IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 19, 2021, accepted May 10, 2021, date of publication June 7, 2021, date of current version July 9, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3087109

Cyber Threat Predictive Analytics for Improving Cyber Supply Chain Security

ABEL YEBOAH-OFORI[1], SHAREEFUL ISLAM[2], SIN WEE LEE[2], ZIA USH SHAMSZAMAN[3], (Senior Member, IEEE), KHAN MUHAMMAD[4], (Member, IEEE), METEB ALTAF[5], AND MABROOK S. AL-RAKHAMI[6], (Member, IEEE)

[1]Department of Computer Science and Engineering, University of West London, Ealing London W5 5RF, U.K.
[2]School of Architecture Computing and Engineering (ACE), University of East London, London E16 2RD, U.K.
[3]Department of Computing and Games, Teesside University, Middlesbrough TS1 3BX, U.K.
[4]Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Department of Software, Sejong University, Seoul 143-747, South Korea
[5]Advanced Manufacturing and Industry 4.0 Center, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia
[6]Research Chair of Pervasive and Mobile Computing, Information Systems Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: Mabrook S. Al-Rakhami (malrakhami@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research at King Saud University through the Vice Deanship of Scientific Research Chairs: Chair of Pervasive and Mobile Computing.

ABSTRACT Cyber Supply Chain (CSC) system is complex which involves different sub-systems performing various tasks. Security in supply chain is challenging due to the inherent vulnerabilities and threats from any part of the system which can be exploited at any point within the supply chain. This can cause a severe disruption on the overall business continuity. Therefore, it is paramount important to understand and predicate the threats so that organization can undertake necessary control measures for the supply chain security. Cyber Threat Intelligence (CTI) provides an intelligence analysis to discover unknown to known threats using various properties including threat actor skill and motivation, Tactics, Techniques, and Procedure (TT and P), and Indicator of Compromise (IoC). This paper aims to analyse and predicate threats to improve cyber supply chain security. We have applied Cyber Threat Intelligence (CTI) with Machine Learning (ML) techniques to analyse and predict the threats based on the CTI properties. That allows to identify the inherent CSC vulnerabilities so that appropriate control actions can be undertaken for the overall cybersecurity improvement. To demonstrate the applicability of our approach, CTI data is gathered and a number of ML algorithms, i.e., Logistic Regression (LG), Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT), are used to develop predictive analytics using the Microsoft Malware Prediction dataset. The experiment considers attack and TTP as input parameters and vulnerabilities and Indicators of compromise (IoC) as output parameters. The results relating to the prediction reveal that Spyware/Ransomware and spear phishing are the most predictable threats in CSC. We have also recommended relevant controls to tackle these threats. We advocate using CTI data for the ML predicate model for the overall CSC cyber security improvement.

INDEX TERMS Cyber threat intelligence, machine learning, cyber supply chain, predictive analytic, cyber security, tactic techniques procedures.

I. INTRODUCTION

Cyber Supply Chain (CSC) security is critical for reliable service delivery and ensure overall business continuity of Smart CPS. CSC systems by its inherently is complex and vulnerabilities within CSC system environment can cascade from a source node to a number of target nodes of the overall cyber physical system (CPS). A recent NCSC report highlights a list of CSC attacks by exploiting vulnerabilities that exist within the systems [1]. Organizations outsource part of their business and data to the third-party service providers that could lead any potential threat. There are several examples for successful CSC attacks. For instance, Dragonfly, a Cyber Espionage group, is well known for targeting CSC organization [2], [3]. The Saudi Aramco power station attack halted its operation due to a massive cyberattack [1]. There are

The associate editor coordinating the review of this manuscript and approving it for publication was Po Yang.

existing works that consider CSC threats and risks but a lack of focus on threat intelligence properties for the overall cyber security improvement. Further, it is also essential to predict the cyberattack trends so that the organization can take the timely decision for its countermeasure. Predictive analytics not only provide an understanding of the TTPs, motives and intents of the threat actors but also assist situational awareness of current supply system vulnerabilities.

This paper aims to improve the cybersecurity of CSC by specifically focusing on integrating Cyber Threat Intelligence (CTI) and Machine Learning (ML) techniques to predicate cyberattack patterns on CSC systems and recommend suitable controls to tackle the attacks. The novelty of our work is threefold:

- Firstly, we consider Cyber Threat Intelligence(CTI) for systematic gathering and analysis of information about the threat actor and cyber-attack by using various concepts such as threat actor skill, motivation, IoC, TTP and incidents. The reason for considering CTI is that it provides evidence-based knowledge relating to the known attacks. This information is further used to discover unknown attacks so that threats can be well understood and mitigated. CTI provides intelligence information with the aim of preventing attacks as well as shorten time to discover new attacks.
- Secondly, we applied ML techniques and classification algorithms and mapped with the CTI properties to predict the attacks. We use several classification algorithms such as Logistic Regression (LG), Support Vector Machine (SVM), Random Forest (RF) and Decision Tree (DT) for this purpose. We follow CTI properties such as Indicator of Compromise (IoC) and Tactics, Techniques and Procedure (TTP) for the attack predication.
- Finally, we consider widely used cyberattack dataset to predict the potential attacks [6]. The predication focuses on determining threats relating to Advance Persistent Threat (APT), command and control and industrial espionage which are relevant for CSC [7]–[9]. The result shows the integration of CTI and ML techniques can effectively be used to predict cyberattacks and identification of CSC systems vulnerabilities. Furthermore, our prediction reveals a total accuracy of 85% for the TPR and FPR. The results also indicate that LG and SVM produced the highest accuracy in terms of threat predication.

The rest of the paper is organised as follows: Section 2 presents an overview of related works including CSC security, cyber threat intelligence and Machine Learning for CSC. Section 3 provides the concepts necessary for the proposed approach and the meta model. Section 4 provides an overview of the proposed approach including the integration of CTI and ML. Section 5 presents the underlying process for the threat analysis and predication. Section 6 implements the process for the threat predication using the widely used Microsoft malware datasets. Section 7 discusses the results and compares the work with the existing works in the literature. Finally, Section 8 provides conclusion and future direction of the work.

II. RELATED WORK

There exists several widely used CTI and ML models in cyber security domain. This section presents the existing works that are relevant with our work.

A. CYBER SUPPLY CHAIN(CSC) SECURITY

The CSC security provides a secure integrated platform for the inbound and outbound supply chains systems with third party service provider including suppliers, and distributors to achieve the organizational goal [10]. Cybersecurity from supply chain context involves various secure outsourcing of products and information between third party vendors, and suppliers [11]. This outsourcing includes the integration of operational technologies (OT) and Information technologies (IT) running on Cyber Physical Systems (CPS) infrastructures. However, there are threats, risks and vulnerabilities that are inherent in such systems that could be exploited by threat actors on the operational technologies and information technologies of the supply inbound and outbound chains systems. The outbound chain attacks include data manipulations, information tampering, redirecting product delivery channels, and data theft. The IT risks include those attacks on the cyber physical and cyber digital system components such as distributed denial of service (DDoS) attacks, IP address spoofing, and Software errors [12]. Regarding CSC security, NIST SP800 [13] proposed a 4 tier framework approach for improving critical infrastructure cybersecurity that incorporates the cyber supply chain risk management framework into it as one of its core components. Tier 1 considers the organizations CSC risk requirement strategy. Tier 2 considers the supply chain associated risk identifications including products and services in the supply inbound and outbound chains. Tier 3 implementation considers the risk assessments, threats analyses, associated impacts and determine the baseline requirements for governance structure. Tier 4 consider real-time or near-time information to understand supply chain risk associated with each product and service. However, the approach and tiers considered risks management but did not emphasize on ML and threat prediction for future trends in the CSC domain. Additionally, [14] proposed a supply chain attack framework and attack patterns that structured and codifies supply chain attacks. The goal of the framework was to provide a comprehensive view of supply chain attacks of malicious insertion across the full acquisition lifecycle to determine the associated threat and vulnerability information.

B. CYBER THREAT INTELLIGENCE (CTI)

Cyber threat intelligence (CTI) gatherings and analysis have become one of the relevant actionable intelligences used to understand both known and unknown threats [4]. The impact of cyberattacks and emerging threats on CSC systems and

its devastating effects on business process, data, Intellectual Property, delivery channel, and cost of recovery has increased the surge for CTI approach. The CTI process includes identification, threat analysis and information disseminating to stakeholders. Considering CTI for cybersecurity, ENISA in [4] explored the opportunities and limitations of current threat intelligence platforms by considering CTI implementation process and threat intelligence programs (TIP) from strategic, tactical and operational goals. The authors proposed a threat intelligence program model that collects, normalize, enrich, correlate, analyse and disseminate threat related information to stakeholders. The strategic CTI goals consider factors that support executive decision makings, tactical goals consider the CTI process and TIP programs that identifying intelligence gap and prioritizing them for risk reduction. The operational goals provide a process that provides an understanding of the threat actors motives, modes of operation, intents, and TTPs and capabilities. However, the processes do not incorporate ML threat predictions. Additionally, [15] proposes a threat intelligence-driven security model that considers six CTI phases and processes lifecycle required to identify intelligence goals. The CTI phases include direction, collection, process, analysis, dissemination, and feedback. The author incorporated internal sources such as network traffic, logs, scans; external sources such as vulnerability database, threat feeds; and human sources such as the dark web and social media into the model for the threat intelligence modelling. The threat intelligence driven security model emphasizes on using network traffics, logs and scans and not ML algorithms for the prediction. Further, [16] develop cyber threat Intelligence metrics that consider assets, requirement business operations, adversary, and consumer intelligence places emphases on value and organizational benefits. The author's approach considers four key stages in the threat intelligence process including intelligence requirements, information collection, analyses, dissemination, and intelligence usage. However, the approach does not consider machine learning for predicting invisible attacks. Furthermore, [17] proposed a CTI model that operationalizes and analyses adversarial activities across the lifecycle of an organization business process to determine actions taken by the attacker. The author's approach was based on the organizational intelligence requirements, information gathering, analyses and disseminate to protect assets for strategic, tactical and operational understanding and situational awareness. However, the works emphasized more on attacker motive and intent and not on ML for the threat predictions. The CTI functional process is to collect metrics and trend analysis for the business risk assessment, prioritization, and decision support with less emphasis on ML for CSC security.

C. MACHINE LEARNING IN CSC SECURITY

There are several works that consider Machine Learning classifiers in various cybersecurity application domains such as spam filters, antivirus and IDS/IPS to predict cyberattack trends [18], [23], [24]. Considering ML for Security [11], proposed ML classification of HTTP attacks using a decision tree algorithm to learn a dataset for performance accuracies and automatically label a request as valid or attack. The authors developed a vector space model used commonly for information retrieval to build a classifier to automatically label the request as malicious in the URL. The approach achieved high precision and recall comparatively. However, the work did not focus on ML and threat prediction in the CSC environment. Further, [20] carried out the feasibility of a study on machine learning models for cloud security to test the models in diverse operation conditions cloud scenarios. The authors compared Logistic Regression, Decision Tree, Naïve Bayes, and SVM classification algorithms techniques to learn a dataset for performance accuracies. The algorithms represent supervised schemes and are used in network security. The result shows an accuracy of 97% in anomalous packet detections. However, the work did consider CSC security from threat prediction in the supply chain environment. Furthermore, [21] surveyed data mining and ML methods for cybersecurity detection methods for cyber analytics in support of intrusion detection in cybersecurity applications. The authors used Artificial Neural Network, Association rules, Fuzzy Association rules and Bayesian Networks classifiers to learn the datasets and provided comparison criteria for the machine learning and data mining models to recognize the types of the attack (misuse) and for detection of an attack (intrusion). However, the techniques and methods used are not ML models and did not focus on ML and threat prediction in the CSC environment. Additionally, [22] review the cybersecurity dataset for ML algorithms used for analysing network traffic and anomaly detection. The author compared the machine learning techniques used for experiments, evaluation methods and baseline classifiers for comparison of the dataset. The results show significant flaws in some dataset during feature selection and are not relevant for modern intrusion detections datasets. However, the review did not stress on the current dataset we used from the Microsoft Malware Threat Prediction website for the prediction. Moreover, [23] explored the classification of logs using ML techniques on a decision tree algorithm to learn a dataset that models the correlation and normalization of security logs. The goal of the ML techniques is to evaluate if the algorithm can predict the performance of classification as an attack or not after a training phase. The dataset used contains anomalous and some identified attacks. The result shows that the DT algorithm was model on internet logs to develop a framework for the normalization and correlation of the classify with an accuracy of 80%. However, the classification model did not compare other classification algorithms such as SVM, LR and RF that are relevant for ML better performance accuracies and threat analysis.

Another initiative [24] explores the viability of using machine learning approaches to predict power systems disturbance and cyberattack discrimination classifiers and focuses specifically on detecting cyberattacks where deception is the core tenet of the event [24]–[30]. The authors in [24]

evaluated the classification performances on, NNge, OneR, SVM, RF, JRpper and Adaboost algorithms to learn the dataset and focused specifically on detecting cyber attacks where deception is the core tenet of the event. For example, in [25], the authors proposed a SCADA power system cyber-attack detection approach by combining a correlation-based feature selection (CFS) method and K-Nearest-Neighbour (KNN) instance-based learning (IBL) algorithm. The combination was useful to reduce the extremely large number of features and to maximize cyberattack detection accuracy with minimum detection time cost. In [26], an ensemble-learning model for detecting the cyberattacks of SCADA-based IIoT platform is proposed. The model was based on the combination of a random subspace (RS) learning method with random tree (RT). The authors in [29] proposed a deep-learning, feature-extraction-based semi-supervised model for cyberattack protection in the trust boundary of IIoT networks. The proposed approach was adaptive to learn unknown attack. However, the works did not consider CSC attacks from supplier inbound and outbound chains.

Regarding ML predictive analytics on various datasets, [28] predicted cybersecurity incidents using ML algorithms to distinguish between the different types of models. The authors used text mining methods such as n-gram, bag-of-words and ML techniques to learn dataset on Naive Bayes and SVM algorithms for classification performance. The experiment was to predict classification accuracies of malware incidents response and actions. The approach did not consider CTI and ML in the CSC system environment. Further, [29] proposed a risk teller system that analyses binary file appearance logs of a machine to predict which machines are at risk of experiencing malware infection in advance. The authors used a random forest algorithm and semi-quantitative methods to build a risk prediction model that creates a profile to capture usage patterns. The results associate each level of risk to a machine infection incident with 95% true positive precision. Besides,[30] characterize the extent to which cybersecurity incidents can be predicted based on externally observable properties of an organization's network. The authors used Verizon's annual data breach investigation report to forecast if an organization may suffer cybersecurity incidents in future. A random forest classifier was used against over 1000 incident reports taken from various datasets. The predictive result achieved an overall accuracy of 90% true positives. However, the work did not provide any inference and map the prediction to existing attacks. All these works above are important and contributed towards the improvement of cyber security by using various ML techniques. However, there is a lack of focus on the overall CSC security context. A limited works emphasize on threat intelligence data for the attack predication. For instance, due to the invisibility nature of cyberattacks, an attack on the CSC system network node has the potential to cascade to other nodes on the supply chain system. Therefore, it is necessary to use ML analytics to predict cyberattacks, threats and the underlying vulnerabilities. Additionally, there is a need to understand an organisational context for the threat analysis. CTI can effectively support to achieve that goal. This work contributes towards this direction. We have integrated CTI for threat gathering and analysis with the ML for the threat prediction so that organizations can determine the suitable control measure for the overall CSC security improvement.

III. FRAMING CONCEPTS

This section presents the conceptual view of the proposed approach by combining concepts from both CTI and CSC.

A. CSC THREAT MODELLING CONCEPTS

This section considers the concepts that are necessary to determine CSC vulnerabilities, goals, requirements, attacks the cyber supply inbound and outbound chains security and the CTI domain [2]. Threat modelling provides a systematic approach to identify and address the possible threats based on a specific context. It provides an understanding of threat actor who can attack the system and possible assets which can be compromised. The proposed approach considers a list of concepts that aid understand the threts and possible mitigation. The concepts provide a view of the relationships between organizational and security goal, requirements, threat actors, attacks, vulnerability, TTPs and indicators of compromise for understanding of the threat. An overview of the concepts is given below:

Goal: A goal represents the strategic aim of an organization. Properties for the goal include the organizational goal, the tangible assets required such as infrastructures to achieve the goal and intangible asset such as credit card information, health record, and other sensitive data for the security goal. The organizational goal is the process, product or service that is carried out. The assets are tangible and intangible assets including the network infrastructures. The security goal is the mechanism, configuration, and control put in place to achieve the goal.

Actor consists of perpetrators, system users, the systems, the third-party vendors, and companies whose services and networks systems are attached to the main organization's supply chain system. The threat actors are those consist of users, agents, cybercriminals, and other systems that aims at compromising the CSC systems and the security goal [8]. The threat actor could be an internal or external attacker. The CSC system includes the various integrations of network nodes that make up the supplier chain system. The third-party vendors include the organization on the supplier inbound and outbound chains that could be attacked, manipulated, or compromised.

Inbound and Outbound Supply Chain: In a CSC environment, the network nodes and communication channels are those that integrate with the inbound and outbound supply chains systems. These are vendors, SMEs, suppliers, and distributors that are on the supply chain. The inbound suppliers are those with external remote access to the CSC system. The outbound chains are those that the organization distributes including individuals, institutions,

and vendors. The organization can experience attacks on the supply inbound and outbound chain that supports the application processes [8]. The threat actor could initial injection attacks or insert a redirect script into the vendor's website and breach the software developed by the manufacturer that is used by the organization's internal employers to distribute services to vendors and individuals. The goal of the attack could be to manipulate, alter or divert products and services after gaining access into the system.

Vulnerabilities: CSC vulnerabilities are the loopholes and configuration flaws that exist on the supply chain system and network nodes that could be exploited by an attack, threat actor or a threat agent. These network vulnerabilities [36] are those that exist on the supply inbound and outbound chains including the network nodes, switches, IP addresses, and firewalls. The vulnerable spots on the CSC system could be identified from various sources including the software, the network, website, the user, processes, the application, and configuration or the third-party vendor. Properties include asset type, source, node, effect and criticality.

Attack: An attack is any deliberate action or assault on the supply chain system with the intent to penetrate a system, to be able to gain access then manipulate and compromise processes, procedures, and delivery channels of electronic products, the information flows, and services [2]. Properties include the type of attack, pattern, prerequisites, and vectors. We consider attack inputs and outputs parameters for our study and the attack concepts for our prediction. Inputs of attack include the tools, capabilities, vectors and knowledge of the vulnerabilities of the domain to exploit. Outputs of the attacks are the patterns, access gained by the threat actor, the methods deployed, TTPs, the loopholes exploited, and the extent of malware propagation and cascading effects. This includes those attacks on cyber physical and cyber digital systems such as hardware, network, IP addresses, and software. The OT and IT delivery mechanisms could be manipulated before the product gets to the consumer [8].

Tactics, Techniques and Procedures (TTPs) consist of the specific adversary behaviour exhibited in an attack [14]. It leverages on resources such as tools, infrastructures, capabilities and personnel. It provides information on the victim's target (who, what or where), that are relevant to exploit targets being targeted, intended effects, kill chain phases, handling guidance and resources of the TTP information [8], [9]. Threats actors' mode of operation is to commit attacks such as Hijacking, social engineering, and footprints, privilege escalation, and reconnaissance penetrate a supply chain.

CSC Requirement: CSC requirements are the constraints and security expectations for the system required to support CSC stakeholders and business needs. The data gathered from stakeholders inform business processes, system infrastructures, internal and external user expectations required for the supply chain system developments and operations [2]. The requirements process and constraints that are generated during the requirements engineering phase forms the basis for the system constraints and statements that support the user and system requirements used to achieve the organizational goal. The requirements consist of attributes such as user categories, stakeholders, description, user ID, acceptance criteria, time constraints, owners and sources. The requirements concepts include properties such as organizational requirements, business requirements, system, user, and operational requirements. The organizational requirements describe the organizational high-level objectives that must be performed to achieve the organizational goal. The business requirements explain the requirement specifications and the properties include customer needs and expectations that must be integrated to meet the system requirements. Systems requirements demand specific properties of the application, architecture and the technical requirements need to be able to describe the features and how the system must function. These system requirements properties include the constraints, assumptions and acceptance criteria and the external entities that will be interacting with the system. They include supply chain systems processes and constraints that are generated during the requirements engineering phase that forms the basis for the system.

Indicators: Indicators are parameters that express an attack of this type, whether it is imminent, in progress or has occurred [32]. Properties required to determine the indicators of compromise includes incident type, source, date & time, impact, motive and intents. The properties are used to determine threat activities, adversary behaviours, TTPs, risky events, or state of the incident to determine what could serve as an indicator of compromise. CSC attack incidents and course of actions provide intelligence about the nature of cyberattack indicators and TTPs that can be deployed on the supply chain especially from the third-party vendor's perspective. Indicators convey specific observable patterns combined with contextual information intended to represent artefacts and or behaviours of interest within a cybersecurity context.

Cyber incident report: Cybersecurity incident is defined as a breach of system security to affect its integrity or availability. It includes unauthorized access or attempted to access a system or causing a disruptive event to essential services. Cybersecurity incident reporting platform provides individuals and organizations with a system to reports cyber incidents they have experienced unexpectedly or any unusual network issues, or suspected fraud or cybercrime activities [31]. Properties for cyber incident reporting include attack type, date and time of the incident, source of the attack, cause of an attack, duration, impact on service, impact on staff and public safety Cyber incident report system is required for cyber threat analysis and to determine the threat level and categorizing. It is used to predict cyberattacks and generate intelligence require to mitigate cyberattacks and for threat information sharing.

Threat information sharing: Threat information sharing is used to provide information necessary to assist an organization in identifying, assessing, monitoring, and responding to cyber threats [32]. Cyber threat information includes

indicators of compromise, tactics, techniques, and procedures used by threat actors, security alerts and threat intelligence reports. It provides findings from the analysis of cyber incidents and suggests actions to take to prevent cyber-attacks, detect, protect, contain, and mitigate cyber incidents. Properties for cyber threat information sharing include information-sharing goals, information sources, scope, sharing community and support. Some rules govern and protect information sharing, such as information sensitivity and privacy, sharing designations, and tracking procedures [32]. It provides a basis for an organization to leverage their combined knowledge, information, experience, and competencies to gain intelligence and understanding of potential threats for remediation and controls.

Controls: Controls are security mechanisms that are put in place to secure organizational business operations and processes. They are security strategies and measures formulated and implemented to ensure that the organizational goal and objectives are achieved [2], [13]. These controls include directive, detective, preventive, corrective and recovery. Directive controls are more strategic and relevant with the specific supplier inbound and outbound chain requirements. These are intended to align organizational and security goals with that of supplier and third-party vendors on the supply chain and provide guidelines for system usage and processes. Preventive controls are policies that are put in place for the technical and physical infrastructures protection. These are derived from standard measures intended to preclude actions violating policy or increasing third party risks to the supply chain system resources. Detective Controls use supply chain attack indicators to identify practices, processes, and tools that identify and possibly react to security violations. These include Firewall, IDS, IPS and the various configurations required for the supply chain systems. Corrective controls involve physical, administrative, and technical measures. Recovery controls includes backup plans, regular updates and contingency planning to ensure integrity or availability of the CSC in the event of an incident. Once an incident occurs on the CSC system that results in the compromise of integrity or availability, the implementation of recovery controls is necessary to restore the system or operation to a normal operating state. These include countermeasures, backups, segmentation, and an incidence response strategy.

The meta-model in Figure 1 explains relationships among the concepts. The organizational goal is determined by the product and services that are produced. The security goal is to ensure that the supply chain systems that support these products and services are secured. CSC organization needs a list of requirements to satisfy for achieve its goals. The TTP as a CTI properties exploits both inbound and outbound vulnerabilities for a successful attack. Cyber incident report provides a detailed about the incident including vulnerability, indicator and incident time frame. This report needs to share among the CSC stakeholders. There are controls which are required to tackle the threats.

IV. THE PROPOSED APPROACH

This section discusses the proposed approach that aims to improve the CSC security. It includes an integration of CTI and ML and a systematic process (presented in the Section 5). Additionally, the underlying concepts of the proposed approach such as actor, goal, TTP, vulnerability, incident, and controls, is also mentioned in Section 3. The approach considers both inbound and outbound chains for the vulnerability so that CSC organisation can focus on the possible system flaws. The approach adopts the CTI process to gather and analyse the threat data and ML techniques to predicate the threat. ML techniques are used on classification algorithms to learn a dataset for performance accuracies and predictive analytics. The rationale for integrating CTI and ML for threat prediction is that the CTI lifecycle process supports input parameters for detecting known attacks whereas ML provides output parameters for predicting known and unknown attacks for future trends.

A. INTEGRATION OF CTI AND ML

The approach combines CTI processes with ML techniques for cyber threat predictive analytics. The goal is to detect vulnerabilities and indicators of compromise on CSC network system nodes using known attacks to predict unknown attacks. We apply the CTI techniques to gather threats (Known attacks) and ML techniques to learn the dataset to predicate cyber threats (unknown attacks) on CSC systems. The inputs are the attacks and TTP that are deployed by threat actors to compromise a system. The attack feature uses properties such as attack type, pattern, attack vectors, and prerequisites to determine the nature of the attack that was deployed. The TTP consists of attack patterns and attack vectors deployed by the threat actor. The TTP parameter includes the capabilities of the threat actor and threat indicators. The threat actor feature uses properties such as user, system and third-party vendors to determine the vulnerable spots and type of tools used for the attack to determine the attack pattern. Tools are the attack weapons or software codes used by the threat actor for reconnaissance and to initiate an attack. For instance, the threat actor could use Nmap tool for scanning a network, Kali Linux tool for penetration and, Metasploit tool for exploiting loopholes in a network. The output parameters are the vulnerabilities and indicators of compromise that are used as threat intelligence. The capability of the threat actor could be determined by the ability to penetrate a system and course Advance Persistent threat (APT) attack and take command and control C&C) the extent of propagation is used to determine the indicators. Finally, we consider various controls such as directive, preventive, detective corrective and recovery required to secure the CSC system.

The rationale for our predictive analytics approach is based on the premise that the cyberattacks phenomenon includes a lot of invincibility, and uncertainties and the makes the threat landscape unpredictable. Similarly, due to the changing organizational requirements, various integrations, varying business processes and the various delivery mechanisms,

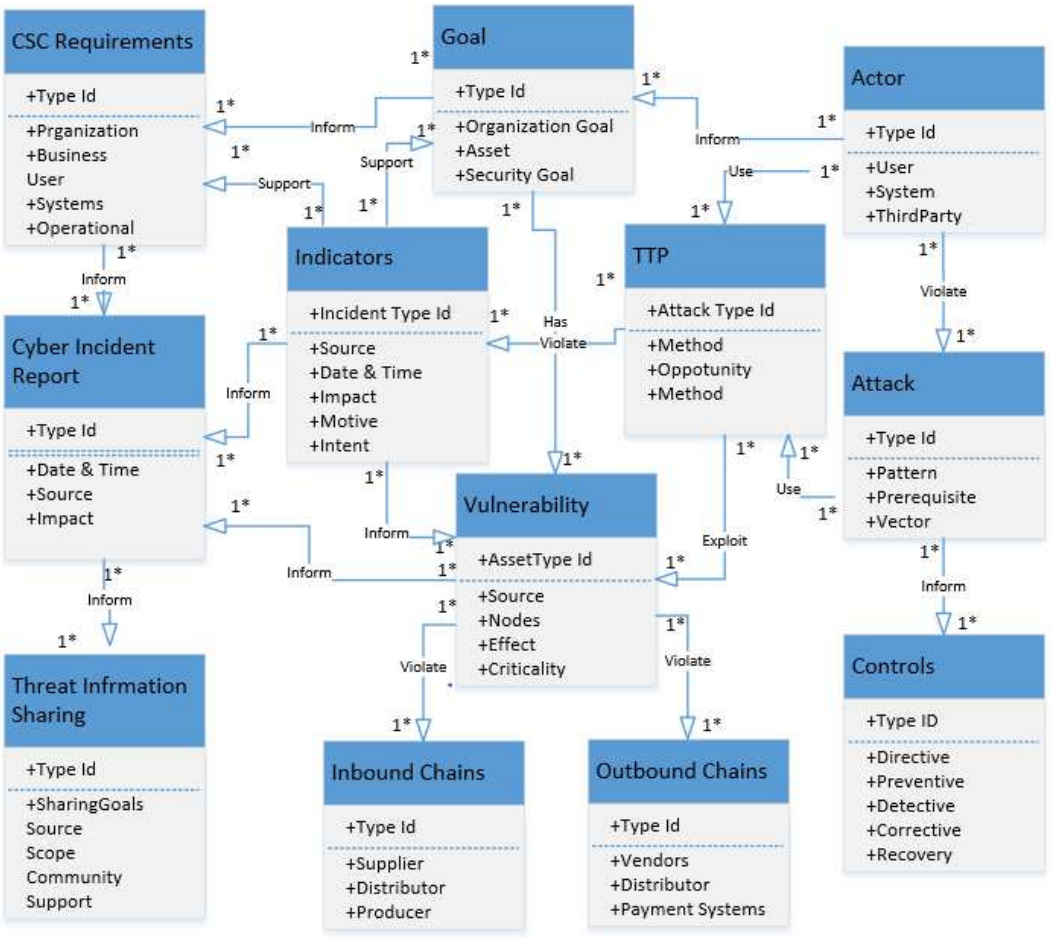

FIGURE 1. Meta-model for the proposed conceptual view of CSC system security.

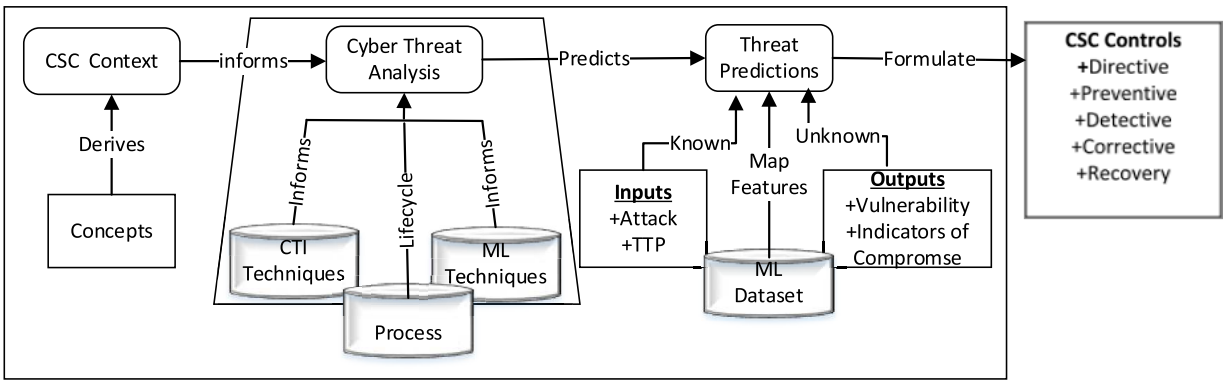

FIGURE 2. Applying CTI and ML for threat intelligence and predictive analytics.

predicting cyberattacks in the CSC organization context has been challenging. To achieve that, first, the proposed approach considers relevant related works and the meta-model concepts to model the CSC attacks and CTI phases. For instance, we identify supply inbound and outbound chain attack indicators and integrate them into CTI phases. Further, the concepts are analysed using the CTI process lifecycle and ML techniques to learn the dataset for our prediction. Furthermore, we use the input and output parameters as indicators for our threat prediction. Finally, the threat prediction results are evaluated to provide informed intelligence regarding the various attacks and future threats that are unknown for appropriate control mechanisms. Figure 2 indicates the proposed approach.

V. THREAT ANALYSIS AND PREDICTION PROCESS

This section discusses the overall process for the CSC threat analysis, prediction, and control in line with the proposed

approach in Section 3. The process includes four sequential phases. It follows a methodical approach and a causal process for each phase to determine strategy, threat analysis, threat prediction, and controls. Each phase includes steps and activities required to achieve the purpose of the phases as shown in Figure 3. The activities include identifying the organization's CSC and security strategy, ML classifications, infrastructures, attack context, input and output parameters for our prediction. The activities for the threat analysis phase include the identification and gathering of threat information, risk assessment and analysis to determine the threat actor, threat profile, TTP and IoC. The activities for the threat prediction phase consider the input parameters for the ML algorithms, predict threats and for performance evaluation by using ML techniques to learn datasets. The control activities include identifying required controls for the CSC systems including internal and external audits to formulate security policies and control mechanisms. We expound on the phases and process further by following the process flow as shown in Figure 3.

A. PHASE 1: DETERMINE STRATEGY

CSC security strategy combines CTI and cybersecurity risk strategy including mechanisms, resources and plans to determine how security goals and controls will be formulated, implemented, and achieved in line with organization goal and objectives. It includes identifying, analysing, reviewing and evaluating organizational assets including infrastructures, resources and implementation procedures. CSC security strategy combines, CTI and cybersecurity risk assessment strategy to gather intelligence and formulate policies. Strategic, tactical and operational management roles and responsibilities are recursive and support each other to ensure security goals are achieved. Strategic management uses intelligence decision to support plans that determine security goals and assign responsibility including executive authorization of blueprints and budget allocation. Tactical management decision regarding the execution of strategic management blueprints including security requirements capturing, third party audit, configuration management plans, uses indicators of compromise to determine controls and validations. The operational level managers ensure the day-to-day implementation of the security goals including monitoring, determining TTPs and escalating threat alerts for remediation and controls. CTI Strategy provides management evidence-based knowledge gathered about threats actors, attacks, patterns, vectors, vulnerabilities, TTPs, motives, intents and capabilities of the adversary. Risk Assessment Strategy considers the organizational goal and assets and develops an overall CSC risk strategy that determines the policies required to guide the organizational business processes. It includes risk assessment, CSC requirements capturing and business function. The risk strategy also considered implementation strategies and procurement policies for OT and IT acquisitions and integrations of assets.

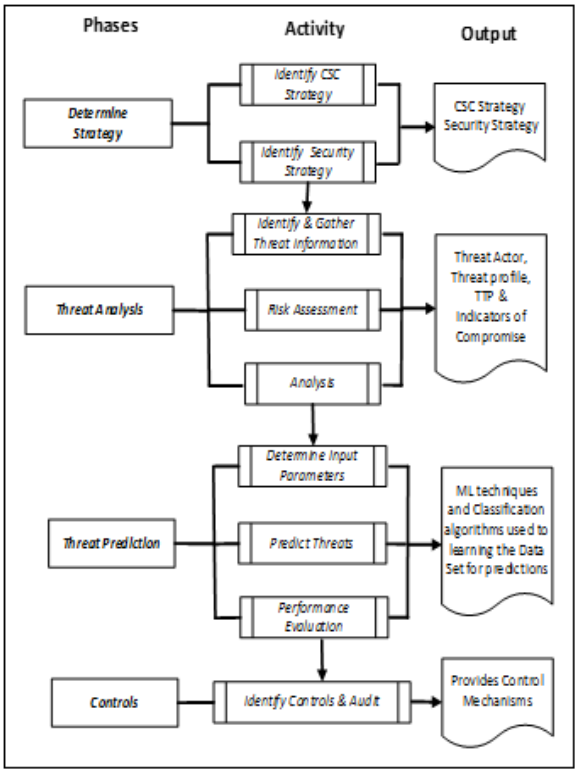

FIGURE 3. Predictive analytics process.

B. PHASE 2: THREAT ANALYSIS

This threat analysis phase follows the CTI techniques to determine and analyse the threats of the CSC context. It requires the CSC strategy information for his purpose and includes three activities.

Activity 1: Identify and Gather Information

This step identifies all vulnerable spots on the supply inbound and outbound chains on the meta-model that is used as indicators for an attack. For instance, in case of a malware attack, this activity looks for the relevant information such as the source of the attack, the tools, patterns and the attack vectors from the analysis of the malware attack that used as our indicator. To determine the indicators of an attack, we use threat activities, adversary behaviours, risky events, or state of the incident to determine what could serve as an indicator. The indicators may be used to identify any inherent vulnerabilities that could be exploited by a threat actor. If necessary, the activity carrying out penetration testing, vulnerability assessment test and threat propagation exercises to determine the supply inbound and outbound chains on the OT and IT by following the below stages [2].

Activity 2: Identify and Gather Information

This step identifies all vulnerable spots on the supply inbound and outbound chains on the meta-model that is used as indicators for an attack. For instance, in case of a malware attack, this activity looks for the relevant information such as the source of the attack, the tools, patterns and the attack vectors from the analysis of the malware attack that used as our indicator. To determine the indicators of an attack,

we use threat activities, adversary behaviours, risky events, or state of the incident to determine what could serve as an indicator. The indicators may be used to identify any inherent vulnerabilities that could be exploited by a threat actor. If necessary, the activity carrying out penetration testing, vulnerability assessment test and threat propagation exercises to determine the supply inbound and outbound chains on the OT and IT by following the below stages [2].

- Stage 1. Reconnaissance: The threat actor uses APT methods to gather intelligence and searches the organization's websites to gather footprints and identify vulnerable spots on the network nodes.
- Stage 2. Experiment: The threat actor uses penetration testing and vulnerability assessment methods various attack patterns, TTP methods, and tools to explore vulnerable spots. The attacks include spear phishing malware or Remote Access Trojan.
- Stage 3. Exploit: the threat actor initiates attack to gain access to the system and other resources of the system. The attack could manipulate, alter and redirect deliveries or initiate and propagate malware.
- Stage 4. Command and Control**:** The threat actor maintains a continuous presence on the system and can change his password to maintain a presence on the CSC using advanced persistent threat attack, remote access command to steal intellectual properties and cause cyber espionage attacks. Most organizations use automated password changing system that prompts users to change their password periodically and that could be exploited by the threat actor. The threat actor can change the password and obfuscate in a Command & Control environment [2].

Activity 3: Risk Assessments

The risk assessment activity includes the process to mitigate CSC risks by determining the probability and impact of CSC attacks and threats as well as the vulnerable spots that could be exploited within the cyber supply inbound and outbound chains and third-party organizations. It identifies all threats that may pose a risk on the system. Risk assesses the CSC security domain and analyse risks access spots that are capture captured. Develop mitigating techniques to control the risks by identifying risks posed by auditing the third-party organizations. Classify them based on their service provisions and levels of integration to the various supply chain network system.

Activity 4: Analysis

This activity focuses on analysis of the threats to determine the actual source of the attack, the type of attack, the attack pattern, the TTP and attack vectors. This will assist to assign the IoC required and what controls are needed. The threat analysis techniques include:

- Stage 1. Threat Activity: Determine the nature of attack, pattern and sources of penetration on the CSC.
- Stage 2: Threat Manipulation: Determines the nature of cybercrimes committed and the extent of the penetration to understand the capabilities, motives and intents of the attacker.
- Stage 3: Threat Impact: Determines the severity of the attack, malware propagation and the cascading effects on the supply chain. These determinants influence the risk factors and the degree of severity of the attacks.

C. PHASE 3: THREAT PREDICATION

The phase considers CSC system nodes that are vulnerable to cyberattacks by integrating CTI and ML to obtain attack predictions of known and unknown attacks using three sequential activities.

Activity 1: Determine Input Parameters

The input parameters mainly consider the attack and TTP to demonstrate how the attackers penetrate a system. In particular, threat actors' properties such as capability and attack vector, tools are used for the input parameters.

- Step 1: Feature Selection: This step includes different ML techniques to select the available features that exist in the data. These feature selection techniques include dimensionality reductions in large datasets for effective and reliable training, testing and prediction. The features we use for our prediction are malware, spyware, spear phishing and Rootkit attacks.
- Step 2: Choosing a Classifier and Performance Metrics: We classify the various algorithms such as LR, DT, SVM and RF in VM to determine (1) the different types of responses based on an attack and (2) different types of response give the TTP deployed. For our study, we use the binary classification as it supports AUC-ROC in distinguishing between the probabilities of the given classes. Further, its precisions can predict correct instances, provides a harmonic mean of precision and recall for the F-score. Determining the right performance metrics to evaluate the algorithms, influences the performance measures and how the algorithm are compared with others. Not using the right metrics could cause overfitting problems and impact on how we evaluate our predictions.

Activity 2: Predict Threats

This activity aims to predicate vulnerabilities and IoC as output feature. The vulnerabilities provide the organization intelligence about areas that are exploitable and the IoC provides the indicators of penetrations, cybercrimes compromises, APTs and C&Cs. Using the cyber threat analysis and the inputs features, we use ML techniques and dataset to predict the output features. The vulnerable spots include network nodes, firewalls, antivirus and anti-malware. The IoC includes the unknown attacks and the extent of cybercrime manipulations, alteration, deletions, exfiltration and redirections that the threat actor could deploy on the system. The stealthy nature of such attacks is so uncertain it cannot be determined on the face value. This includes gathering various attack probabilities and their propagation effects on the CSC using ML techniques to train and test dataset to learn and to gain accurate predictions. The process involves:

- Applying ML techniques to learn the data events from IDS/IPS and firewall logs to collect signatures, threat indicators and, antimalware logs from the various supply chain endpoints. The ML techniques consider LR, SVM, DT, RF and MV algorithms to determine the accuracies of our predictions.
- Determining false positives and false-negative rates.
- Analyse ML results, logs and alerts to understand the attack trends as identified in the initial process to gather intelligence as to what happened, how, why, when, who and where the attack is initiated from.

Activity 3: Performance Evaluation

The performance of the models will be evaluated based on the following values: True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Further, the FP and FN will be determined based on the elements of the confusion matrix. We follow the following steps for the performance evaluation.

Step 1: Using Confusion Metrics to Determine TP and FP Outcomes

A confusion matrix is a two-dimensional matrix that evaluates the performance of a classification model with respect to a specific test dataset. It basically compares the actual target values with those predicted by the machine learning model. It provides a better understanding of the values by calculating the data in the matrix and analyse them to determine any positive or negative classifications. Four outcomes are determined when classifying the instances of the dataset. These include True Positive (TP), True Negative (FP), False Positive (FP) and False Negative (FN) rates. For instance, in an event where an instance is positive, and the outcome is classified as positive, its TP else its FP. Where the instance is negative and the outcome is classified as negative, it is counted as TN, else it is FN [15]. We consider the following method to understand the confusion matrix. The accuracy of the confusion metric is the proportion of the total number of predictions that are considered as accurate. We use the following equation below to determine the TPR, TNR, FPR, FNR and the entropy.

$$AC = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

The recall or true positive rate (TPR) is the proportion of the total number of correct predictions. We consider the equation as:

$$TPR = \frac{TP}{FN + TP} \tag{2}$$

Finally, precision (P) is the proportion of the predicted positive cases that were determined as correct. Hence the formula:

$$P = \frac{TP}{FP + TP} \tag{3}$$

F-measure of F1 – Score (F) is used as the harmonic mean to determine the combinations of precision and recall. We use the formula as:

$$F = \frac{2(Precision\ X\ Recall)}{Precision + Recall} \tag{4}$$

Step 2: Determine Mean Absolute Error (MAE) and Mean Square Error (MSE)

MAE determines the sum of the absolute mean or normal curve of the difference vector between predicted and real values. Whereas MSE determines the mean or normal difference by taking the absolute value of the square root of the mean and convert the units back to the original unit of the output variable and provide a gross idea of the magnitude of the error. For us to predict real numbers or regressions, we used MAE and MSE. The activities include Import AUC-ROC Function, Import Mean Absolute Error, Import Mean Square Error, and Set Entropy Criterion. Entropy is a concept used in information theory to determine the measure of uncertainty about the source of data. It is a unique function that satisfies the four uncertainties axions in a confusion matrix and gives us the degree of disorganization in our data. In an event where a given set of data may contain random collections of unstructured data, and entropy formula is used to separate the positive and negative rates as follows:

$$\text{Entrophy(E)} = -a \log_2 a - b \log_2 b \tag{5}$$

where a = Proportion of positive examples and b = Proportion of negative examples. We use the formula to determine the results in our experiment. We ask the following question to derive the answer from the performance.

- TP = Did the model predicted correctly for the positive class as positive?
- TN = Did the model predicted correctly for the negative class as negative?
- FP = Did the model predicted incorrectly the negative class as positive?
- FN = Did the model predicted incorrectly the positive class as negative?

D. PHASE 4: CONTROL

This final phase aims to identify a list of controls that are to tackle the threat. The controls should ensure that the required security strategic and mechanism are put in place to mitigate the threats. This includes identifying security requirements, internal and external audit as well as threat monitoring and reporting. The process includes identification and review of existing controls, third-party audit and finally information sharing.

VI. IMPLEMENTAION

This section follows the implementation of the proposed approach to determine the applicability of our threat prediction. We only follow threat identification, prediction, and control phases for the implementation.

A. THREAT ANALYSIS

Threat analysis phase uses CTI approach to gather threat. We identify vulnerabilities on the network nodes, IP address,

IEDs and the threats that are linked to the organizational goal that provide us with threat indicators. This includes the TTP used by threat actors and their modes of operations. For our analysis, we adopt the attack concepts and the properties from the meta-model to determine the attack pattern and the TTP deployed on the CSC. The phase involves gathering sources of attacks, vulnerable spots, risks TTPs. Data are gathered from firewalls logs, collecting a signature, threat indicators and events from IDS/IPS, antimalware logs from the various endpoints.

B. THREAT PREDICATION

Further to the discussion in Section 4, threat prediction involves using ML techniques to learn dataset for threat predictions of known and unknown attacks. We follow the ML process for our threat prediction.

1) DESCRIPTION OF DATA

We have considered the widely used dataset from a Microsoft Malware website for the implementation [6]. The dataset is about malware attacks in the Microsoft endpoint system. The data was collected by Microsoft Windows Defender with over 40,000 entries, with 64 columns and each row represents different telemetry data entries. The data represents malware attacks identified on various endpoint nodes from different locations with machine identities, timestamps, organizational identifier and default browser identifiers designed to meet various business requirements. The rationale for using the dataset is that the dataset does not represent Microsoft customer's machine only as it has been sampled to include a much larger proportion of malware infection machines. Therefore, we used this dataset for our predictive analytics as CSC systems integrate various network infrastructures for the business process and interoperability.

The feature description includes MachineIdentifier that considers individual machine ID on the network, GeoNameIdentifier, provides IDs for the geographic region a machine is located in. DefaultBrowsersIdentifier, provides ID for the machine's default browsers. OrganizationIdentifier, provides ID for the organization the machine belongs in. Is protected, provides a calculated field derived from the Spynet Report's AV Products field. Processor considers the process architecture of the installed operating system. HasTpm, indicates true if the machine has TPM (Trusted Platform Module). Over, looks at the version of the current operating system. OsBuild, information indicating the build of the current operating system. Census_DeviceFamily AKA DeviceClass, indicates the type of device that an edition of the OS is intended for desktop and mobile. Firewall, this attribute is true (1) for Windows 8.1 and above if windows firewall is enabled, as reported by the service [6].

2) DATA PREPARATION

The activity involves uploading the data from a website APIs or an HTML file and selecting the data we need then save it as CSV file. We prepare the data by converting the average of the columns of the dataset. Furthermore, we loaded the data from a pre-prepared dataset by calling the categories of the machine learning identifier: The output generated 40,000 training datasets with 62 variables. Handling NaN (Not a Number) in training set by using a command that removes all the NaN in the training set into the dictionary and prints the output. Furthermore, we create a NaN dictionary to handle all the unwanted duplicate data. The output prints $62 - 8 = 54$. (8 columns removed).

3) FEATURE SELECTION

The main features are identified from the primary dataset that are relevant to our work. There were 62 features in the primary data and the focus is on the concepts of attacks, tools and vulnerabilities from our previous work. We characterized threat actor activities, including presumed intent and historically observed behaviour, for the purpose of ascertaining the current threats that could be exploited. Further, we identified eight vulnerable spots and their probability that the cyber attacker could exploit those spots namely the: Firewall, IDS/IPS, Vendors CSC system, Network, IP Addresses, Database, Software, and Websites.

4) BUILDING NEW FEATURES INTO THE DATASET

The features considered as input parameters for the predictions are the attack and TTP as discussed in Section 3.2. To achieve that, we determine the types of attack, tools, vectors, and capabilities for the input. we build the features in line with the existing dataset feature description in [6]. Further, features for predicting the attack inputs and outputs are identified by deriving new features that are in line with the existing datasets and features [6] in Table 2. These features and variables are related to the dataset for our work. Attack patterns are an abstract mechanism for describing how a type of observed attack is executed [32]. The output parameters are determined after our evaluation using the attack pattern, TTPs, vulnerabilities as indicators of compromise. Furthermore, the attack profiles for the ML prediction are built-in dataset. The main goal of our work is to be able to build attack profiles for our ML to predict which node is vulnerable and likely to be attacked. We may not be able to use exact features, but we consider characteristics that are correlated with them and are relevant to represent how the attacks are initiated and the vulnerabilities are exploited for our future prediction. Hence, many features that we analysed were chosen to represent the CTI and security awareness of the stakeholders.

5) CHOOSING AN OPTIMIZATION ALGORITHM FOR THE CLASSIFIERS

For us to choose the classifiers as discussed in Section 4.1.3. activity 1, step 2. we used a pipeline to connect the various classifications. We use the 10-Fold cross-validation to determine the parameter estimation. The 10-Fold cross-validation run and validate the parameter ten times on each algorithm as the values may change and may not generate the accurate

TABLE 1. Matrix to compute the accuracy, precision, recall and the F-score.

Number = 185	Predicted Yes	Predicted No
Actual Yes	TP = 180	FN = 20
Actual No	FP = 40	TN = 120

result when we run it only ones. For the test, we used 10-fold cross validation for more accurate predictive results. The GridsearchCV provides an exhaustive search over specified parameter values for an estimator. We combine all the four algorithms using Majority Voting (MV) algorithm in the classifiers to determine the mean score of the total results. Finally, we use ROC-AUC to distinguish between the accuracies of the binary classification for the predictions [32].

6) EVALUATING THE ACCURACY OF THE THREATS

We consider the following method to understand the confusion matrix as discussed in Section 5. The accuracy of the confusion metrics is the proportion of the total number of predictions that are considered as accurate. Using the equation in Section 5, we evaluate the accuracies (AC) of the metrics to answer the performance of the TP, TN, FP, FN rates in (V) as follows:

$$AC = \frac{180 + 120}{180 + 120 + 40 + 20} = 0.83 \tag{6}$$

Using the Table 3, and the algorithm, we answer the following question to derive the values for the performances. The False positive rate (FPR) determines the rate of negative cases that were incorrectly classified as positive.

• FP = Did the model predicted incorrectly the negative class as positive rates?

$$AC = \frac{40}{120 + 40} = 0.23 \tag{7}$$

The result indicates that FPR of 0.25 negative cases were incorrectly classified as positive. Whereas the true negative rate (TNR) is defined as the number of negative cases that were classified.

• TN = Did the model predicted correctly for the negative class as negative?

$$TNR = \frac{120}{40 + 120} = 0.75 \tag{8}$$

The result indicates TNR of 0.75 were the number of negative cases that were classified as negative.

Further, the false negative rate (FNR) is the proposition of positive cases that were incorrectly classified as negative.

• FN = Did the model predicted incorrectly the positive class as negative?

$$TNR = \frac{20}{180 + 20} = 0.1 \tag{9}$$

The results indicate that the FNR of 0.1 was the proposition of positive cases that were incorrectly classified as negative. The recall or true positive rate (TPR) is the proportion of the total number of correct predictions. We consider the equation as:

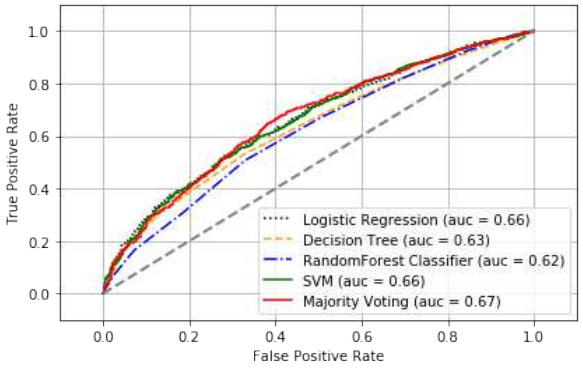

FIGURE 4. Plot the accuracy of all the algorithms in ROC curve for the LG, DT, RF, and SVM in MV.

• TP = Did the model predicted correctly for the positive class as positive?

$$TPR = \frac{180}{180 + 20} = 0.9 \tag{10}$$

The result indicates that the Recall or TPR of 0.9 was the proportion of the total number of instances that were identified correctly from the positive classes. To predict positive cases, we use precision (P) to determine the number of the proportion of instances is considered as correct. Hence the formula:

$$TPR = \frac{180}{180 + 40} = 0.81 \tag{11}$$

The final precision (P) of 0.81 was determined as the proportion of the total number of positive instances that were predicted correctly. The results show that the precision, recall and F-Score used to determine the accuracy and precision of the predictions are considered as accurate between the positive and negative rates. The result indicates that the F-Score of 0.85 was the harmonic mean between precision and recall. The Entropy is 0 if all member of E belongs to the same class, or 1 if they have the same number of samples in each group. The function entropy varies in range from 0 or 1.

7) ACCURACY OF THE ALGORITHMS IN ROC-AUC

Figure 4 depicts the ROC curve that determines the binary classifier system that determines the thresholds of the algorithms. We used AUC_ROC (Area Under Curve – Receiver Operating Characteristics) to model the selection metric for the bi-multiclass classification problem to distinguish between the probabilities of the given classes. AUC_ROC determines the True Positives Rates and False Negatives Rates. We plot the accuracy of all the algorithms in ROC. A 10-fold cross validation was used to determine the accuracy of the LR, DT, SVM and RF algorithms in the ROC. The black, orange, blue and green colours represent the algorithms. The x-axis represented as True Positive Rate and y-axis as False Positive rate. We used a python script to plot the graph as given in Figure 4:

8) 10-FOLD CROSS-VALIDATION

- [*ROCAUC* : 0.66 (+/ − 0.02) *LogisticRegression*]
- *ROCAUC* : 0.63 (+/ − 0.02) [*DecisionTree*]
- *ROCAUC* : 0.62 (+/ − 0.02) [*RandomForest*]
- *ROCAUC* : 0.66 (+/ − 0.02) [*SVM*]
- *ROCAUC* : 0.67 (+/ − 0.02) [*MajorityVoting*]

The results indicate that LG and SVM produced the highest results after we have used the ROC-AUC.

9) DETERMINING THE F-SCORE USING RECALL AND PRECISION RATES

For us to determine the precision, recall, and F-score, we answer the following questions regarding Table 1. Precision: how many positive instances were predicted correctly? Recall: how many instances were identified correctly from the positive classes? F-score: what is the harmonic mean between precision and recall? Using the results from evaluations in (I), we determine the F-Score and used the figures from the recall (0.9) and precision (0.81) to calculate the harmonic mean.

$$F = \frac{2 * 0.81 * 0.9}{0.81 + 0.91} = 0.85 \tag{12}$$

10) INCORPORATING ML AND CASE STUDY FOR EXPERIMENTATION

For us to determine the level of penetration, manipulation and the probability of an attack. We used a case study scenario of the remote CSC attack in [2] as below. The percentages figures were determined using the formula for calculating conditional probabilities in [2] from a low of 1 to a high of 100. The percentage figures in the penetration list are used for the result. The following is the scenario and the table from [2].

11) SCENARIO 1. REMOTE ATTACK ON THE CSC SYSTEM

The organization security team found that an adversary had intruded in the CSC system. The threat actor had compromised the workstation of the CMS that interfaced with suppliers, distributors, and third-party vendors. The organization's electronic products had been altered for some time. The CMS generated inaccurate customer electricity consumptions, which compromised the amount the customers were paying for their utility bills, their online payments, and third-party vendor systems. The organization used two types of payment systems, the prepaid system and post-paid system, that were all integrated into the CMS and HEMS. Using the formula for calculating conditional probabilities [2] and Activity 1 and Table 4, we determined the vulnerable spots, the severities of manipulation in percentages, and threat indicators. The percentages figures were calculated using the formula for calculating conditional probabilities. Further, the figures in penetration list are used to calculate the precision, recall and F-Score in Section 6 for the results.

VII. EXPERIMENTAL RESULTS

This section presents and analyses the results of the threat prediction. We follow a number of assessment parameters such as attack probability, TTP, vulnerable spots, and IoC for this purpose. The attack probability figures are derived from Table 2. The propagation is determined using a probability scale of 0–100%. A percentage score was given after calculating the degree of severity of each manipulation. Form low (≤15%), medium (16% to 59%), or high (above 60%).

- *Prediction of an attack probability.*

Table 3 presents the performance of the classifications of LR, DT, SVM, RF algorithms in identifying the various responses of cyberattacks based on the given malicious attack. From the table, LR achieved an accuracy of 66%, DT, 63% SVM 62% and RF 66%. Comparing the performance of the classifiers, LR and RF both performed better for the Precision, Recall and F-Score, whilst DT and SVM received a low precision, recall and F-score. Comparing that to the attack's categories signifies that Malware, Ransomware and spyware attacks identified different types of responses with 85% accuracy.

- *Prediction of TTP deployed based on the response of the cyberattacks.*

Table 4 presents the performance of the classification algorithms in identifying the various TTPs deployed, and responses based on the given attack vectors. Comparing the TTPs against the attack categories, XSS, session hijacking and RAT attack, DT and SVM achieved a low content for the low precision recall and F-score. However, LR received the highest precision and F-score for malware attack with 83% accuracy for TTPs deployed. Furthermore, ransomware and spyware attacks identified different types of responses for the TTPs with 83% accuracy for the harmonic mean in identifying the attack vectors being rootkit, email attachments and RAT.

- *Prediction of vulnerable spots based on the different types of responses of cyberattacks*

Table 5 presents the performance of the various classifications of the LR, DT, SVM and RF algorithms in identifying the vulnerable spots based on the different types of responses of cyberattacks. The vulnerable spots were identified from the CSC system probable threats table in [2] and used the manipulations figures for precision, recall and F-Score. LR and RF achieved a similar accuracy of 87% for the precision and F-score the successful attacks that signify the probability of exploits on the network nodes. Further, attacks such as malware and ransomware received higher precision based on the exploits and TTPS deployed with 92% accuracy. Whilst spear phishing, session hijacking and DDoS performs lower with the DT and SVM classifiers.

- *Predication of indicators of compromise (IoC).*

Table 6 presents the performance variations of the various classifications algorithms that identify what constitutes as indicators of compromise. With DDoS attack, RF presented the highest precision values of 83% compare to SVM indicating the extent of compromises on the network. LR received

TABLE 2. Probability and threat indicators.

Scenario	Vulnerable Spots	Penetration	Manipulation (%)	Probability	Threat Indicators
1	Firewall	Y	70	High	Wrong Firewall
2	IDS/IPS	Y	60	High	Configuration
3	Vendor	Y	80	High	Audit
4	Network	Y	40	Medium	Sub-netting
5	IP	Y	55	Medium	Segmentation
6	Database	Y	75	High	Sanitizations
7	Software	Y	75	High	Reprogram
8	Website	Y	90	High	SSL/TLS

TABLE 3. Predict the probability of an attack from the various endpoints.

ALGORITHMS	R			DT			SVM			RF		
ACCURACY (%)	**66**			**63**			**62**			**66**		
ATTACKS	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
XSS/Session Hijacking	0.88	0.38	0.65	0.58	0.42	0.68	0.55	0.38	0.63	0.88	0.38	0.65
Spyware/Ransomware	0.90	0.55	0.75	0.85	0.37	0.70	0.65	0.45	0.63	0.90	0.55	0.75
Spear Phishing	0.81	0.17	0.71	0.55	0.28	0.66	0.58	0.36	0.63	0.81	0.17	0.71
Session Hijacking	0.73	0.36	0.62	0.48	0.35	0.61	0.55	0.38	0.63	0.73	0.36	0.62
Rootkit/DDoS	0.56	0.37	0.65	0.57	0.33	0.58	0.53	0.35	0.63	0.56	0.37	0.65
RAT/Island Hopping	0.68	0.30	0.73	0.55	0.22	0.69	0.51	0.25	0.63	0.68	0.30	0.73
Ransomware/Malware	0.88	0.53	0.60	0.59	0.26	0.71	0.54	0.31	0.63	0.88	0.53	0.60
Malware/Spyware	0.81	0.48	0.68	0.58	0.51	0.73	0.55	0.45	0.63	0.81	0.48	0.68
DDoS	0.78	0.36	0.65	0.55	0.33	0.55	0.51	0.32	0.53	0.78	0.36	0.65

TABLE 4. Identify the different TTP deployed based on the response of the cyberattacks.

ALGORITHMS	LR			DT			SVM			RF		
ACCURACY (%)	**66**			**63**			**62**			**66**		
ATTACKS	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
XSS/Session Hijacking	0.82	0.26	0.55	0.55	0.31	0.61	0.55	0.27	0.56	0.82	0.26	0.55
Spyware/Ransomware	0.88	0.51	0.71	0.65	0.33	0.62	0.65	0.31	0.61	0.88	0.51	0.71
Spear Phishing	0.71	0.23	0.61	0.53	0.22	0.56	0.58	0.36	0.59	0.71	0.23	0.61
Session Hijacking	0.63	0.26	0.58	0.52	0.28	0.52	0.56	0.38	0.48	0.63	0.26	0.58
Rootkit/DDoS	0.51	0.27	0.63	0.51	0.31	0.58	0.48	0.35	0.57	0.51	0.27	0.63
RAT/Island Hopping	0.68	0.28	0.68	0.54	0.21	0.61	0.51	0.25	0.58	0.68	0.28	0.68
Ransomware/Malware	0.86	0.44	0.66	0.58	0.22	0.65	0.59	0.31	0.62	0.86	0.44	0.66
Malware/Spyware	0.79	0.41	0.67	0.65	0.51	0.63	0.55	0.45	0.61	0.79	0.41	0.67
DDoS	0.71	0.36	0.61	0.55	0.33	0.55	1.55	0.32	0.53	0.71	0.36	0.61

TABLE 5. Predict vulnerable spots based on the different types of responses of cyberattacks.

ALGORITHMS	LR			DT			SVM			RF		
ACCURACY (%)	**66**			**63**			**62**			**66**		
ATTACKS	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
XSS/Session Hijacking	0.63	0.60	0.61	0.65	0.61	0.62	0.61	0.59	0.60	0.62	0.59	0.61
Spyware/Ransomware	0.85	0.83	0.80	0.86	0.81	0.83	0.82	0.79	0.81	0.83	0.78	0.80
Spear Phishing	0.68	0.62	0.66	0.63	0.59	0.61	0.64	0.60	0.62	0.63	0.61	0.68
Session Hijacking	0.66	0.61	0.64	0.65	0.61	0.64	0.62	0.59	0.60	0.63	0.60	0.62
Rootkit/DDoS	0.64	0.60	0.61	0.63	0.61	0.58	0.61	0.57	0.59	0.64	0.38	0.58
RAT/Island Hopping	0.64	0.61	0.63	0.65	0.62	0.64	0.64	0.61	0.62	0.64	0.33	0.58
Ransomware/Malware	0.84	0.81	0.82	0.85	0.81	0.84	0.61	0.58	0.60	0.75	0.55	0.62
Malware/Spyware	0.82	0.77	0.81	0.86	0.83	0.85	0.85	0.81	0.83	0.66	0.45	0.69
DDoS	0.65	0.61	0.62	0.64	0.60	0.63	0.62	0.59	0.61	0.75	0.33	0.62

the highest precision and F-score for malware and spyware attacks, whereas RF and LR received the similar precision, recall and F-score.

VIII. DISCUSSIONS

The results for the predictive analytics are analysed in AUC_ROC as indicated in Figure 4. A 10-Fold cross-validation was used to run each algorithm to determine the parameter estimation and validated the accuracies. The evaluation of the accuracies of the metrics to answer the performance of the TPR, TNR, FPR, FNR as shown in Table 3. We determine the harmonic mean for the proportion of the total number of accuracies for the precision, recall, and F-score. The proportion for the precision is 220 for the

TABLE 6. Indicators of compromise (IOC). FOR performance variations of the various classifications algorithms.

ALGORITHMS	LR			DT			SVM			RF		
ACCURACY (%)	66			63			62			66		
ATTACKS	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
XSS/Session Hijacking	0.68	0.63	0.66	0.55	0.42	0.61	0.51	0.38	0.63	0.68	0.37	0.71
Spyware/Ransomware	0.80	0.8	0.75	0.85	0.55	0.70	0.65	0.45	0.63	0.78	0.52	0.76
Spear Phishing	0.81	0.17	0.71	0.55	0.65	0.70	0.55	0.45	0.63	0.77	0.17	0.68
Session Hijacking	0.73	0.66	0.62	0.55	0.65	0.70	0.55	0.45	0.63	0.73	0.65	0.62
Rootkit/DDoS	0.56	0.37	0.60	0.55	0.65	0.70	0.55	0.45	0.63	0.56	0.37	0.59
RAT/Island Hopping	0.68	0.30	0.33	0.55	0.65	0.70	0.55	0.45	0.63	0.68	0.30	0.63
Ransomware/Malware	0.70	0.33	0.62	0.55	0.65	0.70	0.55	0.45	0.63	0.72	0.33	0.60
Malware/Spyware	0.74	0.48	0.65	0.55	0.65	0.70	0.55	0.45	0.63	0.71	0.48	0.65
DDoS	0.68	0.56	0.65	0.55	0.65	0.70	1.55	0.45	0.63	0.68	0.56	0.57

TABLE 7. Mapping the attack category and predictive analytics.

Attack Category	CSC Attack Features	Threat Descriptions for Probable Cause of Attack	Threat Predictions (%)
1	XSS/Session Hijacking	Default Browser vulnerabilities and injecting a code in the URL or website	80
2-5	Spyware/Ransomware	Outdated Antivirus/Patches that are not updated regularly	90
6-7	Spear Phishing	Use Reconnaissance to identify vulnerable spots and attach email with a virus	80
8-9	Session Hijacking	Exploit Unchanged Hard-Coded password in software bought off the shelf	75
10-14	Rootkit/DDoS	Attack on BIOS or attach a virus to a USB key to cascade when booting.	80
15-20	RAT/Island Hopping	Attacks from Vendor systems to gain access to the organizational system	70
21-28	Ransomware/Malware	Exploiting outdated OS versions and encryptions especially TLS/SSL	60
29-35	Malware/Spyware	Packet injection and Resonance attacks	70
36-38	DDoS	Exploit IP Address Systems and Packet injections	55

number of positive instances that were predicted correctly. The proportion of recall (0.9) instances was identified correctly from the positive classes. The F-score of (0.85) was the harmonic mean between precision and recall. Hence, an accuracy of 85% is the total number of predictions that are considered accurate for the TPR and FPR. Further, we have a slight variation in our predictions of the TPF and FPR comparing the LR, DT, SVM, and RF algorithms in the pipeline and using MV for running them. However, the accuracy of the proportion of the total number of predictions remains accurate with an average of 65% and 30% as the combine values for the TPR and FPT respectively. Additionally, the results indicate that LG and SVM produced the highest results after we have used the ROC-AUC. The predictive analysis of our evaluation after we have used the CTI to gather information, gain knowledge and understanding of the organizational context and the situational awareness remains acceptable as compared to other literature that focused on ML only for predictions. The Table 7 shows the list the attack categories and threat predictions.

Table 6 combines the probability of attacks identified from previous work and map them with the feature descriptions of the threats to explains the predictive analytics [2]. The mapping includes attack categories, CSC attack features, and the threat describes for probable cause of attacks from the telemetry data and Microsoft endpoint protection threat report for the predictions. The attack categories were determined from the dataset of various threat descriptions from the telemetry data [23] that contains the properties of the various families of malware generated by the Windows defenders. The CSC attack features were derived from the various families of malware that has the probability of infecting the various CSC endpoint nodes. The threat descriptions were gathered by the threat report collected by the Microsoft Windows Defender [23]. The results specify that spyware/ransomware scored 90%. All the attack categories that score 80% indicated that an XSS or session hijacking could be deployed on the CSC website as uses public facing IPs it connects to various vendors. These could lead to spear phishing, rootkit and DDoS attacks. The rest of the threat prediction scores are explained in Table 7.

The paper reveals several observations made from the CSC attacks to using CTI lifecycle processes for intelligence gatherings, and ML for predictive analysis for the overall Smart CPS security improvement. The study revealed that several challenges are facing the organization in securing their systems as attackers are executing arbitrary commands on the supply chain systems remotely and manipulating systems.

A. MAPPING CYBERATTACKS ON CSC FOR PREDICTIVE ANALYTICS OF INDICATORS OF COMPROMISE

Table 8 provides details of how we mapped the cyberattacks on the CSC system for predictive analytics to determine the indicators of compromise. We used the threat modelling concepts in Section 3, and the properties to identify the

TABLE 8. Output parameters for indicators of compromise.

Cyberattack	Attack Pattern	Vulnerability	TTPs	IoCs
Malware	Insert a program in software	Untested Software	Insert Rootkit in code to hide in the system	Cascade to other networks nodes/ bypass antimalware
RAT	Hide in executable program, Backdoor code in an email attachment. HTTP Request Splitting, downloads	Network, Web and application server, Social Engineering, Phishing	Inject entry point identifier in the Explore Phase	Downloads itself when the user opens an email and provides access to the attacker
XSS	Embed malware in web browser content.	Programs that allow the remote host to execute codes and scripts.	Inject XSS payload and response split syntax in the user control input or URL	Injected scripts cascade to resources accessed by the applications
Ransomware	Social Engineering, Trojan, Botnets and Exploit kits to encrypt system files	Targets outdated antivirus and unpatched MS Windows application system	Map user environment, with documents, pictures and recycle bin and report content to C&C.	Calculate entropy of all file contents on the various systems, encrypt and propagate
Session Hijacking	Uses unauthentic HTTP cookies request from users.	Unencrypted websites, HTTP sessions, and open Wi-Fi connections	Insert network traffic that is not encrypted. Man-in-the-Middle attacks	Gain access and commits, APT, C5C and industrial espionage attacks.

TABLE 9. CSC security controls.

CSC Control	Descriptions	Asset	Approach	Implementation
Directive	Strategic management controls derived from the CTI and ML processes intended for policy formulation.	Identify Critical Assets and Security Framework that meet organizational goal	Map CTI gatherings and ML predictive analytics results to security goal	Assign controls to security teams to oversee the implementation. Adopt a framework or standard to support the development
Preventive	Proactive measures that are required to be implemented. Financial, physical, and technical measures intended to preclude actions violating policy or increasing risk to system resources.	Determine attacks that can exploit assets. Assign risks and threat levels to assets using CSCRM.	Determine Mitigations goals including internal and external audit controls	Create awareness by organize training and workshops to train users
Detective	Develop business impact assessment. Involve the use of practices, processes, and tools that identify and possibly react to security violations.	Implement periodic and ad-hoc security assessment using penetration testing and vulnerability assessment to pre-empt cyber threats	Use impact analysis and cost benefit analysis to determine the cost of alternatives of not investing in detection tools	Configure devices and automate passive tools on CSC systems to flag threats, run and monitor reports of firewalls, IDS/IPS, anti-malware and system updates
Corrective	Involve configurations and countermeasures designed to react to the detection of an incident to reduce or eliminate the zero-day attacks.	Design security policies that inform what must be done in the event of an incident	Develop Asset Inventory of all network nodes connected to the CSC organizational network including DHCP security	Implement Policies and business continuity plan to repair CSC systems, hard drive, patches systems, quarantine CSC systems
Recovery	Recovery strategy, Incident response and back up plans, regular updates, and contingency planning to ensure integrity or availability of the CSC system	Design policies and business impact assessment that can assist to restore the system or operation to a normal operating state upon any compromise as soon as possible.	Develop disaster recovery plan that will restore system to its operational state.	Form a team and Organize training and workshops to train staff to understand and be aware of the DRP implementations.

cyberattack, the attack pattern that were used, the vulnerable spots that were exploited, and the TTPs that are deployed by the threat actor on the CSC systems as the indicators of compromise (IoC). Indicators of compromise are parameters used

to express whether an attack-type is imminent, in progress or has occurred. Refer [2] further reading on threat modelling. Threat actors use sophisticated and stealthy methods to inject a virus, worms, bugs or a Trojan into software or in an HTTP request in an 'Island Hopping' attack. The intent is to penetrate the network or gain access to the webserver when a request is being processed. The motive could be to manipulate the vulnerable spots, alter the software and delivery channels and maintain APT and command & control presence.

Using the C&C methods, the attacker can modify products during manufacturing, manipulate it during distributions and the various domain attacks. These attacks could cascade to other nodes on the supply inbound and outbound chains. The table below provides a matrix that blends the input and output parameters for the prediction. Our observation is that the following vulnerabilities exist in the cyber supply chain system:

- The supply chain variables are accessible to the threat actor due to the business applications used for the supply chain variables and that could be exploited using incorrect user data.
- Information retrieved through inputted data is not configured properly due to poor validation.
- The variables are not well encapsulated to prevent software redirect. For instance, setting an input variable as public in a class when developing the software source codes makes the website open to external attackers.

B. MACHINE LEARNING FOR PREDICTIVE ANALYTICS

Machine learning approach to cybersecurity has been effective in analyzing and predicting future attacks and attack trends. We use ML techniques and classification algorithms including LD, SVM, DT, RF, and MV to develop threat intelligence techniques that can predict which nodes on our CSC system are venerable to attacks. We plot the accuracy of all the algorithms in ROC. AUC_ROC to determine the true positives and true negative rates. The results show that the best parameter result was SVM with an accuracy of 0.66. ML provides us with the ability to combine algorithms to determine which of them produced the highest accuracy and output for the best parameter for our prediction. However, it does not provide us with the ability to understand the threat actor's motives and intents.

C. COMPARING RESULTS WITH EXISTING WORKS

A stated in the related works, there have a lot of attention of using ML classifiers for cyber security. A vector space model is used for information retrieval for HTTP attacks using a decision tree algorithm to automatically label the request as malicious in the URL[11]. A number of classification algorithms LR, DT, NB, and SVM are considered for cloud security and tested the models in diverse operational conditions using cloud security scenarios[20]. Further, [21] used data mining and ML methods on Artificial Neural Network, Association rules, Fuzzy Association rules and Bayesian Networks classifiers for cybersecurity detection and analytics in intrusion detection security applications. Furthermore, [22] compared ML datasets used for analyzing network traffic and anomaly detection relevant for modern intrusion detections datasets. Moreover, [23], explored the classification of logs using a decision tree algorithm that models the correlation and normalization of security logs. Similarly, [24] compared NNge. LF, DT, Naïve Bayes, and SVM classification algorithms performance and ML predictions for power system disturbance and cyberattack discriminations. Then, [25] used an instance-based learning classification algorithm to learn a dataset for feature reduction and detection techniques to detect cyberattacks on smart grid. Additionally, [26] used an ensembled learning model based on the combination of a random subspace with random tree to detect cyberattacks on Industrial IoT networks. Likewise, [28] explored mitigating techniques on IoT cybersecurity threats in a smart city by using ML techniques to learn dataset on LR, SVM, DT, RF, ANN and KNN classifiers for anomaly detections. Further, [29] proposed a novel adaptive trust boundary protection for Industrial IoT network by using deep learning on a semi supervised model for detecting unknown cyberattacks. Furthermore, [30] used deep neural network discriminator on a down sample encoder cooperative data generator train the algorithm to capture actual distribution of attack model on industrial IoT attack surface. Additionally, authors in [31] predicted cybersecurity incidents by using Naive Bayes and SVM algorithms to investigate and analyse various datasets collected from SMEs. Finally, [32] model a risk teller system that used ML to predict which machines are at risk of getting infected or are clean and forecast if an organization may experience cybersecurity incidents in the future. Though all the works are relevant and contribute for the cyber security improvement. However, there is a lack of focus on the overall CSC security and ML classifiers are mainly used datasets for the threat predication. The proposed work presents a conceptual view by integrating relevant concepts from CSC and CTI domain. It provides a systematic threat analysis using the CTI techniques and integrates ML classifiers for the threat predication. Additionally, we considered LG, DT, SVM, RF algorithms in Majority Voting to learn the malware threat prediction dataset.

D. CSC SECURITY CONTROLS

There are various security controls in existence, whose effectiveness are based on existing CSC attacks and risks including CIS Controls 2018 and ISO27002:2011. We recommend the approach to address the CSC security using threat intelligence gathered from known and unknown attacks in line with organizational objectives and provide security recommendations. Some organizations provide a recommendation, however, not all may be relevant to the cyber supply chain organizational objective. Table 9 identifies basic concepts that are required to maintain security controls in the supply chain environment. To incorporate cybersecurity controls into a cyber supply chain system, we use knowledge of actual

CSC attacks that have occurred in the past. A compromised supply chain system provides us with the knowledge of previous attacks to continually learn from and build effective and practical defences mechanisms. To ensure proper CSC security controls, the organization must form a strategic team to identify, investigate, review and evaluate the supply chain system processes and applications.

E. THREAT INFORMATION SHARING

Threat information sharing is essential for any cyber physical system and specifically for the CSC context. It helps supply chain organisations and its stakeholders to aware about the current threat trends so that appropriate control can be identified to tackle the attacks. The CTI information includes threat landscapes, TTPs, tools, and intelligence reports. The threat intelligence is shared amongst the various organizations, institutions, vendors and businesses on the CSC system for strategic management decision making. It designates information and creates situational awareness on the various security alerts, assess and monitor threats, risk and existing controls. Due to the sensitive nature of the intelligence and privacy rules, these organizations are required to sign an agreement to ensure the following:

- Establish Information sharing rules
- Establish security system and audit rules
- Establish rules that govern the sharing of sensitive information
- Establish information classification rules. (Need to Know)

Challenges facing information sharing include the sensitivity nature of cyberattacks and the fact that it could lead to reputational damage, and sometimes legal ramifications. Most organizations are reluctant to share information relevant to CSC security.

IX. CONCLUSION

The integration of complex cyber physical infrastructures and applications in a CSC environment have brought economic, business, and societal impact for both national and global context in the areas of Transport, Energy, Healthcare, Manufacturing, and Communication. However, CPS security remains a challenge as vulnerability from any part of the system can pose risk within the overall supply chain context. This paper aims to improve CSC security by integrating CTI and ML for the threat analysis and predication. We considered the necessary concepts from CSC and CTI and a systematic process to analyse and predicate the threat. The experimental results showed that accuracies of the LG, DT, SVM, and RF algorithms in Majority Voting and identified a list of predicated threats. We also observed that CTI is effective to extract threat information, which can integrate into the ML classifiers for the threat predication. This allows CSC organization to analyse the existing controls and determine additional controls for the improvement of overall cyber security. It is necessary to consider the full automation of the process and industrial case study to generalize our findings. Furthermore, we are also planning to consider evaluating the existing controls and the necessary of future controls based on our prediction results.

REFERENCES

[1] National Cyber Security Centre. (2018). *Example of Supply Chain Attacks*. [Online] Available: https://www.ncsc.gov.uk/collection/supply-chain-security/supply-chain-attack-examples

[2] A. Yeboah-Ofori and S. Islam, "Cyber security threat modelling for supply chain organizational environments," *MDPI. Future Internet*, vol. 11, no. 3, p. 63, Mar. 2019. [Online]. Available: https://www.mdpi.com/1999-5903/11/3/63

[3] B. Woods and A. Bochman, "Supply chain in the software era," in *Scowcroft Center for Strategic and Security*. Washington, DC, USA: Atlantic Council, May 2018.

[4] *Exploring the Opportunities and Limitations of Current Threat Intelligence Platforms, Version 1*, ENISA, Dec. 2017. [Online]. Available: https://www.enisa.europa.eu/publications/exploring-the-opportunities-and-limitations-of-current- threat-intelligence-platforms

[5] C. Doerr, TU Delft CTI Labs. (2018). *Cyber Threat Intelligences Standards—A High Level Overview*. [Online]. Available: https://www.enisa.europa.eu/events/2018-cti-eu-event/cti-eu-2018-presentations/cyber-threat-intelligence-standardization.pdf

[6] Research Prediction. (2019). *Microsoft Malware Prediction*. [Online]. Available: https://www.kaggle.com/c/microsoft-malware-prediction/data

[7] A. Yeboah-Ofori and F. Katsriku, "Cybercrime and risks for cyber physical systems," *Int. J. Cyber-Secur. Digit. Forensics*, vol. 8, no. 1, pp. 43–57, 2019.

[8] CAPEC-437, Supply Chain. (Oct. 2018). *Common Attack Pattern Enumeration and Classification: Domain of Attack*. [Online]. Available: https://capec.mitre.org/data/definitions/437.html

[9] Open Web Application Security Project (OWASP). (2017). *The Ten Most Critical Application Security Risks, Creative Commons Attribution-Share Alike 4.0 International License*. [Online] Available: https://owasp.org/www-pdf-archive/OWASP_Top_10-2017_%28en%29.pdf.pdf

[10] US-Cert. (2020). *Building Security in Software & Supply Chain Assurance*. [Online]. Available: https://www.us-cert.gov/bsi/articles/knowledge/attack-patterns

[11] R. D. Labati, A. Genovese, V. Piuri, and F. Scotti, "Towards the prediction of renewable energy unbalance in smart grids," in *Proc. IEEE 4th Int. Forum Res. Technol. Soc. Ind. (RTSI)*, Palermo, Italy, Sep. 2018, pp. 1–5, doi: 10.1109/RTSI.2018.8548432.

[12] J. Boyens, C. Paulsen, R. Moorthy, and N. Bartol, "Supply chain risk management practices for federal information systems and organizations," *NIST Comput. Sec.*, vol. 800, no. 161, p. 32, 2015, doi: 10.6028/NIST.SP.800-161.

[13] *Framework for Improving Critical Infrastructure Cybersecurity, Version 1.1*, NIST, Gaithersburg, MD, USA, 2018, doi: 10.6028/NIST.CSWP.04162018.

[14] J. F. Miller, "Supply chain attack framework and attack pattern," MITRE, Tech. Rep. MTR140021, 2013. [Online]. Available: https://www.mitre.org/sites/default/files/publications/supply-chain-attack-framework-14-0228.pdf

[15] C. Ahlberg and C. Pace. *The Threat Intelligence Handbook*. [Online]. Available: https://paper.bobylive.com/Security/threat-intelligence-handbook-second-edition.pdf

[16] J. Freidman and M. Bouchard, "Definition guide to cyber threat intelligence. Using knowledge about adversary to win the war against targeted attacks," iSightPartners, CyberEdge Group LLC, Annapolis, MD, USA, Tech. Rep., 2018. [Online]. Available: https://cryptome.org/2015/09/cti-guide.pdf

[17] EY. (2016). *Cyber Threat Intelligence: Designing, Building and Operating an Effective Program*. [Online]. Available: https://relayto.com/ey-france/cyber-threat-intelligence-report-js5wmwy7/pdf

[18] A. Yeboah-Ofori and C. Boachie, "Malware attack predictive analytics in a cyber supply chain context using machine learning," in *Proc. ICSIoT*, 2019, pp. 66–73, doi: 10.1109/ICSIoT47925.2019.00019.

[19] B. Gallagher and T. Eliassi-Rad, "Classification of HTTP attacks: A study on the ECML/PKDD 2007 discovery challenge," Lawrence Liverpool Nat. Lab., Livermore, CA, USA, Tech. Rep., 2009, doi: 10.2172/1113394.

[20] D. Bhamare, T. Salman, M. Samaka, A. Erbad, and R. Jain, "Feasibility of supervised machine learning for cloud security," in *Proc. Int. Conf. Inf. Sci. Secur. (ICISS)*, Dec. 2016, pp. 1–5, doi: 10.1109/ICISSEC.2016.7885853.

[21] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 1153–1176, 2nd Quart., 2016, doi: 10.1109/COMST.2015.2494502.
[22] O. Yavanoglu and M. Aydos, "A review on cyber security datasets for machine learning algorithms," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2017, pp. 2186–2193, doi: 10.1109/BigData.2017.8258167.
[23] E. G. V. Villano, "Classification of logs using machine learning," M.S. thesis, Dept. Inf. Secur. Commun. Technol., Norwegian Univ. Sci. Technol., Trondheim, Norway, 2018.
[24] R. C. B. Hink, J. M. Beaver, M. A. Buckner, T. Morris, U. Adhikari, and S. Pan, "Machine learning for power system disturbance and cyber-attack discrimination," in *Proc. 7th Int. Symp. Resilient Control Syst. (ISRCS)*, Denver, CO, USA, Aug. 2014, pp. 1–8, doi: 10.1109/ISRCS.2014.6900095.
[25] A. Gumaei, M. M. Hassan, S. Huda, M. R. Hassan, D. Camacho, J. D. Ser, and G. Fortino, "A robust cyberattack detection approach using optimal features of SCADA power systems in smart grids," *Appl. Soft Comput.*, vol. 96, Nov. 2020, Art. no. 106658, doi: 10.1016/j.asoc.2020.106658.
[26] M. M. Hassan, A. Gumaei, S. Huda, and A. Almogren, "Increasing the trustworthiness in the industrial IoT networks through a reliable cyber-attack detection model," *IEEE Trans. Ind. Informat.*, vol. 16, no. 9, pp. 6154–6162, Sep. 2020, doi: 10.1109/TII.2020.2970074.
[27] J. Abawajy, S. Huda, S. Sharmeen, M. M. Hassan, and A. Almogren, "Identifying cyber threats to mobile-IoT applications in edge computing paradigm," *Elsevier Sci. Direct Future Gener. Comput. Syst.*, vol. 89, pp. 525–538, Dec. 2018, doi: 10.1016/j.future.2018.06.053.
[28] M. M. Rashid, J. Kamruzzaman, M. M. Hassan, T. Imam, and S. Gordon, "Cyberattacks detection in IoT-based smart city applications using machine learning techniques," *Int. J. Environ. Res. Public Health*, vol. 17, no. 24, p. 9347, Dec. 2020, doi: 10.3390/ijerph17249347.
[29] M. M. Hassan, S. Huda, S. Sharmeen, J. Abawajy, and G. Fortino, "An adaptive trust boundary protection for IIoT networks using deep-learning feature-extraction-based semisupervised model," *IEEE Trans. Ind. Informat.*, vol. 17, no. 4, pp. 2860–2870, Apr. 2021, doi: 10.1109/TII.2020.3015026.
[30] M. M. Hassan, M. R. Hassan, S. Huda, and V. H. C. de Albuquerque, "A robust deep-learning-enabled trust-boundary protection for adversarial industrial IoT environment," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9611–9621, Jun. 2021, doi: 10.1109/JIOT.2020.3019225.
[31] A. Mohasseb, B. Aziz, J. Jung, and J. Lee, "Predicting cybersecurity incidents using machine learning algorithms: A case study of Korean SMEs," in *Proc. INSTICC*, 2019, pp. 230–237, doi: 10.5220/0007309302300237.
[32] L. Bilge, Y. Han, and M. D. Amoco, "Risk teller: Predicting the risk of cyber incidents," in *Proc. CCS*, 2017, pp. 1299–1311, doi: 10.1145/3133956.3134022.
[33] Y. Liu, A. Sarabi, J. Zhang, P. Naghizadeh, M. Karir, and M. Liu, "Cloud with a chance of breach: Forecasting cyber security incidents," in *Proc. 24th USENIX Secur. Symp.*, Washington, DC, USA, 2015, pp. 1009–1024.
[34] *Guide to Cyber Threat Information Sharing*, document NIST 800-150, 2018, doi: 10.6028/NIST.SP.800-150.
[35] S. Barnum, "Standardizing cyber threat intelligence information with the structured threat information expression," V1.1. Revision, STIX, USA, Tech. Rep., 2014, vol. 1. [Online]. Available: https://www.mitre.org/publications/technical-papers/standardizing-cyber-threat-intelligence-information-with-the
[36] A. Yeboah-Ofori, S. Islam, and E. Yeboah-Boateng, "Cyber threat intelligence for improving cyber supply chain security," in *Proc. Int. Conf. Cyber Secur. Internet Things (ICSIoT)*, May 2019, pp. 28–33, doi: 10.1109/ICSIoT47925.2019.00012.
[37] A. Boschetti and L. Massaron, *Python Data Science Essentials*, 2nd ed. Dordrecht, The Netherlands: Springer, 2016. [Online]. Available: https://link.springer.com/content/pdf/10.1007/BF00994018.pdf
[38] A. Yeboah-Ofori, "Classification of malware attacks using machine learning in decision tree," *IJS*, vol. 11, no. 2, pp. 10–25, 2020. [Online]. Available: https://www.cscjournals.org/manuscript/Journals/IJS/Volume11/Issue2/IJS-155.pdf
[39] W. Wang and Z. Lu, "Cyber security in smart grid: Survey and challenges," *Elsevier Comput. Netw.*, vol. 57, no. 5, pp. 1344–1371, Apr. 2013.
[40] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Sep. 1995, doi: 10.1023/A:1022627411411.

ABEL YEBOAH-OFORI received the B.Sc. degree in computing and information systems from UEL, the M.Sc. degree in information security and computer forensics, and the Ph.D. degree in cyber security from the School of Architecture, Computing and Engineering (ACE), University of East London, U.K. He is currently a Lecturer with the University of West London. He holds a Postgraduate Certificate in Higher Educations Practices (PgCert) and a Fellow of the British Higher Education Academy (FHEA). He is a Prince 2 Project Management Practitioner, Certified Cyber Security and Digital Forensics Investigations practitioner. He has published journal articles, reviewed a few articles, and provided consultancy services. He was invited in 2018 to participate in Cyber Security Maturity Assessment Program with the Global Cyber Security Capacity Centre, USA, Oxford University, and the World Bank. He was invited to an Advisory and Review Workshop 2017 on National Cyber Security Policy and Strategy by MoC and Council of Europe (CoE) as part of GLACY+ activities. His research interests include cyber security, digital forensics, cyber threat intelligence, cyber-attack modeling, cyber supply chain security and risks, and machine learning.

SHAREEFUL ISLAM was a Visiting Researcher with the National Institute of Informatics (NII), Japan, and SBA Research, Austria. He is currently working as a Senior Lecturer and a Programme Leader with the Cyber Security and Network Program, School of ACE, University of East London, U.K. His research interests include in the area of cyber security, requirement engineering, information systems, and risk management. He has pioneered work in developing risk assessment and treatment method using business and technical goals, modeling language for cyber security risk management. The works are implemented in various application domain including cloud migration, critical infrastructure, and information system. He has published more than 70 articles (H-index 23) and he has led and/or participated in projects funded by the European Union (FP7), Innovate U.K., FwF, and DAAD. He has experience of acting as an Evaluator for national and international funding bodies, including the EPSRC, FwF, and CHIST-ERA. He is a Fellow of the British Higher Education Academy (HEA) and a certified PRINCE 2 and Management of RISK (MoR) practitioner.

SIN WEE LEE received the B.Eng. degree (Hons.) in electronics and computing from Nottingham Trent University, U.K., and the Ph.D. degree in neurocomputing from Leeds Beckett University, U.K. He is currently working with the School of Architecture, Computing and Engineering (ACE), University of East London, U.K. He has published more than 40 refereed articles in high-quality journals and international conferences in neural networks, data analytics, and machine learning. His main research interest and field of expertise are in the neural networks and machine learning for data analytics.

ZIA USH SHAMSZAMAN (Senior Member, IEEE) received the Master of Engineering (M.Eng.) degree from the Department of CICE, Hankuk University of Foreign Studies, South Korea, and the Ph.D. degree from the Insight Centre for Data Analytics, National University of Ireland Galway, Ireland. He is currently working as a Senior Lecturer in computer science with the Department of Computing and Games, Teesside University, U.K. He was involved in several research projects funded by FP7, SFI, Cisco Inc., and ETRI. He worked in the ICT industry over seven years and also achieved few professional certifications, such as CEH, CDCP, CCNA, and JNCIA-ER. His research interests include the IoT, the social IoT, CPS, cybersecurity, artificial intelligence, deep learning, semantic web, and ontologies. He is an Advisory Panel Member in Elsevier.

KHAN MUHAMMAD (Member, IEEE) received the Ph.D. degree in digital contents from Sejong University, Seoul, South Korea, in 2019.

He is currently an Assistant Professor with the Department of Software and the Director of the Visual Analytics for Knowledge Laboratory (VIS2KNOW Lab), Sejong University, Seoul. His research interests include intelligent video surveillance (fire/smoke scene analysis, transportation systems, and disaster management), medical image analysis, (brain MRI, diagnostic hysteroscopy, and wireless capsule endoscopy), information security (steganography, encryption, watermarking, and image hashing), video summarization, multimedia data analysis, computer vision, the IoT/IoMT, and smart cities. He is serving as a reviewer for over 100 well-reputed journals and conferences, from IEEE, ACM, Springer, Elsevier, Wiley, SAGE, and Hindawi publishers. He is an associate editor of four journals and an editorial board member of five journals.

METEB ALTAF received the Ph.D. degree from Brunel University London, London, U.K., in 2009. Since 2009, he has been with the KACST as an Assistant Research Professor. He was appointed as the Director Assistant for Administrative Affairs and the Director Assistant for Scientific Affairs with the National Center for Robotics and Intelligent Systems. After that, he was appointed as the Director of the National Robotics Technology and Intelligent Systems Center before it become known as the National Center for Robotics Technology and Internet of Things. He has been promoted as a Research Associate Professor. In the meantime, he became the Director of the Innovation Center for Industry 4.0, King Abdulaziz City for Science and Technology. He is currently the Director of the Advanced Manufacturing and Industry 4.0 Center. During his career life, he published number of articles in different well-known ISI journals and in well recognized conferences as well as he is lecturing at the Biomedical Technology Department, King Saud University. He has supervised more than 20 research projects locally and internationally as technology transfer projects.

MABROOK S. AL-RAKHAMI (Member, IEEE) received the master's degree in information systems from King Saud University, Riyadh, Saudi Arabia, where he is currently pursuing the Ph.D. degree with the Information Systems Department, College of Computer and Information Sciences. He has worked as a Lecturer with King Saud University, Muzahimiyah Branch, and taught many courses, such as programming languages in computer and information science. He has authored several articles in peer-reviewed IEEE/ACM/Springer/Wiley journals and conferences. His research interests include edge intelligence, social networks, cloud computing, the Internet of Things, big data, and health informatics.