IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Date of current version 6 December 2022.

Digital Object Identifier 10.1109/TETC.2022.3210449

Guest Editorial: Special Section on "To be Safe and Dependable in the Era of Artificial Intelligence: Emerging Techniques for Trusted and Reliable Machine Learning"

SHANSHAN LIU, (Member, IEEE), PEDRO REVIRIEGO, (Senior Member, IEEE), FABRIZIO LOMBARDI, (Life Fellow, IEEE), AND PATRICK GIRARD, (Fellow, IEEE)

During the last decade, advances in areas such as convolutional neural networks, deep learning, and hardware accelerators among others, have enabled the widespread and ubiquitous adoption of machine learning in real-world systems. This trend is expected to continue and expand in coming years leading to a world that depends heavily on machine learning-based systems.

To be safe and dependable in this new era of artificial intelligence requires that these innovative systems must be reliable and secure. This poses many research challenges. For example, fault tolerance is commonly achieved by redundant design, but the implementation of deep neural networks is already challenging, so there is little room to add additional elements for fault tolerance. Similarly, understanding the vulnerabilities of advanced machine learning (ML) systems is a complex issue as shown by recent attacks on image classification implementations. Therefore, it is essential to "learn" how to build ML systems that cannot be manipulated or corrupted by malicious attackers and that can operate reliably when its underlying hardware or software suffers from errors.

As Guest Editors, we are delighted and honored to introduce to the readership of *IEEE Transactions on Emerging Topics in Computing (TETC)* this Special Section on "To be Safe and Dependable in the Era of Artificial Intelligence: Emerging Techniques for Trusted and Reliable Machine Learning." It consists of six papers that have been accepted through a rigorous peer-review process. Each submission has been reviewed by three experts in its research area, and only papers that have consistently shown the highest quality standards, have been accepted.

The first paper is titled "On the Properness of Incorporating Binary Classification Machine Learning Algorithms into Safety-Critical Systems" and authored by M. Gharib, et al [1]. ML algorithms are being incorporated into many systems due to its capability to solve complex problems. Specifically, Safety-Critical Systems (SCS) have used ML to meet stringent requirements in their operation. However, performance of ML algorithms usually relies on metrics that were not developed with safety in mind. This paper debates on accounting for the distribution, so not just based on the number of so-called False Negatives when assessing ML algorithms to be integrated into SCS. The properness of incorporating ML-based components (anomaly-based intrusion detectors) into SCS is empirically assessed using both traditional and novel metrics that focus on the numbers as well as the distribution of False Negatives.

The second paper of this Special Section is by T. Wang et al. with title "Secrecy driven Federated Learning via Cooperative Jamming: An Approach of Latency Minimization" [2]. This paper considers Federated Learning (FL) to provide a promising framework for enabling distributed ML based services without revealing users' private data. The scenario of wireless FL is taken as primary application; in this case it encounters an eavesdropping attack when the parameter-server sends the model-data to the wireless devices; hence, a secrecy driven FL is proposed via cooperative jamming. In this scheme, wireless devices cooperatively provide jamming to the eavesdropper to enhance secure throughput based on the measure of physical layer security. Numerical results are provided to validate the effectiveness of the proposed algorithms as well as the performance advantage of the proposed secrecy driven FL via cooperative jamming; these results show that this scheme can reduce latency by more than 35% compared to the case with no jamming.

In this Special Section, the third paper is given by "An Interrelated Imitation Learning Method for Heterogeneous Drone Swarm Coordination," (with B. Yang et al as authors) [3]. As the proliferation of small drones has boosted diverse intelligent services, swarm coordination plays a vital role in enhancing their execution efficiency; however, due to unreliable air communication and heterogeneous computation capabilities, coordinated actions (particularly in distributed scenarios with incomplete observations) will likely encounter many difficulties. In this paper, the generative adversarial

imitation learning model is employed to coordinate the drones' maneuvers by imitating the peer's demonstrations. The proposed approach is evaluated by the drones' formation control task; experimental results show that performance metrics (such as imitation accuracy, execution time and energy) are significantly improved.

Next, the paper "Improving the Reliability of Network Intrusion Detection Systems through Dataset Aggregation" (by R. Magan-Carrion, et al.) deals with a novel methodology for ML based Network Intrusion Detection Systems (NIDSs) [4]. This new methodology allows ML models to work on integrated datasets, so empowering the learning process with diverse information from different datasets. The proposed scheme targets the design of more robust models, that generalize better than traditional approaches. Also, two well-known ML models (linear and non-linear) based on the information of three of the most common datasets in the literature, are built. The results that the proposed methodology offers show that these two ML models (trained as a NIDS solution) significantly benefit from this approach, so being able to generalize better when training using the approach here presented for dataset integration.

The paper "An Approximate Memory based Defense against Model Inversion Attacks to Neural Networks," (by Q. Xu, et al.) deals with a different aspect of reliable neural network operation [5]. It is well known that a diverse and comprehensive training data is critical in building robust ML models. However, model inversion attacks (MIA) have demonstrated that an ML model can leak important information about its training dataset. This work examines existing MIAs and proposes a hardware-oriented solution to protect the training data from such attacks. Experiments show that the proposed approach can efficiently provide protection of training data from run-time adversarial attacks.

The last paper of this Special section is "The Role of Explainability in Assuring Safety of Machine Learning in Healthcare (by Y. Jia, et al.) [6]. Current approaches to assuring safety-critical systems and software are difficult to apply to systems employing ML because there is no clear, pre-defined specification against which to assess validity. Little work has been reported to explicitly investigate the role of explainability for safety assurance in the context of ML development. This paper identifies ways in which Explainable AI (XAI) methods can contribute to safety assurance of ML-based systems; it then uses a concrete ML-based clinical decision support system (concerning weaning of patients from mechanical ventilation) to demonstrate the way XAI methods can be employed to produce evidence to support safety assurance. Therefore, XAI methods have a valuable role in safety assurance of ML-based systems in healthcare, but that they are not sufficient in themselves to assure safety.

We are confident that you will find these papers to be of invaluable help in your ongoing research investigations and appreciating them as we have enjoyed organizing this Special Section.

SHANSHAN LIU, *Guest Editor*
New Mexico State University, Las Cruces, 88001 USA
PEDRO REVIRIEGO, *Guest Editor*
Universidad Politécnica de Madrid, 28040, Madrid, Spain
FABRIZIO LOMBARDI, *Guest Editor*
Northeastern University, Boston, 02215 USA
PATRICK GIRARD, *(Corresponding TETC Editor)*
Laboratory of Informatics, Robotics and Microelectronics of Montpellier, 34095, Montpellier, France

APPENDIX AND RELATED WORK

[1] M. Gharib, T. Zoppi, and A. Bondavalli, "On the properness of incorporating binary classification machine learning algorithms into safety-critical systems," *IEEE Trans. Emerg. Topics Comput.*, to be published, doi: 10.1109/TETC.2022.3178631.

[2] T. Wang, Y. Li, Y. Wu, and T. Q. S. Quek, "Secrecy driven federated learning via cooperative jamming: An approach of latency minimization," *IEEE Trans. Emerg. Topics Comput.*, to be published, doi: 10.1109/TETC.2022.3159282.

[3] B. Yang, C. Ma, and X. Xia, "An interrelated imitation learning method for heterogeneous drone swarm coordination," *IEEE Trans. Emerg. Topics Comput.*, to be published, doi: 10.1109/TETC.2022.3202297.

[4] R. Magán-Carrión, D. Urda, I. Diaz-Cano, and B. Dorronsoro, "Improving the reliability of network intrusion detection systems through dataset aggregation," *IEEE Trans. Emerg. Topics Comput.*, to be published, doi: 10.1109/TETC.2022.3178283.

[5] Q. Xu, M. T. Arafin, and G. Qu, "An approximate memory based defense against model inversion attacks to neural networks," *IEEE Trans. Emerg. Topics Comput.*, to be published, doi: 10.1109/TETC.2022.3179980.

[6] Y. Jia, J. McDermid, T. Lawton, and I. Habli, "The role of explainability in assuring safety of machine learning in healthcare," *IEEE Trans. Emerg. Topics Comput.*, to be published, doi: 10.1109/TETC.2022.3171314.

Shanshan Liu (Member, IEEE) received the PhD degree in microelectronics and solid-state electronics from the Harbin Institute of Technology, Harbin, China, in 2018. She was a post-doctoral researcher with the Department of Electrical and Computer Engineering (ECE), Northeastern University, Boston, USA, from 2018 to 2021, and is currently an Assistant Professor with the Klipsch School of ECE, New Mexico State University, Las Cruces, USA. She serves as an associate editor for the *IEEE Transactions on Emerging Topics in Computing* and the *IEEE Transactions on Nanotechnology*, a guest editor for the *IEEE Transactions on Circuits and Systems-I*, and a technical program committee member for the IEEE International Symposiums on DFT, NANO and IOLTS. Her research interests include fault tolerance design in high performance computing systems, emerging computing, VLSI design, dependable machine learning, error correction codes.

Pedro Reviriego (Senior Member, IEEE) received the MSc and PhD degrees in telecommunications engineering from the Technical University of Madrid, Madrid, Spain, in 1994 and 1997, respectively. From 1997 to 2000, he was an engineer with Teldat, Madrid, working on router implementation. In 2000, he joined Massana to work on the development of 1000BASE-T transceivers. From 2004 to 2007, he was a distinguished member of technical staff with the LSI Corporation, working on the development of Ethernet transceivers. From 2007 to 2018 he was with Nebrija University and from 2018 to 2022 with Universidad Carlos III de Madrid. He is currently with Universidad Politécnica de Madrid working on several topics in computer science with a focus on security, privacy and reliability. He is currently associate editor of the *IEEE Transactions on Emerging Topics in Computing*.

Fabrizio Lombardi (Life Fellow, IEEE) received the BSc degree (Hons.) in electronic engineering from the University of Essex, U.K., in 1977, the master's degree in microwaves and modern optics, the diploma degree in microwave engineering from the Microwave Research Unit, University College London, in 1978, and the PhD degree from the University of London, in 1982. He is currently the International Test Conference (ITC) Endowed Chair Professorship with Northeastern University, Boston, USA. His research interests are bio-inspired and nano manufacturing/computing, VLSI design, testing, and fault/defect tolerance of digital systems. He has extensively published in these areas and coauthored/edited seven books. He was the editor-in-chief of the *IEEE Transactions on Computers* from 2007 to 2010, and the inaugural editor-in-chief of the *IEEE Transactions on Emerging Topics in Computing* from 2013 to 2017, *IEEE Transactions on Nanotechnology* from 2014 to 2019. He is currently the 2022/23 president of the IEEE Nanotechnology Council and a member of the IEEE Publication Services and Products Board.

Patrick Girard (Fellow, IEEE) received the MSc degree in electrical engineering, and the PhD degree in microelectronics from the University of Montpellier, France, in 1988 and 1992, respectively. He is currently a research director with CNRS (French National Center for Scientific Research), and works with the Microelectronics Department of the Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM), France. From 2010 to 2014, he was the Head of the Microelectronics Department. He is the co-director of the International Associated Laboratory <<LAFISI>> (French–Italian Research Laboratory on Hardware-Software Integrated Systems) created, in 2013 by the CNRS and the University of Montpellier with the Politecnico di Torino, Italy. He has supervised 40 PhD dissertations and has published seven books or book chapters, 80 journal papers, and more than 250 conference and symposium papers on these fields. His research interests include all aspects of digital testing and memory testing, with emphasis on critical constraints, such as timing and power. Reliability and fault tolerance are also part of his research activities. He is an associate editor of the *IEEE Transactions on Aerospace and Electronic Systems*, *IEEE Transactions on Emerging Topics in Computing*, and the *Journal of Electronic Testing: Theory and Applications* (Springer).