**INTRODUCTION**

Internet of Things (IoT) devices are increasingly integrated in cyber-physical systems (CPS), including in critical infrastructure sectors such as dams and utility plants. In these settings, IoT devices (also referred to as Industrial IoT or IIoT) are often part of an Industrial Control System (ICS), tasked with there liable operation of the infrastructure. ICS can be broadly defined to include supervisory control and data acquisition (SCADA) systems, distributed control systems (DCS), and systems that comprise programmable logic controllers (PLC) and Mod bus protocols. The connection between ICS or IIoT-based systems with public networks, however, increases their attack surfaces and risks of being targeted by cyber criminals. One high-profile example is the Stux net campaign, which reportedly targeted Iranian centrifuges for nuclear enrichment in 2010, causing severe damage to the equipment. Another example is that of the incident targeting a pump that resulted in the failure of an Illinois water plant in 2011. Black Energy was another campaign that targeted Ukraine power grids in 2015, resulting in power outage that affected approximately 230,000 people. In April 2018, there were also reports of successful cyber-attacks affecting three U.S. gas pipeline firms, and resulted in the shutdown of electronic customer communication systems for several days. Although security solutions developed for information technology (IT)and operational technology (OT) systems are relatively mature, they may not be directly applicable to ICSs. For example, this could be the case due to the tight integration between the controlled physical environment and the cyber systems. Therefore, system-level security methods are necessary to analyze physical behavior and maintain system operation availability. ICS security goals are prioritized in the order of availability, integrity, and confidentiality, unlike most IT/OT systems (generally prioritized in the order of confidentiality, integrity, and availability). Due to close coupling between variables of the feedback control loop and physical processes, (successful) cyber-attacks on ICS can result in severe and potentially fatal consequences for the society and our environment. This reinforces the importance of designing extremely robust safety and security measurements to detect and prevent intrusions targeting ICS. Popular attack detection and attribution approaches include those based on signatures and anomalies. To mitigate the known limitations in both signature-based and anomaly-based detection and attribution approaches, there have been attempts to introduce hybrid-based approaches. Although hybrid based approaches are effective at detecting unusual activates, they are not reliable due to frequent network upgrades, resulting in different Intrusion Detection System (IDS) typologies. Beyond this, conventional attack detection and attribution techniques mainly rely on network metadata analysis (e.g. IP addresses, transmission ports, traffic duration, and packet intervals). Therefore, there has been renewed interest in utilizing attack detection and attribution solutions based on Machine Learning (ML) or Deep Neural Networks (DNN) in recent times.7In addition, attack detection approaches can be categorized into network-based or host-based approaches. Supervised clustering, single-class or multi-class Support Vector Machine (SVM), fuzzy logic, Artificial Neural Network (ANN), and DNN are commonly used techniques for attack detection in network traffic. These techniques analyze real-time traffic data to detect malicious attacks in a timely manner. However, attack detection that considers only network and host data may fail to detect sophisticated attacks or insider attacks. Unsupervised models that incorporate process/physical data can complement a system’s monitoring since they do not rely on detailed knowledge of the cyber-threats. In general, a sophisticated attacker with sufficient knowledge and time, such as a nation state advanced persistent threat actor, can potentially circumvent robust security solutions. Furthermore, most of the existing approaches ignore the imbalanced property of ICS data by modeling only a system’s normal behavior and reporting deviations from normal behavior as anomalies. This is, perhaps, due to limited attack samples in existing datasets and real-world scenarios. Although using majority class samples is a good solution to avoid issues due to imbalanced datasets, the trained model will have no view of the attack samples’ patterns. In other words, such an approach fails to detect unseen attacks and suffers from a high false positive rate [8]. Thus, there have been attempts to utilize DL approaches, for example, to facilitate automated feature (representation) learning to model complex concepts from simpler ones [9] without depending on human-crafted features. Motivated by the above observations, this paper presents our proposed novel two-stage ensemble deep learning-based attack detection and attack attribution framework for imbalanced ICS datasets. In the first stage, an ensemble representation learning model combined with a Decision Tree (DT) is designed to detect attacks in an imbalanced environment. Once the attack is detected, several one-vs-all classifiers will ensemble together to form a larger DNN to classify the attack attributes with a confidence interval during the second stage. Moreover, the proposed framework is capable of detecting unseen attack samples. A summary of our approach in this study is as follows:

* We develop a novel two-phase ensemble ICS attack detection method capable of detecting both previously seen and unseen attacks. We will also demonstrate that the proposed method out performs other competing approaches in terms of accuracy and f-measure. The proposed deep representation learning results in this method being robust to imbalanced data.
* We propose a novel self-tuning two-phase attack attribution method that ensembles several deep one-vsall classifiers using a DNN architecture for reducing false alarm rates. The proposed method can accurately attribute attacks with high similarity. This is the first ML-based attack attribution method in ICS/IIoT at the time of this research.
* We analyze the computational complexity of the proposed attack detection and attack attribution framework, demonstrating that despite its superior performance, its computational complexity is similar to that of other DNN-based methods in the literature.