5th INTERNATIONAL CONFERENCE ON RESEARCH IN HUMANITIES
26-28 August 2022 , Cambridge, United Kingdom

A Real-Time Intrusion Detection System for Bio-Medical and Cyber-Physical Systems

Ajay Kaushik[1], V. Samuel Raj[2]

[1]Department of Computer Science and Engineering, SRM University, Delhi-NCR, India
[2]Professor, Dean Academics and Registrar, SRM University, Delhi-NCR, India

Abstract

The industrial sector has made the shift from embedded systems to cyber-physical systems (CPSs) with the introduction of Industry 4.0 and the Internet of Things (IoT). Manufacturers may now use data collected from on-board sensors to monitor the operation of industrial equipment. Industrial systems may be monitored in this way, and abnormalities can be identified. As the quantity of visible data grows, many firms are forced to explore for new ways to deal with such vast volumes of data. Timely detection of intrusion in industrial and commercial data has emerged as essential enablers for satisfying the scientific community's current expectations for CPSs and bio-medical devices in recent years. Big data technology might be used to build a platform for CPS real-time monitoring at work. This research provides a Bio-medical and CPS-Intrusion Detection System built with Apache Spark (CPS-IDS). Using real-world datasets, the proposed technique is verified. This system's effectiveness, as well as its suitability and scalability for future demand, are evaluated in a real-world situation. The overall efficiency of the equipment has enhanced as a result of this method.

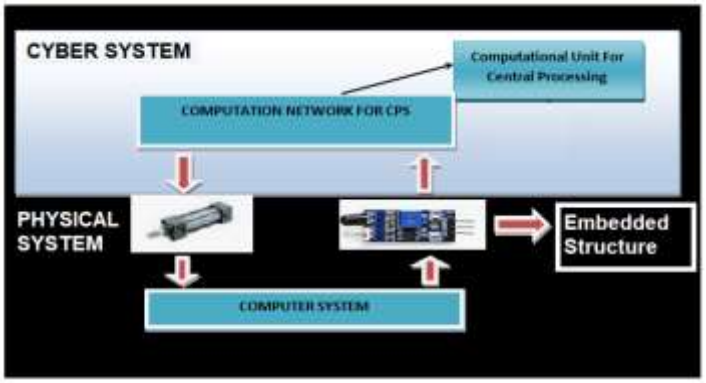
Keywords. Artificial Intelligence, Cyber Security, Cyber Dataset, Machine Learning, IoT

1. Introduction

Cyber Physical Systems (CPS) and IoT have played a crucial role in technological revolution in last decade [1]. Progress has been facilitated by the particular requirements of the industrial manufacturing sector [2]. Businesses saw a technical shift as a consequence of moving away from traditional embedded systems and toward Cyber-Physical Systems (CPSs) [3-4]. Because of the emphasis of this research on manufacturing, the term CPS is assumed to relate to an Industrial Cyber-Physical System (ICPS) rather than a broader variety of systems. Because of the utilisation of ICPSs, manufacturers may apply more sophisticated approaches to improve and optimise production processes across their whole manufacturing system and product lifecycle. As a consequence, not only would the quality of the items improve, but so would productivity and energy consumption. Global companies will invest euro143 billion in this area by 2020, increasing the worldwide M&C market investment to e500 billion [5-6].

It is possible for an ICPS to include a wide range of devices. To enable intelligent manufacturing [7-10], a number of these devices may communicate with one another to make decisions while the system is operating. Since these industrial systems must be monitored cost-effectively, it is essential that all data coming from ICPSs be gathered so that problems may be discovered in a timely manner and production can continue without interruption. The data that has been collected may be used to identify any problems with the system. The early identification of problems is made possible by this anomaly detection. It is necessary to capture and analyse all of the ICPS data in order to do this. As previously said, an ICPS might consist of several devices, resulting in a large amount of data being received. As a consequence, the current data volumes lead in delays and may even result in non-functioning. An industrial equipment problem may also have a negative impact on production if it is discovered too late. Industrial environments need that the system remain active at all times, seven days a week. As a consequence, any interruptions to a smart manufacturing system's equipment, network, and so on must be avoided in order to avoid a decrease in production and financial loss. A typical CPS is shown in Fig. 1

Figure 1 – Cyber Physical System



Increasing amounts of data have been generated due to the rapid development and widespread usage of sensor and IoT frameworks. It is necessary to horizontally expand the ICPS in order to accommodate more computer resources for data import, analysis, and storage as the quantity of data increases. Other issues, such as data partitioning [11] or resource management, may arise if more processing nodes are inducted. As a consequence, scalability is a major issue with these systems. Because cloud infrastructure allow ICPSs to handle large amounts of data quickly, scalable, and fault-tolerant, they are particularly important in this context. For small and medium organisations, cloud-based approaches are ideal since they give on-demand services with minimal impediments and upfront costs [12]. Adding more processing nodes to the current server to distribute the load may lead to other challenges, such as data partitioning [11] or resource management. As a result, these systems struggle with scalability. When it comes to the fast handling of data, the cloud that accompany them are critical [5] [12]. A range of industrial fields have evolved to depend increasingly on physical and digital component communication [13-15]. Smart production is now possible thanks to the widespread use of ICPSs throughout industry [21].

The ability to exchange information between analogue and digital components has become critical in a variety of industrial settings [16-21]. Manufacturing has become more efficient because to ICPSs. In order to effectively monitor industrial systems, an ICPS by itself is insufficient. It is necessary to go through four phases of Internet of Things (IoT) in order to acquire enough data to make an appropriate conclusion. Connecting the devices, monitoring the operational condition of the systems and eventually achieving desired business results are all steps in which data is sent and monitored through real-time monitoring. The ICPS's performance may then be improved with the use of data analytics. Finally, an on-board

intelligence that maximises the commercial value of the information gathered in the preceding steps is needed. Data from industrial systems may be gathered and processed in this way, allowing businesses to get important information for making informed choices, diagnosing problems, and implementing preventative maintenance strategies, among other things [22-25]. Lee et al. [26] argue that algorithms are necessary for drawing conclusions and preventing anomalies because of the necessity to examine data from physical components as rapidly as possible and the vast volume of data collected. Analyzing data in this way is likely to be more efficient than doing it by hand. With a significant amount of data, people are unable to draw conclusions fast and effectively, according to Niggemann et al. [27]. In addition to being inadequate, human-managed systems, they claim to be difficult to maintain.

Security experts estimate that it takes six months to uncover a virus or malware infestation in an organisation, and only if a preset bad activity is detected by the firm's intrusion detection systems [3]. (IDS). Distributed denial-of-service (DDoS) attackers often employ the approach of accessing business networks without generating security alarms to carry out their operations. This paper presents supervised machine learning strategies for early identification of Bio-Medical and CPS intrusions. It is also compared to the results of integrating NetFlow with Hadoop Framework: MapReduce and the latest Hadoop Framework: Spark. The rapid and accurate detection of anomalies in real-time data has made it possible to make advancements in real-time data analysis. For example, the CPS-ID architecture is designed to detect anomalies in active networks within a short period of time and identify anomalies. The proposed paradigm addresses the gap in network security research by including these two fundamental aspects. In addition, a detailed assessment is carried out to guarantee that the suggested framework fits the performance standards originally stated. Evaluation findings were also utilised for benchmarking against those of other systems.

2. Materials and Methods

2.1 Data Acquisition and Mapping

The dynamic nature of NetFlow traffic makes anomaly identification difficult. General network anomaly detection and analysis are the primary topics of this study. The datasets are imported from Kaggle and other online data sources. Data processing should be able to handle increases in data volume and cope with system errors (scalability) (fault-tolerant). As the amount of data increases, so does the need for storage flexibility. In addition, a fast search engine is required in order to access the database quickly. In order to handle both immediate (such as anomaly alerts or the present machine condition) and long-term data, this system must include data

serving mechanisms (i. e., advanced analytics). As discrete components, the procedure is divided into 4 key stages, including:

1. Servers and/or selectors for collecting NetFlow traffic Netflow data pre-processing
2. Spark cluster simulation
3. Detector of anomalies
4. A graph or notice to the concerned body in the form of a message.

2.2 Pre-processing of Data

Network trace preparation and data preprocessing approaches are crucial during preprocessing, which leads to improved outcomes in the detection process. Reviewers found that many research restrict their preprocessing methods and do not go into depth. Preprocessing and anomaly identification will commence as soon as network traffic files arrive at this step. Pretreatment of network-based intrusion detection data consumed 50% of the total effort and provided a better categorization of network traffic as normal or abnormal since data preprocessing is a necessary step in all knowledge discovery activities. Learning-based algorithms benefit from removing strong co-relational, irrelevant, and duplicated information. For this work, we devised an unique preprocessing method to concentrate on the Netflow preprocessing step. We looked at numerous formal process models for KDDM (knowledge discovery and data mining).

2.3. CPS Intrusion Detection System

We believe that intrusion prevention strategies are inherently flawed, and that network intrusion detection systems play a critical role in preventing security breaches (IDSs). Near-real-time detection of suspicious behaviour is the goal of our proposed NIDs. In intrusion-based detection systems, an anomaly may be detected to identify several assaults, resulting in network failures in real time or setting the stage for future disasters. In terms of cost, system response time, and customer happiness, it is better to address anomaly detection early on. Anomaly identification in this research is aided by the following super vised approaches. Random forest, Decision Tree, Support Vector Machines, K-Nearest Neighbors Algorithm.

2.4. Simulation using Spark

As part of our design, Spark takes over once the dataset has been preprocessed. Specifically, we assume in this research that the isolated spark cluster is composed of five worker nodes, including a master node. As an extension to the famous MapReduce architecture, Spark is aimed to provide a near-real-time response rate for new sorts of requests, including interactive searches and stream processing. Faster processing of huge datasets makes a big difference in how quickly you can access and analyse data vs how long you have to wait. Memory-based computing is one of Spark's most important speed-enhancing features. Even so, for disk-based sophisticated applications, this method outperforms MapReduce. There are many different types of tasks that traditionally needed distinct systems. In addition, Apache Spark decreases the administrative overhead of maintaining various tools. A wide range of programming languages are supported by Spark, including R, Python, Scala, and many more.

2.5. Identification

The data flow utilised to monitor industrial systems in real time and spot issues early is discussed in this section. A data acquisition system is used to acquire data from industrial machines. PLCs (Programmable Logic Controllers), which are installed on each piece of industrial equipment, provide data to a local database. In order to determine the significance of the anomaly, a low and a high boundary were drawn around it. These are the predetermined boundaries set by experts. As a result, a yellow flag is generated if the computation's result exceeds the low boundary. A red signal is generated if it surpasses the high threshold of the system. If this is not the case, the flag will be coloured green. In order to raise the severity of an anomaly, flag colours may only be altered. An alert is only sent in this case. In order to reset the alarms to green, a certain procedure must be followed.

3. Results.

The real time traffic is inducted for this research. An overview of detection times for various detection models can be shown in Figure 2. Using the horizontal axis, we plotted the total number of packets. We used seconds as the unit of time on the vertical axis. It is evident that decision tree has the performance in terms of detection time. This is due to its superior classification performance as compared with random forest and SVM.

Figure 2 – Packet size vs Detection Time

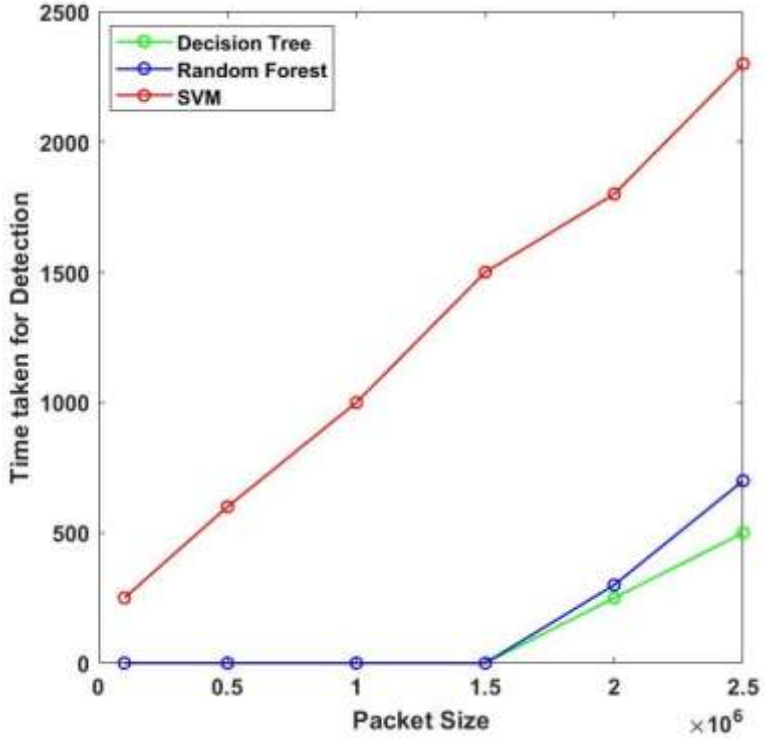
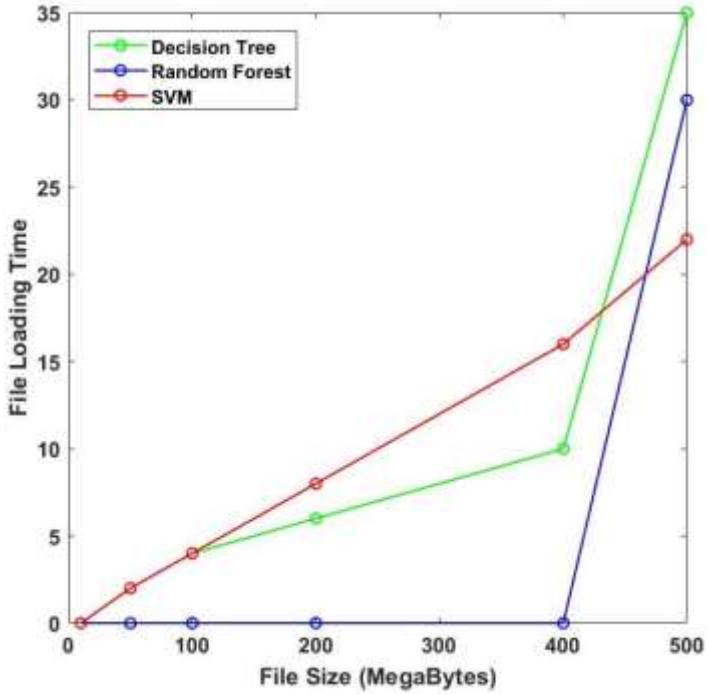


Fig. 3 depicts the time it takes for various models to create a log file, as well as the size of that log file. The log file size was represented on the x plane by [10M, 50MB, 100MB, 400MB, and 500MB]. We measured time in seconds, ranging from one second to one hundred and twenty seconds, on the vertical axis. It is seen that random forest has the best file loading time and SVM possesses the worst computation time for loading the files.

Figure 3 – File size vs File loading time



Using three independent measurements, the time it took for each model to accumulate the entire number of packets is shown in Fig. 4. File size and packet count both increase the amount of time it takes for pre-processing. However, we also discovered that it is not directly proportional to the number of packets that are being sent. While processing a small piece of packet count is reliant on the classification model, processing a big chunk is not. For 100K packets, Decision Tree is still the best pre-processing model, and it took just 0.85 seconds to pre-process, whereas 3000K packets took around 312 seconds. Only [1.14 seconds and 322 seconds] for 100K and 3400K packets count, respectively, RF was the fastest model. Even after all these years, SVM is the worst pre-processing and classifier out there.

Figure 4 – Number of Packets vs Pre-processing Time

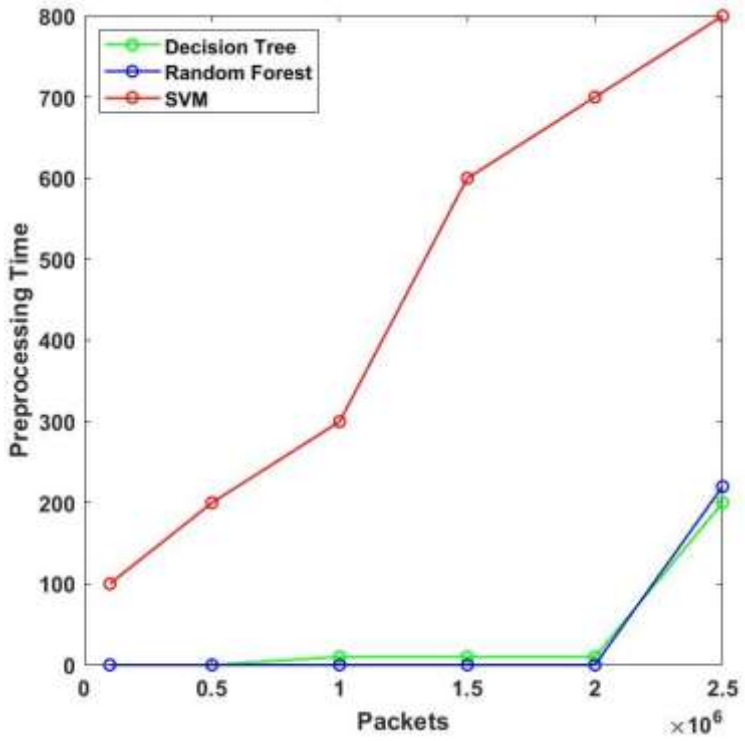


Fig. 5 represents a comparison of file size vs time taken for data transfer by the proposed approach. It is seen that the processing time increases with file size. Specifically, the computation time shows a sharp rise when file size crosses 6 units. Adding all of the necessary elements for optimal operation and increasing accuracy and time consumed by recording and transmitting such traffic volume to the testbed/deployed framework, however, necessitates additional pre-processing effort in terms of time.

Figure 5 – Depiction of File size vs Transfer Time by proposed approach

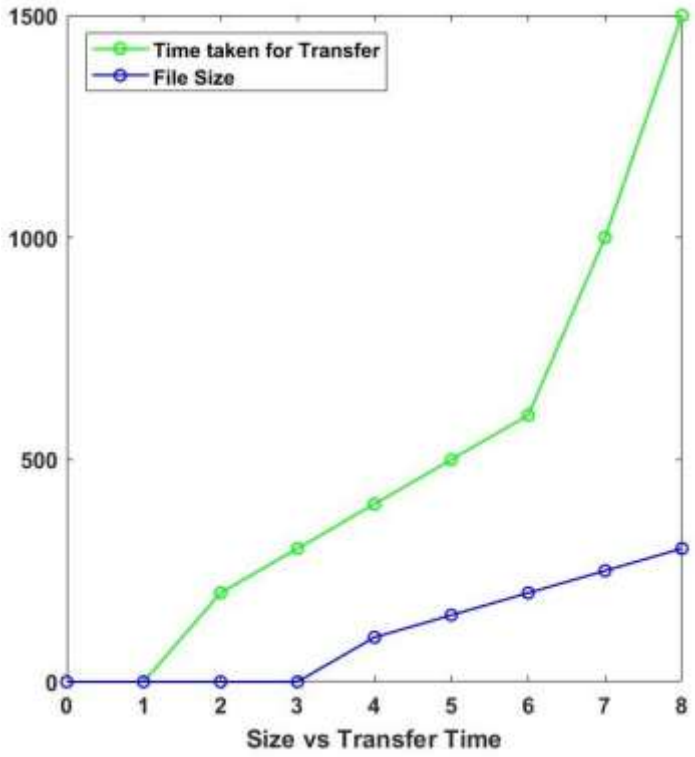
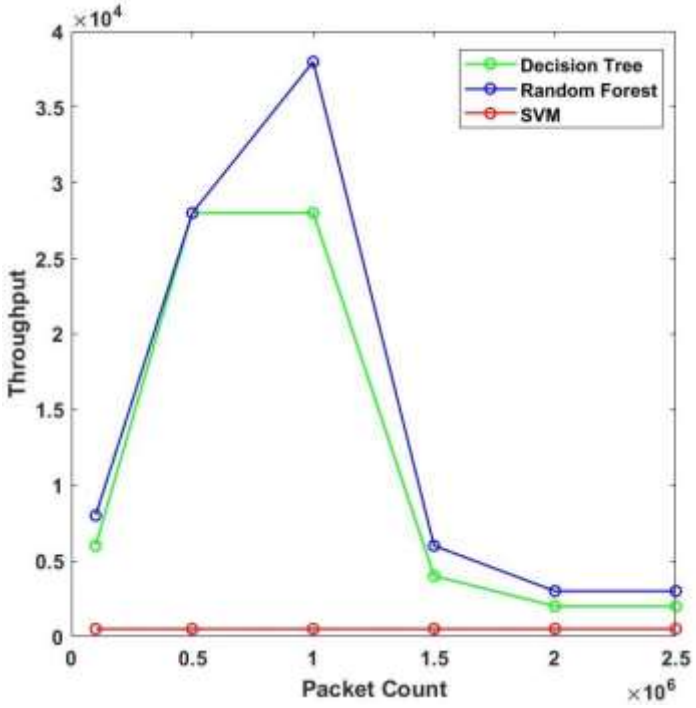


Figure 6 depicts throughput as a function of time. It is evident from Fig. 6 that random forest has the best throughput as compared with Decision tree and SVM. The worst performance is displayed by SVM. Throughput is a measure of overall system performance in terms of anomaly detection. Random forest outperforms decision tree and SVM in terms of overall system performance.

Figure 6 – Packet count vs Throughput



4. Conclusion

This article presents CPS-IDS, Spark-based intrusion detection system that can analyse possible assaults in CPS with no time delays. DDoS flooding assaults may be detected using ML-Spark algorithms, which are implemented in CPS-IDS and used to preprocess real network data. CPS-IDS uses low latency and high efficiency parallel data processing to address the traditional solution's scalability, memory inefficiency, and process complexity concerns. These ICPSs have been successfully monitored in a real-world industrial setting with several press machines using a Big Data technique described in this study. Large-scale industrial environments with huge data volumes, as well as those prone to unanticipated breakdowns, provide unique challenges for implementing Big Data technology. Because of this, our work utilises systems that are fast, scalable, and fault-tolerant for data collection and processing purposes. The performance of industrial equipment and any detected abnormalities will be shown on a dashboard. The results show that the application surpasses the monitoring system's current needs, since data processing remains stable for the current data volume.

References

Deka, S. A., Lee, D., & Tomlin, C. J. (2021). Towards Cyber–Physical Systems Robust to Communication Delays: A Differential Game Approach. *IEEE Control Systems Letters*, *6*, 2042-2047.

Fan, J., Huang, J., & Zhao, X. (2021). Improved interval estimation method for cyber-physical systems under stealthy deception attacks. *IEEE Transactions on Signal and Information Processing over Networks*, *8*, 1-11.

Cheng, D., Shang, J., & Chen, T. (2021). Finite-Horizon Strictly Stealthy Deterministic Attacks on Cyber-Physical Systems. *IEEE Control Systems Letters*, *6*, 1640-1645.

Zhang, J., Pan, L., Han, Q. L., Chen, C., Wen, S., & Xiang, Y. (2021). Deep learning based attack detection for cyber-physical system cybersecurity: A survey. *IEEE/CAA Journal of Automatica Sinica*, *9*(3), 377-391.

Tao, F., & Qi, Q. (2017). New IT driven service-oriented smart manufacturing: framework and characteristics. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, *49*(1), 81-91.

Colombo, A. W., Karnouskos, S., & Bangemann, T. (2014). Towards the next generation of industrial cyber-physical systems. In *Industrial cloud-based cyber-physical systems* (pp. 1-22). Springer, Cham.

Yue, X., Cai, H., Yan, H., Zou, C., & Zhou, K. (2015). Cloud-assisted industrial cyber-physical systems: An insight. *Microprocessors and Microsystems*, *39*(8), 1262-1270.

Lu, Y., & Xu, X. (2019). Cloud-based manufacturing equipment and big data analytics to enable on-demand manufacturing services. *Robotics and Computer-Integrated Manufacturing*, *57*, 92-102.

Atat, R., Liu, L., Wu, J., Li, G., Ye, C., & Yang, Y. (2018). Big data meet cyber-physical systems: A panoramic survey. *IEEE Access*, *6*, 73603-73636.

Faiz, R. B., & Edirisinghe, E. A. (2009). Decision making for predictive maintenance in asset information management. *Interdisciplinary Journal of Information, Knowledge, and Management*, *4*, 23.

Xun, Y., Zhang, J., Qin, X., & Zhao, X. (2016). FiDoop-DP: Data partitioning in frequent itemset mining on hadoop clusters. *IEEE Transactions on parallel and distributed systems*, *28*(1), 101-114.

Cheng, Y., Zhang, Y., Ji, P., Xu, W., Zhou, Z., & Tao, F. (2018). Cyber-physical integration for moving digital factories forward towards smart manufacturing: a survey. *The International Journal of Advanced Manufacturing Technology*, *97*(1), 1209-1221.

Baheti, R., & Gill, H. (2011). Cyber-physical systems. *The impact of control technology*, *12*(1), 161-166.

Tarraf, D. C. (2013). Control of cyber-physical systems. *Proc. of Lecture Notes in Control and Information Sciences*, *449*.

Monostori, L., Kádár, B., Bauernhansl, T., Kondoh, S., Kumara, S., Reinhart, G., ... & Ueda, K. (2016). Cyber-physical systems in manufacturing. *Cirp Annals*, *65*(2), 621-641.

Leitão, P., Colombo, A. W., & Karnouskos, S. (2016). Industrial automation based on cyber-physical systems technologies: Prototype implementations and challenges. *Computers in industry*, *81*, 11-25.

Harrison, R., Vera, D., & Ahmad, B. (2016). Engineering methods and tools for cyber–physical automation systems. *Proceedings of the IEEE*, *104*(5), 973-985.

Gerostathopoulos, I., Bures, T., Hnetynka, P., Keznikl, J., Kit, M., Plasil, F., & Plouzeau, N. (2016). Self-adaptation in software-intensive cyber–physical systems: From system goals to architecture configurations. *Journal of Systems and Software*, *122*, 378-397.

Yue, T., Ali, S., & Selic, B. (2015, July). Cyber-physical system product line engineering: comprehensive domain analysis and experience report. In *Proceedings of the 19th International Conference on Software Product Line* (pp. 338-347).

Eidson, J. C., Lee, E. A., Matic, S., Seshia, S. A., & Zou, J. (2011). Distributed real-time software for cyber–physical systems. *Proceedings of the IEEE*, *100*(1), 45-59.

Narayanan, A. N., Ak, R., Lee, Y. T., Ghosh, R., & Rachuri, S. (2017). Summary of the symposium on data analytics for advanced manufacturing.

Ly, L. T., Maggi, F. M., Montali, M., Rinderle-Ma, S., & Van Der Aalst, W. M. (2015). Compliance monitoring in business processes: Functionalities, application, and tool-support. *Information systems*, *54*, 209-234.

Bu, Y., Howe, B., Balazinska, M., & Ernst, M. D. (2010). HaLoop: Efficient iterative data processing on large clusters. *Proceedings of the VLDB Endowment*, *3*(1-2), 285-296.

Wilber, L. (2012). A Practical Guide to Big Data: Opportunities, Challenges, Tools. *Dassault Systems White Papers*, 4-36.

Chen, Z., Zhang, X., & He, K. (2017, September). Research on the technical architecture for building CPS and its application on a mobile phone factory. In *2017 5th International Conference on Enterprise Systems (ES)* (pp. 76-84). IEEE.

Lee, J., Bagheri, B., & Kao, H. A. (2015). A cyber-physical systems architecture for industry 4.0-based manufacturing systems. *Manufacturing letters*, *3*, 18-23.

Niggemann, O., Biswas, G., Kinnebrew, J. S., Khorasgani, H., Volgmann, S., & Bunte, A. (2015, August). Data-Driven Monitoring of Cyber-Physical Systems Leveraging on Big Data and the Internet-of-Things for Diagnosis and Control. In *DX* (pp. 185-192).