEY TOPICS

In this section, we present the various topics that the workshop focused on through the various presentations and invited talks.

2.1 Real-World Federated Approaches for Healthcare Data Sharing and Analysis

The first keynote with title "Overcoming open issues and unmet needs in healthcare through the sharing, harmonization and federated analysis of unstructured medical data" was presented by Dimitrios I. Fotiadis, Prof. of Biomedical Engineering in the Department of Materials Science and Engineering, University of Ioannina, Greece that focused on overcoming open issues and unmet needs in healthcare through the sharing, harmonization, and federated analysis of unstructured medical data. The underlying heterogeneity and the reduced quality of the existing medical data across different clinical centers obscures the interlinking and co-analysis of such data. In addition, the legal and ethical barriers obscure the sharing of sensitive data and

highlight the need for the development of federated learning strategies to enable the federated analysis of medical data across different countries with inherent data protection policies. In this direction, a framework was presented allowing to overcome open issues and unmet needs through the design and development of (i) automated methods for data curation to address data inconsistencies and improve data quality in terms of relevance, conformity, and completeness, and (ii) hybrid data harmonization pipelines (i.e., pipelines for unifying disparate data fields, formats, dimensions, and columns into a composite dataset). Those are based on a combination of lexical and semantic matching methods with word embeddings which are utilized on top of medical index repositories and external knowledge bases to transform the heterogeneous data into a common standardized format, (iii) data augmentation through the design of robust virtual population generators to enhance the statistical power of databases with insufficient population size and improve the performance of the existing AI models, and (iv) federated AI algorithms to enable the training and evaluation of trustworthy and explainable AI workflows across high-quality and harmonized data stored in federated databases within a cloud environment. Multiple case studies were presented, and conducted across different clinical domains, including primary Sjögren's Syndrome (pSS) and hypertrophic cardiomyopathy (HCM), among others, to demonstrate the efficacy of the proposed framework to address clinical unmet needs.

2.2 Supporting Medical Trials with a Data Lake Federation

The second keynote with title "Supporting Medical Trials with a Data Lake Federation: a Research Perspective" was presented by Letizia Tanca, a full professor in data management at Politecnico di Milano, Italy. The talk focused on how to adopt a data lake federation in order to obtain significant results and benefits for medical organizations. Letizia Tanca started with the fact that the collection of data needed for clinical trials is very critical since a complete picture of the patient's status can only be obtained from real-world data, collected in different clinical institutions during their research and clinical history. A suitable solution to store and process the huge amount of necessary information, often coming from very heterogeneous devices and data sources, is needed to create the above-mentioned tremendous value. Data lake technology appears to be a promising solution for achieving the ability to manage and analyze data in healthcare. The goal is to be able to manage the complexity of the volume and variety of big data by providing data analysts with a self-service environment to which advanced techniques can be applied. Towards this direction, the adoption of a data lake federation, through which the involved medical organizations obtain significant results and new benefits is introduced. The extremely heterogeneous data collected in the data lakes of the federation must be accurately described, in order to document its quality, facilitate its discovery and integration, and define ethical, security, and privacy policies. Based on the experience in the Health Big Data project [5], the proposed architecture to collect and use data in the federation identifies the main IT research challenges we are facing nowadays.

2.3 Continuous Machine Learning for COVID-19

The COVID-19 pandemic caused a flurry of research in employing machine learning techniques for COVID-19 detection, mitigation, and treatment [20, 8, 16]. Avci et al. [4] in their paper entitled "Is My Model Up-to-date? Detecting CoViD-19 Variants by Machine Learning" proposed a continuous machine learning approach to detect COVID-19 variants. Their work is motivated by the need to update machine learning models due to concept drift [22]. They proposed an online deep-learning approach that used two neural networks. The first network called the primary neural network is used for output prediction. The second network called the secondary neural network is used for updating the model in the background. A performance estimator continuously checks for performance degradation. When degradation is detected, the primary neural network is replaced by the secondary neural network. The detection of concept drift is done by evaluating the area under the curve (AUC) metrics. The approach was evaluated using a real dataset containing patients that were tested positive or negative for COVID-19 using an RT-PCR test. It achieved better accuracy than an offline technique by 6% and an online binary classification technique by 5%.

2.4 Predictive Analytics Using Machine/Deep Learning

As AI models are successfully applied in many medical areas, exploring how to integrate AI models with medical domain taxonomies seems natural. In the paper entitled "Diagnosis Prediction over Patient Data using Hierarchical Medical Taxonomies"[9], the authors investigate how hierarchical medical taxonomies can be used to improve AI-based diagnostic tasks. In this direction, patient

graphs are extracted from electronic health records, and then the node embeddings of that patient graph are pre-initialized using information from medical taxonomies such as ICD-9 [2], ATC [1], and LOINC [3]. This step improves the performance of graph convolution network models over the enriched patient graph on diagnostic downstream tasks. Experiments performed using the MIMIC-IV dataset indeed confirm the usefulness of the proposed approach.

In another approach, Massa et al. [14] with their paper "Monitoring Human Attention with a Portable EEG Sensor and Supervised Machine Learning" employ cheap and portable electroencephalography (EEG) sensors for monitoring the human attention level. The acquired signals are processed using machine learning to estimate the attention level of people, enabling the continuous and unobtrusive monitoring of people, especially useful in the fields of rehabilitation and psychology. The authors propose a feature extraction technique based on sliding windows, and supervised machine learning distinguishing between attentive and distracted states, whereas their results show that it is highly beneficial in terms of accuracy to train the model first on the specific subject under investigation.

Another similar study is focusing on the estimation of respiration and heartbeat rates using impulse response ultra-wideband (IR-UWB) radar sensors with title "Respiration and heartbeat rates estimation using IR-UWB non-contact radar sensor recordings: A pre-clinical study" [15]. The authors first propose an architecture composed of simulated chest and heart monitors and they perform extensive recordings of the simulator's displacements, trying to identify the optimal mathematical estimation of the respiratory and heartbeat rates, compared to the initial frequencies given to the simulated procedure. The experimental results show that the proposed architecture is able to estimate sufficient quality rates in an accurate manner and even if these pre-clinical tests are made under ideal conditions, the UWB-radar can further be used for clinical assessment, with promising perspectives.

2.5 Privacy-Preserving Machine Learning

Federated learning has gained a lot of attention in healthcare [17, 23, 12] as it enables collaborative learning of machine learning models without sharing raw data between a client and a server. This protects the privacy of patient data, a critical requirement in healthcare data sharing. Split learning [7] improves upon federated learning wherein the model architecture and weights are not shared between a client and a server. However, split learning is prone to privacy leakage due to the reconstruction of the activation maps in deep learning models. Khan et al. [11] in their paper entitled "Split Ways: Privacy-Preserving Training of Encrypted Data Using Split Learning proposed to apply homomorphic encryption on the activation maps during split learning to protect user privacy. They focused on a U-shaped split learning model where the server only trains the fully connected layers of a convolutional neural network. The other layers are trained on a client. Using encrypted activation maps, they achieved comparable accuracy to training on plain text activation maps on an abnormal heart rhythm dataset. However, the training cost and communication overhead increased significantly. Further research is necessary to lower the overhead of employing homomorphic encryption in split learning.

2.6 Recommendation Systems

Recently, several approaches to recommendation systems exploit health-related information to provide suggestions, e.g., [19, 18]. Following a similar path, the paper entitled "SHARE: A Framework for Personalized and Healthy Recipe Recommendations" [24] proposes a personalized recommendation system that provides suggestions about recipes to users based on their health history and the preferences of similar users. Overall, the SHARE framework combines user tastes and nutritional information about the recipes in order to provide recommendations for recipes that meet the user's preferences and specific health needs. An alternative that offers personalized filtering for the users of the system is also presented. An experiment with real uses was performed and shows the system's ability to provide highly relevant personalized recommendations, using a large real-world data set of recipes.

2.7 Medical Question Answering Systems

Medical question answering systems [13, 6, 10] are growing in popularity due to the success of natural language processing (NLP) techniques and knowledge graphs. Most of the existing literature focuses on the English language. Tsampos et al. [21] with the paper entitled "A Medical Question Answering System with NLP and graph database" proposed a medical question answering system for the Greek language using NLP techniques and a graph database. Medical data in Greek are processed using sentence segmentation, tokenization, parts-of-speech tagging, and dependency parsing. The parse trees of sentences are represented as a graph and stored in an open-source graph database. In the

graph, the tokens are represented as nodes, each dependency is represented as a relation, and morphological features are represented as node properties. Questions are translated into graph queries and executed by the graph database. However, due to a lack of detailed evaluation, it was hard to judge the effectiveness of the proposed system.

3. CONCLUSIONS

A number of key observations and research directions emerged in the discussions during the workshop. Particular attention was given to large language models (LLMs) which are AI tools typically used to process and generate text. LLMs can improve the user-friendliness of a system since they can offer answers to questions in natural language, or they can summarize or even translate text on a level that is understandable from human capabilities. This way, users, or patients in our case, can actively interact with systems using LLMs in a transparent way, facilitating access to healthcare. Still, a systematic and comprehensive overview of the potentials and limitations of LLMs in the health domain is missing. Overall, HeDAI 2023 was successful in attracting international researchers to share their findings in health data management in the era of AI.

4. REFERENCES

[1] Atc/ddd index 2023. `https://www.whocc.no/atc_ddd_index/`. Accessed: 2024-01-15.

[2] International classification of diseases,ninth revision, clinical modification (icd-9-cm). `https://www.cdc.gov/nchs/icd/icd9cm.htm`. Accessed: 2024-01-15.

[3] The international standard for identifying health measurements, observations, and documents. `https://loinc.org/`. Accessed: 2024-01-15.

[4] O. Avci and G. Pozzi. Is My Model Up-to-date? Detecting CoViD-19 Variants by Machine Learning. In G. Fletcher and V. Kantere, editors, *Proceedings of the Workshops of the EDBT/ICDT 2023 Joint Conference*, 2023.

[5] S. Bahri, N. Zoghlami, M. Abed, and J. M. R. S. Tavares. Big data for healthcare: A survey. *IEEE Access*, 7:7397–7408, 2019.

[6] A. Ben Abacha and P. Zweigenbaum. Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information Processing Management*, 51(5):570–594, 2015.

[7] O. Gupta and R. Raskar. Distributed learning of deep neural network over multiple agents. *Journal of Network and Computer Applications*, 116:1–8, 2018.

[8] L. O. Hall, R. Paul, D. B. Goldgof, and G. M. Goldgof. Finding COVID-19 from chest X-rays using deep learning on a small dataset. *arXiv preprint arXiv:2004.02060*, 2020.

[9] E. R. Hansen, T. Sagi, and K. Hose. Diagnosis prediction over patient data using hierarchical medical taxonomies. In G. Fletcher and V. Kantere, editors, *Proceedings of the Workshops of the EDBT/ICDT 2023 Joint Conference, Ioannina, Greece, March, 28, 2023*, volume 3379 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023.

[10] Z. Jiang, C. Chi, and Y. Zhan. Research on medical question answering system based on knowledge graph. *IEEE Access*, 9:21094–21101, 2021.

[11] T. Khan, K. Nguyen, and A. Michalas. Split ways: Privacy-preserving training of encrypted data using split learning. *CoRR*, abs/2301.08778, 2023.

[12] A. Korkmaz, A. Alhonainy, and P. Rao. An evaluation of federated learning techniques for secure and privacy-preserving machine learning on medical datasets. In *2022 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–7. IEEE, 2022.

[13] M. Lee, J. Cimino, H. R. Zhu, C. Sable, V. Shanker, J. Ely, and H. Yu. Beyond information retrieval—medical question answering. In *AMIA annual symposium proceedings*, volume 2006, page 469. American Medical Informatics Association, 2006.

[14] S. M. Massa, G. Usai, and D. Riboni. Monitoring human attention with a portable EEG sensor and supervised machine learning. In G. Fletcher and V. Kantere, editors, *Proceedings of the Workshops of the EDBT/ICDT 2023 Joint Conference, Ioannina, Greece, March, 28, 2023*, volume 3379 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023.

[15] A. Pentari, D. Manousos, T. Kassiotis, G. Rigas, and M. Tsiknakis. Respiration and heartbeat rates estimation using IR-UWB non-contact radar sensor recordings: A pre-clinical study. In G. Fletcher and V. Kantere, editors, *Proceedings of the Workshops of the EDBT/ICDT 2023 Joint Conference, Ioannina, Greece, March, 28, 2023*, volume 3379 of *CEUR Workshop*

Proceedings. CEUR-WS.org, 2023.

[16] S. Prasanna and P. Rao. A data science perspective of real-world COVID-19 databases. In L. Gruenwald, S. Jain, and S. Groppe, editors, *Leveraging Artificial Intelligence in Global Epidemics*, pages 133–163. Academic Press, 2021.

[17] N. Rieke, J. Hancox, W. Li, F. Milletari, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, et al. The future of digital health with federated learning. *NPJ digital medicine*, 3(1):119, 2020.

[18] M. Stratigi, H. Kondylakis, and K. Stefanidis. Fairgrecs: Fair group recommendations by exploiting personal health information. In *DEXA*, 2018.

[19] M. Stratigi, H. Kondylakis, and K. Stefanidis. Multidimensional group recommendations in the health domain. *Algorithms*, 13(3):54, 2020.

[20] H. B. Syeda, M. Syed, K. W. Sexton, S. Syed, S. Begum, F. Syed, F. Prior, and F. Yu Jr. Role of machine learning techniques to tackle the COVID-19 crisis: systematic review. *JMIR medical informatics*, 9(1):e23811, 2021.

[21] I. Tsampos and E. I. Marakakis. A medical question answering system with NLP and graph database. In G. Fletcher and V. Kantere, editors, *Proceedings of the Workshops of the EDBT/ICDT 2023 Joint Conference, Ioannina, Greece, March, 28, 2023*, volume 3379 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2023.

[22] A. Tsymbal. The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, 106(2):58, 2004.

[23] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5:1–19, 2021.

[24] K. Zioutos, H. Kondylakis, and K. Stefanidis. SHARE: A framework for personalized and healthy recipe recommendations. In *Proceedings of the Workshops of the EDBT/ICDT 2023 Joint Conference*, 2023.