Super Learner Ensemble for Anomaly Detection and Cyber-risk Quantification in Industrial Control Systems

Gabriela Ahmadi-Assalemi, Haider Al-Khateeb, Gregory Epiphaniou, Amar Aggoun.

***Abstract*— Industrial Control Systems (ICS) are integral parts of smart cities and critical to modern societies. Despite indisputable opportunities introduced by disruptor technologies, they proliferate the cybersecurity threat landscape, which is increasingly more hostile. The quantum of sensors utilised by ICS aided by Artificial Intelligence (AI) enable data collection capabilities to facilitate automation, process streamlining and cost reduction. However, apart from operational use, the sensors generated data combined with AI can be innovatively utilised to model anomalous behaviour as part of layered security to increase resilience to cyber-attacks. We introduce a framework to profile anomalous behaviour in ICS and derive a cyber-risk score. A novel super learner ensemble for one-class classification is developed, using overlapping rolling windows with stratified, k-fold, n-repeat cross-validation applied to each base-learner followed by majority voting to derive the best learner. Our approach is demonstrated on a liquid distribution sensor dataset. The experimental results reveal that the proposed technique achieves an overall F1-score of 99.13%, an anomalous recall score of 99% detecting anomalies lasting only 17 seconds. The key strength of the framework is the low computational complexity and error rate. The framework is modular, generic, applicable to other ICS and transferable to other smart city sectors.**

Keywords*— *Machine Learning, Cyber-Physical Systems, cyber security, digital forensic and incident response, Supervisory Control and Data Acquisition, SCADA, Programmable Logic Controllers, PLC, Human Machine Interface, HMI, Industry 4.0, Internet of Things, IoT, Smart City, Insider Threat, cyber resilience

I. Introduction

Industrial Control Systems (ICS) such as water treatment, water distribution plants, manufacturing, power grids, wind turbines, and transportation are integral components of smart cities. ICS have a massive impact on the wider society which, if disrupted, could result in devastating consequences [1-4]. Cyber-attacks exploiting the Internet of Things (IoT) and critical infrastructure are an attractive target for threat actors. The threat landscape is becoming more hostile, shifting towards organised cyber-crime and Advanced Persistent Threat (APT) [5-11]. The motivation for this paper arises from the opportunity to address emerging and increasing threats to ICS. ICS are complex, interconnected and distributed networks that consist of segments including corporate networks, logic and physical control [12]. The complex interconnectivity and the prevalence of cyber components within these segments such as Supervisory Control and Data Acquisition (SCADA), Human-Machine-Interface (HMI), Programmable Logic Controllers (PLC) and the quantum of sensors underpinned by communication networks make ICS vulnerable to cyber-attacks [12, 13]. Cyber-attacks against ICS could potentially result in real-world damage with significant and hazardous impact on communities [13]. The impact could have a monetary impact on businesses, loss of Intellectual Property (IP), threat to national security including socio-economic consequences on entire ecosystems [1, 2, 14][13]. Therefore, it is critical to protect ICS from cyber-attacks, pointing to the significance of research to develop protective mechanisms. ICS sensor-generated data is used for operational monitoring. ICS sensor-generated data can be innovatively utilised to improve the defence-in-depth thus increase resilience to cyber-attacks. There is little or no research contributing towards other aspects enabled by anomaly detection.

In this study, we address the research questions which can be expressed as: How can we form a framework which addresses anomaly detection in Cyber-Physical Systems (CPS) such that it is optimised, and security-process driven? How can this framework be utilised to quantify the cyber-risk in CPS? And how can this framework support Digital Forensic and Incident Response (DFIR)? We attempt to address the problem of a security process-driven proactive protective mechanism as part of layered defence-in-depth in CPS utilising Machine Learning (ML) techniques. Anomalous behaviour detection from sensor data has key advantages. Anomalous behaviour detection is attainable from sensor data hence previously unknown attacks are detectable including external threat actors and smart-cyber insiders. The key contributions of this study are as follows:

- A novel **S**uper learner **E**nsemble **A**nomaly detection cyber-**R**isk quantification (SPEAR) framework is introduced. The SPEAR framework provides a solution for resilient profiling of anomalous behaviour in ICS from sensors generated data and cyber-risk quantification in the prevalence of anomalous behaviour.
- A super learner ensemble model is constructed using overlapping Rolling Windows (RW) (also referred to as sliding windows in literature, in this study we use the term rolling windows) to create a robust predictor for anomalous behaviour detection in ICS. The resulting best learner in the stack is based on majority voting. The model achieves an overall F1-score of 99.13% and an anomalous recall score of 99.00% for binary classification of one against all in a short data segment including detecting anomalies lasting only 17s.
- A Bayesian Belief Network (BBN) model for cyber-risk quantification is proposed leveraging ICS sensor data. The model supports post-incident investigations as part of DFIR to objectively quantify the cyber-risk value in the prevalence of anomalous behaviour.

- The SPEAR framework's applicability to support DFIR is theoretically discussed as part of forensic readiness and to improve defence-in-depth in ICS.
- The reviewed scientific literature suggests that although cyber-risk management is an area of scientific interest, research focusing on a quantification of cyber-risk value based on anomalous behaviour detection in CPS remains limited. To the best of our knowledge, this research is the first study to combine attempts to address quantifying cyber-risk value when anomalous behaviour is prevalent as part of Security-by-Design (SbD) and to support DFIR

In the remainder of this paper, we discuss the related work in Section II, the SPEAR framework including the methodology is presented in Section III. Section IV shares an ICS case study reporting results of our experiment and the Discussion is presented in Section V. Finally, we conclude our study and future works in Section VI.

II. Related Work

A. Inherent and Emerging Threats in ICS

Ubiquitous sensor networks are transformational to the operations of ICS. They are integral segments of smart cities due to the level of control and intelligence gained from their sensing, processing, and communication capabilities. Broadly, CPS are subject to cyber-attacks such as targeting authentication through compromised key attacks [15], compromising the confidentiality and integrity of CPS data by targeting the CPS data storage, communication channels, actuators' controls and end-points [16]. Threats specific to ICS are often more basic such as outdated security measures. For example, in brownfield implementations where legacy systems coexist with innovative sensing technology, equipment is exposed through vulnerabilities resulting from outdated security updates. A false sense of security is provided by securing physical aspects of CPS while wireless and remote connectivity surpasses the physical boundaries. Poor configuration, lack of appropriate network segregation, compromised credentials targeting cloud-based ICS systems, backdoors, remote access channels, software vulnerabilities and smart-cyber insiders are attractive attack vectors for threat actors [3]. However, compared with the well-established field of Information Security (IS), ICS security is a less-well understood discipline and the attacks remain poorly described. [17].

Notably, the numbers of widely acknowledged and reported high-profile attacks on Critical National Infrastructure (CNI) are limited. For example, Solar Sunrise 1998 was one of the earliest multi-stage cyber-attacks against critical infrastructure which systematically exploited a vulnerability in the Sun Solaris operating systems targeting the United States (US) Department of Defence networks [2, 18]. Another substantial incident was Stuxnet [19] where a malware attack that targeted an Iranian nuclear plant. Norsk Hydro a renewable energy supplier was targeted by the LockerGoga ransomware [20]. The attack on the Ukrainian power grid compromised the SCADA system [21]. The attack on the Kemuri Water Company compromised the sensors monitoring the plant and the levels of chemicals in a water treatment plant were altered [22]. In the recent attack against Florida's Oldsmar's water treatment facility, the attackers briefly increased the amount of sodium hydroxide a hundred-fold. The chemical is the main ingredient in drain cleaners. The facility supplies water to commercial establishments and about fifteen thousand residents. This attack could have had profound consequences on the community [23]. Interconnected systems are subject to attacks and it may not be possible to establish the source or the motive [2, 24]. Thus, it is critical to establish an intelligence-based defence-in-depth mechanism and understand the threat models posed against ICS.

B. Threat Modelling

We consider ICS related cybersecurity attacks discussed in the literature and covered in the previous section [2, 5-11, 18-23] in addition to those listed in Table XI. Social challenges such as accidental insiders, disgruntled employees and social engineering are underestimated and difficult to detect. These challenges fall outside of traditional cyber defence measures such as firewalls, access control, network, and host security. We suppose an attack vector where a disgruntled employee or an external contractor have authorised unmonitored access coupled with knowledge of the operational infrastructure, the software systems, and environmental data configurations. In this situation, access to the computer systems could result in unauthorised manipulation including accidental disruption. Moreover, such compromise could be performed locally or remotely. Physical access to operational infrastructure and tampering with the physical process in the system could lead to physical damage and alteration to the expected functioning of the operational infrastructure.

Likewise, open standards and interconnectivity with corporate and public networks in ICS create new attack vectors [3, 12, 13, 15, 16]. Resourceful attack actors such as APT will adapt their tactics, techniques and procedures to exploit these opportunities [7]. The initial attack vectors could include the attacker's ability to compromise a device on the corporate network or an internet-exposed ICS component leveraging unpatched or a zero-day vulnerability [12]. Exploiting legitimate account credentials coupled with poorly designed or bypassed security controls, could enable the attacker gain access to the ICS operational infrastructure [25]. We assume an attack vector where the attacker gains access to the logic control and the physical control layers. Attackers could inject code altering the sensor values creating a difference registered by the PLC compared with the real states of the physical process. Furthermore, attackers could inject command to gain control of the actuators creating a difference between the expected and registered state. Therefore, pertinent to this study we consider attacks to the operational infrastructure at the physical control layer listed in Table XI.

C. Application of the learning techniques

ML techniques utilised by domains including social media, medical analysis, computer vision and gaming are applied in cyber defence measures in smart city sectors such as transportation, healthcare, buildings and ICS [1, 26-30]. For example, ML techniques are utilised as cyber defence measures for anomalous behaviour detection. One of the advantages of ML techniques over signature, statistical or rule-based approaches is detection of previously unseen attacks. According to [29], ML is frequently applied in intrusion detection, malware analysis, phish and spam detection. The utilised ML approaches are categorised in two main domains, shallow and deep learning. They both include supervised, unsupervised, and semi-supervised learning models. In supervised learning, each instance has a pre-assigned class. The classifier is trained to apply the labelling of the target feature on new unseen data whereas in unsupervised learning the classifier is looking for the presence of patterns if [31, 32].

This poses an important question; which one is the most suitable learning method. There is no ultimate de-facto classifier, the choice depending on several factors not least. the problem being solved. Other factors include distinct types of classifiers perform differently [31, 32], the types of datasets available, organisational business and risk models. The following study [33] proposes a statistical testing procedure for algorithm comparison. Repeated training and testing are asserted in other scientific literature [31]. Another study [34] proposed an ensemble anomaly detection generic framework using RW for energy consumption in buildings. To protect IoT network traffic, [35] utilised an ensemble learning method. The proposed method consisted of three ML techniques; Naïve Bayes (NB), Decision Tree (DT) and Artificial Neural Network (ANN) NB based on the AdaBoost classifier with majority voting. Furthermore, consideration should be given to the type of classifier for the scale and range of the investigated cyber-attacks, the classifier's performance in detecting the anomaly [29], and the algorithm's generalisation ability [32]. Different approaches were proposed for anomalous behaviour detection. Algorithms were utilised individually or as part of an ensemble such as Support Vector Machine (SVM) [34, 36], Principal Component Analysis (PCA) [34], Random Forest (RF) [34], Autoencoder (AC) [34], ANN [35], DT [35], NB [35], Isolation Forest (IF) [37, 38]. It is not the aim of this study to solve the classifier problem but to apply a robust model to the problem outlined in this paper and present a direction for future research.

D. Approach to quantifying the cyber-risk value

The Confidentiality, Integrity and Availability (CIA) triad have been considered fundamental to good security practice. The CIA triad has been adopted and driven by the IS community; however, it does not sufficiently address the security aspects in ICS. For example, understanding of control and safety facets are important in ICS due to their complexity, fragmentation, real-time interactions. Likewise, ICS can be geographically dispersed and potentially owned by multiple legal entities and jurisdictions. To overcome the limitations of the traditional CIA approach and address the challenges in ICS we investigate the use of the Parkerian Hexad (PH) as a forward-looking alternative to converge engineering and IS good practices [1, 39, 40]. Safety is considered as a seventh dimension by the authors of the following study [39] who assert that the creation and use of the data should not be harmful. Furthermore, the authors emphasize that the safety dimension provides context for cybersecurity risk assessment and impacts situational awareness. We do not challenge this approach and consider the safety aspect of significance to quantifying the cyber-risk value in Cyber-Physical-Natural (CPN) ecosystems [1].

Scientific literature shows little evidence of deviation from the conventional risk formula. Only a few studies propose enhancements such as the architectural perspective of risk [41]. Studies such as those listed in Table I and others [42-44] focus on estimating the risk in control systems. For example, the following study [44] predicts the risk level at a particular time whereas other studies [42, 43] dynamically and timely calculate the risk. Another study [45] investigates ways to enhance the resilience of power systems against cyber-attacks. The method of assessing cyber-risk remains a significant shortcoming in CPS. ICS are highly automated, designed for safety, reliability and availability and not developed with the SbD approach. Therefore, ICS have limited consideration to understand the cyber-risk value, the scale of the impact caused by cyber-attacks [46] or support for DFIR as shown in Table I. Most empirical studies including [47-51] address static risk without necessarily quantifying the cyber-risks. In this paper, we use the term "cybersecurity risk" and "cyber-risk" interchangeably.

Furthermore, the researchers acknowledge the scientific efforts to improve the cybersecurity posture of ICS including through frameworks such as the National Institute of Standards and Technology (NIST) voluntary Framework for Improving Infrastructure Cybersecurity [52]. We do not challenge existing approaches. However, we lean on the characteristics outlined in this section, which quantify the cyber-risk value and contextualise situational awareness comprehensively for CPS. We seek to address the cyber-risk value quantitatively from CPS datasets in the prevalence of detected anomalies.

Table I
Comparison of Risk Assessment (RA) approaches of similar studies in ICS and support for DFIR

Studies	Address ICS	Dynamic RA	Support for DFIR	Empirical Study addressing RA
This Study	✓	✓	✓	✓
[42]	✓	✓	×	✓
[43]	✓	✓	×	✓
[53]	×	✓	×	✓
[50]	×	×	×	✓
[47]	✓	×	×	✓
[51]	×	×	×	✓
[49]	×	×	×	✓
[48]	×	×	×	✓
[54]	✓	✓	×	✓
[55]	×	✓	×	✓
[56]	✓	✓	×	✓
[57]	×	✓	×	✓
[25]	✓	×	×	✓
[45]	✓	✓	×	✓
[46]	✓	✓	×	✓

III. THE SPEAR FRAMEWORK

A. *The SPEAR Framework Overview*

This study presents the SPEAR Framework shown in Fig. 1, to facilitate proactive anomalous behaviour detection in ICS. Our approach is motivated by a study from the field of genetics and molecular biology [58]. The authors construct a fast learner using a weighted combination of several candidates and utilise V-fold cross-validation to avoid overfitting. Their approach is aimed to generalise to any parameter. As part of our framework in this study, leveraging supervised learning we construct a super learner ensemble. Using overlapping RW, we derive the best predictor for anomalous behaviour detection for the datasets. Our approach differs from other studies such as [34, 35, 58]. We do not rely on a single classification model for the base-learners [31], we use a stack of base-learners, overlapping RW and apply stratified k-fold n-repeat Cross-Validation (CV) to each individual base-learner [32] at the time of training the model. The choice of the best learner in the stack is based on majority voting.

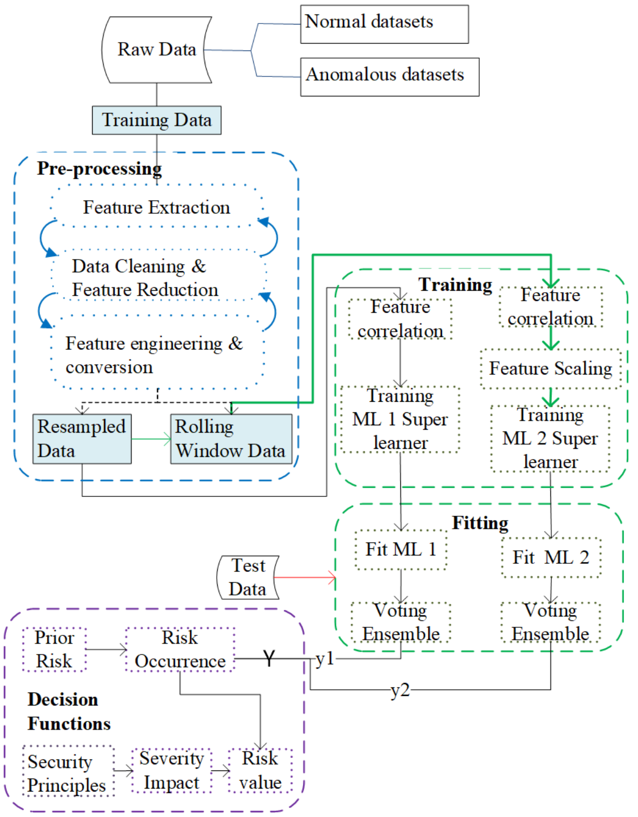

Fig. 1 SPEAR Framework, which consists of the data pre-processing, model training, model fitting and decision function stages.

B. *Procedure Design*

Firstly, during the pre-processing phase, the temporal dataset is transformed, and features are extracted to solve the problem as a supervised model as shown in Fig. 4. Contextual features contained within the date and timestamp are introduced including a feature to represent the elapsed time from the beginning of the event. This research study does not seek to establish the date and time of the events. Its interest is to uncover a behavioural anomaly as a temporal event in a sequence of events. Next, irrelevant, missing, or duplicated feature values or instances could skew the learning algorithm performance. Such features are addressed during the data cleaning phase utilising several data cleaning techniques. Feature engineering introduces additional features which contribute to the learning model's performance.

To evaluate the performance of the ML model, consideration was given to the train, test and validation data subsets. The anomalous behaviour varies from a few seconds to several minutes as presented in Table XI. Therefore, consideration to handle imbalanced datasets is factored in. For the one-class binary classification for the outlier detection, the normal and anomalous datasets are combined into a single dataset. The base-learners are trained on random subsets of the total training data and are fitted with test data. The same approach is applied to individual attacks, creating a set of imbalanced datasets. The stratify parameter is used to retain the train-test split ratio for the train and test sets, setting aside 30% of the dataset for testing. Grid-search is utilised for hyperparameter optimisation. To avoid overfitting or significantly reducing the number of samples in the train or test sets, repeated stratified 10-fold, 3-repeat CV is applied to the base-learners during the model training. The meta-learner is trained from the outputs of the sub-models utilising a list of defined estimators from the stack as input arguments. Majority voting ensemble $\hat{y} = \mathit{mode}\{ \lambda_{b1}(x), \ldots \lambda_{b1n}(x)\}$ is applied before the final prediction is produced as illustrated in Fig. 4 [31, 32]. The performance scores are derived from the confusion matrix, see Fig. 2.

Data Class		Predicted Class	
		Normal	Anomalous
Actual Class	Normal	**True Positive (TP)**	**False Negative (FN)**
	Anomalous	**False Positive (FP)**	**True Negative (TN)**

Fig. 2 Confusion Matrix

C. *Piloting*

Pilot experimental work focused on a small subset of the individual attack scenarios. Pilot experimentation helped to test and evaluate the instrument, the procedure and the formal experiment's optimal time window.

D. *The Feature Extraction*

The timestamp feature is rearranged to the 'Date' format [dd/mm/yyyy hh:mm:ss.sss]. The single 'Register' feature has all sensor types as the feature's values. The dataset features are rearranged according to the algorithm in Table II, such that each sensor type is represented as an individual register feature $f_{v1}...f_{vn}$ labelled R1…Ri in time series. Feature extraction by sensor type separates sensors by their functions while their time segments are unchanged. Furthermore, this feature extraction enables the grouping of different sensor types into learning sub-models which is an interesting approach, similar to another study [13].

Table II

Algorithm 1: SPEAR Framework Feature Extraction Algorithm

Input raw dataset of instances $i_1...i_n$ with features $f_{r1}...f_{rn}$, of values $[v_1....v_n]$

Output labelled dataset of instances $i_1...i_n$ and class (normal data [0], anomalous data [1]), with features $f_{v1}...f_{vn}$ of value $[v_1....v_n]$

Step 1: Load raw dataset into dataframe

```
    for $f_r$ identify unique values
        extract v into $f_v$ using index 'Date'
        label Class for $i_1 \dots i_n$
    end for
```

E. Data Cleaning

Recording of sensor readings may become corrupted or erroneous for several reasons during the data collection process such as malfunctioning sensors, malicious activity, disruptions in network connectivity or the data collection infrastructure. This could result in noisy, missing, or duplicated observations within the dataset, in real-world data and large datasets the likelihood of erroneous data increases. Therefore, data cleaning is essential for a meaningful analysis of the dataset which we handle according to the algorithm in Table III. We identify missing values, duplicate instances, unique feature values, single-value and low variance features. In this dataset, features that have a single value or very few unique values, have zero or low variance of <=0.001% are not likely to contribute to the predictive model's performance. Therefore, such features are removed from the dataset. Missing values are often marked with a placeholder such as 'NaN' or left blank. However, not all algorithms have the resilience to deal with missing values, particularly predictive techniques [59]. To minimise the loss of data, missing values are marked. Instances are dropped where missing values <=0.5% of the dataset, otherwise, values would be imputed using the forward-fill method propagating the last observed non-null value forward until the next non-null value is reached.

Table III

Algorithm 2: SPEAR Framework Data Cleaning and Feature Reduction

```
Input raw dataset of features $f_{v1} \dots f_{vn}$, values [$v_1 \dots v_n$] and instances
$i_1 \dots i_n$
Output cleaned dataset with feature-set $f_{vc1} \dots f_{vcn}$ of value [$v_1 \dots v_n$] and
instances $i_{c1} \dots i_{cn}$
Step 1: identify missing $f_v$
    for i replace missing $f_v$[v==NaN]
        count $f_v$[v==NaN]
        print summary of missing $f_v$
        if missing $f_v$[v <0.5%]
            remove missing instances
        else
            impute missing $f_v$[v, impute method == ffill]
            verify missing $f_v$[v]
        then go to step 2
    end for
Step2: identify duplicate instances $i_d$
    for i:
        calculate $i_d$
        remove $i_d$
        then go to step 3
    end for
Step 3: Identify features with single value, few values and near zero
    variance predictors:
    for i in range $f_v$[v]:
        print $f_v$[v], len==unique
        if len==unique where (unique $f_v$[v]/total i*100)<=0.001%)
            drop $f_v$
        else
            cleaned dataset of feature-set $f_{vc1} \dots f_{vcn}$ of value [$v_1 \dots v_n$]
            and instances $i_{c1} \dots i_{cn}$
    end for
```

F. Feature Engineering and Visualisation

This part of the pre-processing phase introduces additional features to the dataset, according to the algorithm in Table IV. To apply an ML algorithm to train the dataset, the time-series dataset is transformed, so that it can be modelled as a supervised problem. Contextual features based on the date and timestamp are introduced. While information about business hours, public holidays, years' seasons, part of the week could be extracted and enhance the performance of a learning algorithm, in this dataset using the date would not likely help the learning algorithm and could result in inferior performance. The dataset is resampled using seconds as the smallest time unit, and the mean values for each sensor using the default label bucket and bin interval values. Furthermore, the seasonality and the trend characteristics of the discrete, pumps and ultrasound sensors in the dataset were established, the normal dataset's repeatable patterns are shown Fig. 3. The test for a null hypothesis whether the dataset is stationary, the statistical Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test [60] for stationarity around a deterministic trend is applied to the sensors. The probability score, the p-value of the test is >0.05 (significance level), confirming the KPSS null hypotheses, and showing the sensors' stationarity around a constant as shown in Table V. The equation used to calculate the lags [60], where 'n' represents the length of the series:

$$int(12*(n/100**(1/4))$$

No other environmental information is introduced. Further analysis of the dataset uses the seasonal data decomposition function to verify the stationarity around a deterministic trend and decomposes the data into four components: level, trend, seasonality, and noise. The components are structured as outlined in the following equations, where 'y(t)' is the time series dataset over some time, Level (L), Trend (T), Seasonality (S), Noise (N):

$$y(t) = L+ T + S + N$$

Furthermore, 3s, 5s, 10s size rolling mean windows and parameters including minimum period and window types are utilised to test and evaluate the model's performance.

Table IV

Algorithm 3: Feature Engineering for the SPEAR Framework

```
Input: cleaned dataset of features $f_{vc1} \dots f_{vcn}$ of value [$v_1 \dots v_n$] and
instances $i_{c1} \dots i_{cn}$, index [Date]
Output: pre-processed dataset with features $f_1 \dots f_n$, values [$v_1 \dots v_n$] and
instances $i_1 \dots i_n$
for i
    set index $f_{vc}$==datetime ['%d/%m/%Y%H:%M:%S.%f']
    transform $f_{vc}$ datetime to new features where 'second'[v==%S],
     'minute'[v==%M], 'hour'[v==%H]
    then resample $i_{c1} \dots i_{cn}$, f index (v==datetime [%d/%m/%Y
     %H:%M:%S']), $f_1 \dots f_n$ ([v==mean])
end for
for i of sensors $f_{vc}$
    do kpss stationarity test
end for
Apply Algorithm 2 Step1
for i
    apply rolling window (interval[s], min_periods, win_type, mean)
end for
reset index
drop $f_{vc}$ 'datetime'
```

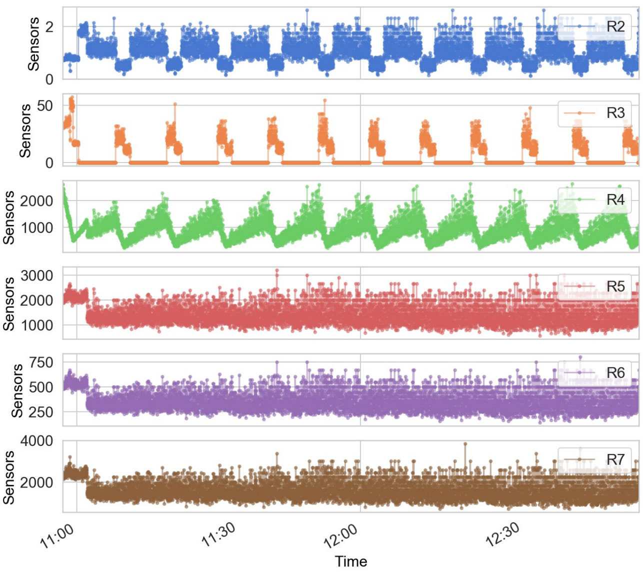

Fig. 3 Sensors' temporal distribution – normal dataset.

Table V
KPSS test output stationarity test – normal dataset

Test Output	**Sensor Output Values**		
Sensors	**Discreet**	**Pumps**	**Ultrasound**
Test Statistics:	0.14212411 940017958	0.214696034 58547087	0.1368645250 9501715
p-value:	0.1		
Critical Values:	'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739		
num lags:	36		
Stationarity	Series is Stationary		

G. The SPEAR Framework Learning Algorithms

This section introduces the proposed detection scheme presenting two ML algorithms for the framework.

The supervised ML model uses the concept of a super-learner ensemble for classification algorithms for anomalous behaviour identification in CPS. This model consists of nine stacked base-learners. The base-learners are typically investigated independently to gain the best performance based on the optimal set of features and classifiers. The aim is to avoid selecting a suboptimal classifier to solve the problem, to improve the predictive performance and increase the generalisation performance of the algorithm. The learning algorithm, as shown in Table VI, is based on the general framework of several ensemble algorithms [32]. Scientific studies accept that meta-learners may not produce better results than any of the classifiers used individually, nonetheless their use mitigates the risk of using an inefficient classifier [31, 61].

The learning model is trained and tuned using resampling and resampling with rolling windows techniques. The stacked base-learners are trained on random subsets of the total training data, they are fitted with test data and produce accuracy scores. The meta-learner is a heterogenous ensemble derived from the base-learners consisting of different algorithms. The meta-learner is trained from the base-learners' outputs, utilising a list of defined estimators from the stack as input arguments. The meta-learner applies the majority voting method before the final prediction is produced, see Fig. 4**.** [31, 32]. The labelled dataset uses the stratify parameter to retain the train-test split ratio and is split into train-test sets, setting aside 30% of the dataset for testing. A dictionary of parameter values is defined for hyperparameter optimisation and uses a grid-search technique to determine the best parameter set. To avoid overfitting or significantly reducing the number of samples in the train or test sets, repeated stratified 10-fold, 3-repeat CV is applied. The two models are trained independently applying 1s resampling and a 10s rolling window to the dataset, as shown in Fig. 4**.**

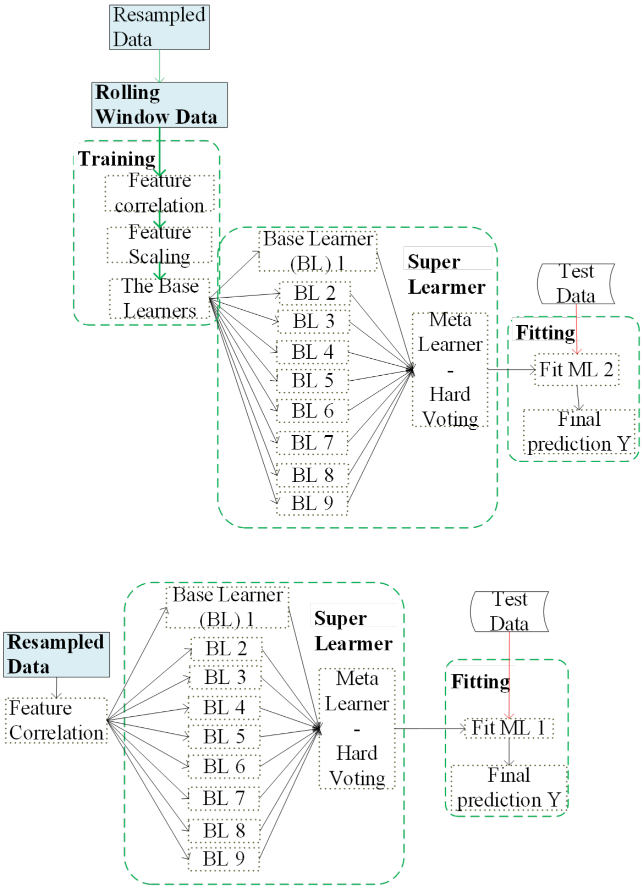

Fig. 4 Supervised ML, super learner ensemble model.

Table VI
Algorithm 4: Supervised Learning Ensemble Super Learner for the SPEAR Framework based on the general framework of ensemble algorithms [32]

```
Input: Pre-processed dataset $D = \{(x_1, y_1), …., (x_n, y_n)\}$,
base-learners algorithm $\lambda_{t1},… \lambda_{tn}$, meta-learner algorithm $\lambda$
Output: $H(x)$
Other Definitions: $h_t$ =base-learner, T = Number of learning
algorithms, h' = meta-learner
#Train the base-learners by applying the base-learner learning
#algorithms to the pre-processed original training dataset
for t in $(t_1...t_T)$:
    $h_t = \lambda_t(D)$;
end for
#Produce a new dataset for training the meta-learner,
#The output of the base-learners is the input for the meta-learner
#Original labels are retained
$D' = \Theta$;
for i in $(i_1...i_n)$:
    for t in $(t_1...t_T)$:
        $z_{it} = h_t(x_i)$;
    end for
    #The new dataset is produced from the cross-validated the total
    #number of base-learners. The meta-learner is applied where h' will
    #become the function of $z_{it1}, …. z_{itT}$ for y.
      $D' = D' \cup ((z_{it1}, …. z_{itT}), y_i)^{in}_{i=1}$;
end for
#train the meta-learner h' by applying the meta learner algorithm λ to
#the newly generated dataset D'.
   h'=λ(D');
Output: $H(x) = h'(h_{t1}(x), ….., h_{tT}(x))$
```

The unsupervised ML model covered in Fig. 4 and Fig. 5, uses the concept of outlier detection to identify anomalous behaviour in CPS as its main algorithm. This model uses IF which is an unsupervised ML ensemble. ML methods such as statistical, clustering or classification-based algorithms require the normal behaviour profile established first. Unlike other unsupervised ML methods, IF defines anomalies as few and different [32, 62] and uses isolation to determine anomalous behaviour. It does not require a profile of the normal behaviour first [62] making it a fast algorithm with low demand on memory. The IF creates an ensemble of isolation trees trained on a random data subset '$d_{max\text{-}samples}$' from the main dataset '$d_{max\text{-}samples}$' ⊂ D of the maximum number of features '$f_{max\text{-}features}$', as shown in Table VII. The IF, with several randomly created partitions, isolates the anomalies through recursive binary splitting completed by each of the created iTrees and randomly selects a split feature 'q_f' and a split value 'p_v' from the input dataset D' generating a left D_l' node and a right D_r' node until all the samples are isolated, as presented in Table VIII. The splitting required for sample isolation starts at the internal root node and terminates at the external leaf node with several internal interim nodes produced if there is a possible split remaining until the maximum path depth is reached. Accepting that the anomalies are few and different, they can be isolated such that they have a shorter path. Therefore, anomalies are isolated nearer the root of the tree while normal measurements are isolated near the leaf nodes of the formed iTree. Left and right interim nodes are created at each point that a split occurs until the final external node is reached at the point which cannot create any further nodes. A density-based approach utilising Local Outlier Factor (LOF) and a distance-based approach using SVM were added to the IF algorithm to investigate performance variations of an unsupervised multi-learner ensemble model.

Table VII
Algorithm 5: IF Forest training phase of the unsupervised learning ensemble for SPEAR Framework, based on the [62]

```
Input:  Pre-processed Dataset $D = \{x_1, …., x_n\}$,
Number of tree estimators $\varepsilon_{n\text{-}estimators}$, data sub-set $d_{max\text{-}samples}$ data sub-set
features $f_{max\text{-}features}$
Output: new dataset iTree D'
Initialise Forest
#for the number of trees
for i=1 to $\varepsilon_{n\text{-}estimators}$:
    #The maximum number of samples which represent the data sub-set
    and the maximum number of features in the data sub-set to train the
    tree
    $d_{max\text{-}samples} \leftarrow$ sample (D, $d_{max\text{-}samples}$ , $f_{max\text{-}features}$)
    Forest ← Forest ∪ iTree(D')
end for
return Forest
```

In this study, although setting the contamination level can be achieved utilising subject matter expert knowledge, the labels for the dataset are known and they are used to set the contamination level 'C_o' given the anomalous instances i_α {1, …., α}, and normal instances i_n{1, …., n} in the dataset for the ground truth and validation of results as follows:

$$C_o = i_\alpha / i_n$$

The labels are removed and not used by the algorithm for anomalous behaviour detection.

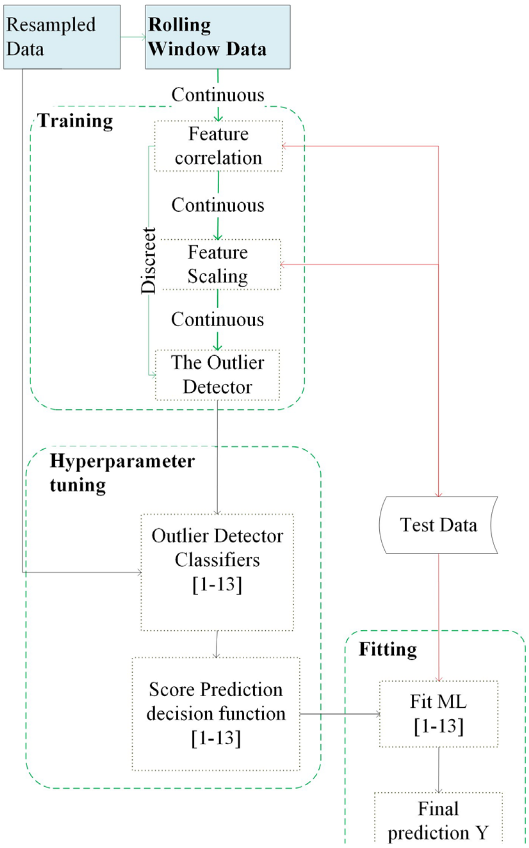

Fig. 5 Unsupervised ML, multi-learner ensemble model.

Table VIII:
Algorithm 6: IF iTree training phase of the unsupervised learning ensemble for SPEAR Framework based on the [62]

```
Input:  D'
Output: iTree
If D' cannot be split:
    external leaf node;
elif:
   let Q be D' features
   randomly select $q_f \in Q$
   randomly select a split $p_v$ between the min and max of $q_f$ in D'
  $D'_l \leftarrow$ filter (D', $q_f > p_v$)
  $D'_r \leftarrow$ filter (D', $q_f \leq p_v$)
  return interim node {Left ← iTree($D_l$), Right← iTree($D_r$),     feature
split ← $q_f$, value split ← $p_v$ }
end
```

H. The Cyber-risk Value Quantification (CRVQ) Model

The CRVQ model aims to objectively quantify the cyber-risk value utilising intelligence learnt from CPS datasets in the prevalence of anomalous behaviour such that it is trustworthy, testable, and repeatable. Our approach to quantifying the overall Cyber-risk Value (CRV_t) is inspired by the Common Vulnerability Scoring Systems (CVSS) [63]. CVSS is an open framework for communicating the severity and attributes of vulnerabilities in software that consists of base, temporal and environmental metrics.

We introduce a concept of quantifying the CRV_t for a materialised cyber-risk. This is achieved by producing initial scores for risk occurrence and risk severity impact combined with an updated score derived from the performance metrics of the detected anomalous behaviour. The metrics in the CRVQ model consist of three phases. The constant base metric group of attributes, the temporal metrics group which is expected to change over time and the environmental metrics group which is anticipated to vary between organisations and smart sectors. The attributes in the three metric groups are utilised to derive the risk occurrence score, the risk severity impact, and the safety scores.

In phase one, the anomalous behaviour is identified using ML techniques which produce a set of performance scores based on the ML's predictions. Two confidence scores are derived, the Report Confidence Accuracy (RC_A_s) and the Report Confidence Anomalous Behaviour Detection (RC_ABD_s) which are linked to the outcome of the ML models' prediction 'Y', as shown in Fig. 5. Each base-learner, y_1-y_n, detects anomalies independently using base-learners followed by a meta-learner for the final prediction. The RC_A_s is expressed as:

Anomalies in dataset + overall accuracy - (1 – anomalies accuracy)

$$A_d + A_t - (1 - A_a)$$

The RC_ABD_s is expressed as:

Anomalies in dataset + weighted F1-score - (1- anomalies F1-score)

$$A_d + F1_w - (1 - F1_a)$$

where contamination level of anomalies A_d is derived as:

Anomalous instance in the dataset / Normal instance in the dataset

$$i_{Ad} / i_{Nd}$$

and the anomalies' F1-score '$F1_A$' is derived as:

2*(Anomalies' Recall * Anomalies' Precision) / (Anomalies' Recall + Anomalies' Precision)

$$2*(R_A * P_A) / (R_A + P_A)$$

Phase two utilises the concept of the CVSS framework. In addition to a subset of CVSS attributes [63] adds traits specific to the CRVQ model in the environmental and temporal metric groups, as shown in Fig. 6. New attributes are introduced to express an actual value derived quantitatively based on the ML predictions of detected anomalies. We produce confidence scores of the occurring anomaly in combination with the prior knowledge set by the base score. The update factor is based on the actual occurrence of the cyber-risk derived from anomalous behaviour detection. The safety factor is combined with the Initial Risk Occurrence (IRO) to produce the overall risk occurrence value. The two temporal metric attributes, RC_ABDs and RC_A_s are mandatory, their values shown in Table IX. The Attack Vector (AV_b) use Network (N) and Physical (P) annotations. The Attack Complexity (AC_b) and Privilege (Pr_b) base metrics use Low (L) and High (H) annotation. The Scope utilises Unchanged (U) and Changed (C) annotation. The User Interaction (UI_b) base metric uses the None (N) and Required (R) annotations. The temporal metrics group replaces the 'Report Confidence' metric with the RC_A_s and the RC_ABD_s metrics, generated by the ML predictions, using the Confirmed (C) and Unknown (U) annotation. The values could be expressed as a combined report confidence metric; however, the aim is to report their values independently. In addition, the Collateral Damage (CD_e) uses the None (N) and Confirmed (C) annotation. The attributes in the base metrics are mandatory and not expected to change. Whereas the temporal metrics are expected to change over time across environments and are therefore used as an update factor. The metrics' dependencies between the variables are presented in Fig. 7.

We introduce the PH attributes of Possession, Utility and Authenticity in addition to the CIA as shown in Fig. 6**.** to derive the risk severity impact. A binary state is utilised for authenticity, possession, and utility metrics are covered in Table X with the following considerations:

- The Base and Environmental Metric Groups in this model use the Low (L), High (H) binary annotation. The 'Required' column in Table X shows the CVSS mandatory setting for the metric and the one in our model, respectively. The setting of zero (0) means that the rating is not present. The setting is annotated to N when it is not mandatory and annotated to Y if it is mandatory.
- The **authenticity score** (**Au_{er}**) is related to the expected normal state of operation and attribution to the source of the data. As anomalies are detected, the authenticity cannot be attributed, and operation is not considered to be in a normal state.
- The **possession score** (**P_{er}**) is related to the control of the CPS or their components producing the data. Literature refers to the physical disposition of media where the data is contained [64]. In CPS, possession can be lost if a CPS component producing the data is compromised in which case the control over the specific CPS component is considered lost. Therefore, possession is breached if the data does not reflect the status consistent with normal operation.
- The **utility score** (**U_{er}**) relates to the usefulness of the data during the normal state of CPS operations which in the presence of anomalous behaviour is considered compromised. This metric does not consider the threat actor's efforts, or the computational complexity needed to compromise the data utility. The utility score is greatest for utility compromise.

In addition, to express the **safety**, we use the CD_e attribute introduced in the environmental metric [63]. The attribute is expected to differ between organisations, as shown in Fig. 6 and Fig. 7. The attribute describes the condition related to the data in the presence of detected anomalous behaviour as not considered for safe use. Therefore, having potential consequences to the organisation such as cascading effect resulting in loss of life, damage to equipment or monetary loss.

Table IX
Metrics and attributes used to derive the risk occurrence.

Required	CVSS Metric Group & Attributes	Rating	Value
	Base		
Y→Y	Attack Vector AV_b	N/P	0.85/0.62
Y→Y	Attack Complexity AC_b	L/H	0.77/0.44
Y→Y	Privileges Required Pr_b	L/H	0.62/0.27
Y→Y	Scope S_{bi}	U/C	0.06/0.23
Y→Y	User Interaction UI_b	N/R	0.85/0.62
	Temporal		
0→Y	Report Confidence RC_ABD_e	C/U	96.00/85.00
0→Y	Report Confidence A_{cs}	C/U	96.00/85.00
0→Y	Collateral Damage CD_e	N/C	0.10/0.90

Table X
Base and Environmental Metric Groups and used to derive the Risk Severity Impact using the annotation of Low (L) and High (H) ratings.

Required	Metric Groups & Attributes	Rating	Value
	Base		
Y→Y	Confidentiality Impact C_b	L/H	0.22/0.56
Y→Y	Integrity Impact I_b	L/H	0.22/0.56
Y→Y	Availability Impact A_b	L/H	0.22/0.56
	Environmental		
N→Y	Conf. Requirement C_{er}	L/H	0.50/1.00
N→Y	Integrity Requirement I_{er}	L/H	0.50/1.00
N→Y	Availability Requirement A_{er}	L/H	0.50/1.00
0→N	Possession Impact P_{er}	L/H	0.50/1.00
0→N	Utility Impact U_{er}	L/H	0.50/1.00
0→N	Authenticity Impact Au_{er}	L/H	0.50/1.00

In phase three, the cyber-risk score is derived from the three metric groups using a BBN. In this study, the base metric group attributes' values are used as the input for the BBN as the previous knowledge to quantify the prior distribution value deriving the IRO and the Initial Risk Severity Impact (IRSI).

The Bayes theorem is utilised to derive the CRV_{ro} and the CRV_{rs} values. According to [65] the Bayes theorem can be used to derive conditional probability, where in generalised terms the probability [P] of random variables 'x' given 'y' can be expressed as:

$$P(x|y) = P(y|x) * P(x) / P(y).$$

Furthermore, according to Fig. 7, to derive the IRO the following applies:

- $P(AV_b)$ is not dependent.
- UI_b depends on AV_b, expressed as $P(UI_b)=P(UI_b|AV_b)$.
- AC_b depends on AV_b, expressed as $P(AC_b)=P(AC_b|AV_b)$.
- Pr_b conditionally depends on the AV_b and AC_b, and AV_b and AC_b are also internally dependent, thus $P(Pr_b) = P(Pr_b|Av_b,AC_b)$. If $P(Pr_b|Av_b,AC_b)$ is generalised as $P(x|y_1, y_2)$ the theorem is applied as:

$$P(x|y_1, y_2) = P(x) *P((y_1, y_2)|x) / (y_1, y_2)$$

Then

$$P(x|y_1, y_2)| = P(y_1) + P(y_2) - P(y_1) * P(y_2)$$

However, we aim to solve the challenge of deriving a value that is based on actual detected anomalous behaviour, is accurate, trustworthy and improves cyber-risk situational awareness. Therefore, the Risk Occurrence Update Factor (ROUF) and the Risk Severity Update Factor (RSUF) are produced. The ROUF is based on the ML model's performance metrics and safety factor whereas the RSUF is based on the PH Au_{er}, P_{er}, and U_{er} which are integrated into the Confidentiality (C_{er}), Integrity (I_{er}), and Availability (Av_{er}) requirement environmental metrics attributes such that combined value is used in this model for simplicity. The Cyber-risk Value of risk occurrence (CRV_{ro}) is derived as:

$$IRO = \int P\ (AC_b, AV_b, Pr_b, UI_b).$$
$$ROUF = \int P\ (RC\_ABD_s, RC\_A_s, CD_e)$$
$$CRV_{ro} = \int IRO \times ROUF$$

The Cyber-risk Value of risk severity (CRV_{rs}) is derived as:

$$IRSI = \int P\ (C_b, I_b, Av_b).$$
$$RSUF = \int P\ (C_{er}, I_{er}, Av_{er}).$$
$$CRV_{rs} = \int IRSI \times RSUF$$

The overall Cyber-risk Value CRV_t is derived as:

$$\int CRV_{ro} \times CRV_{rs}$$

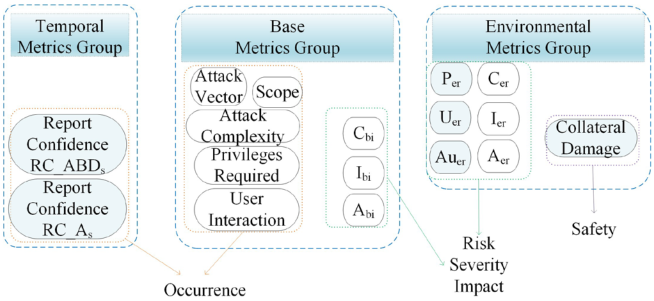

Fig. 6 The metrics used in the CRVQ model.

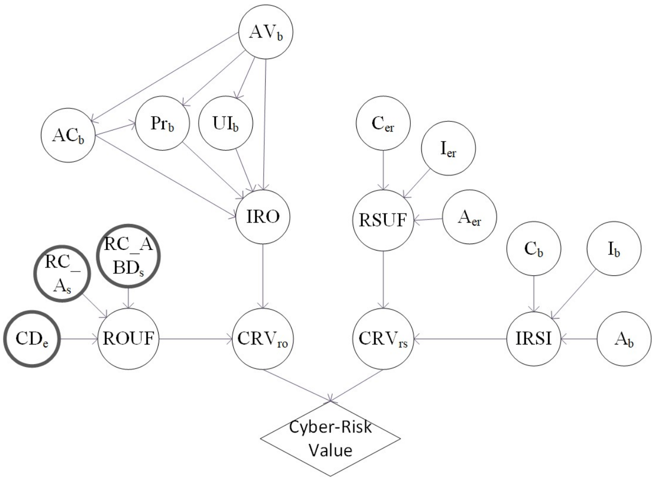

Fig. 7 BBN CRVQ model.

IV. Case Study: ICS Liquid Distribution

The case study illustrates the SPEAR framework piloted as a proof-of-concept on a simple ICS Liquid Distribution test bed [66].

A. Experiment Design for Piloting the SPEAR Framework

The experimental environment consisted of two liquid containers, two pumps, an ultrasound sensor, four discreet liquid level sensors, automated controls, and infrastructure for the data acquisition, as presented in Fig. 8. The schematic diagram shows the main tank, the positioning of the sensors and their corresponding liquid levels. Each liquid level is coupled with the decimal representation of the value that each sensor assumes based on the PLC register's binary state. The secondary tank shows the ultrasound sensor and the depth of the liquid. The liquid depth is divided into 10,000 equal segments with 0 representing a full tank and 10,000 an empty tank. Based on the discreet and the ultrasound triggers the pumps assumed ON or OFF states alternatively or in combination. This was reflected by the values recorded in the dataset. The diagram shows the registers' Least Significant Bit (LSB), the PLC registers [R2-R7] and the dataset features allocated to the bit segment [0-15] within each PLC register. The testbed functions in manual or automated modes using a touch-screen command and remote network connectivity. The pumps were activated and deactivated depending on the liquid reaching a pre-determined level. The activation of the pump which fills the main tank depended on the ultrasound sensor values. The pumps and the registers indicate the binary state of the sensors assuming two states; an ON state represented as 1.0 and an OFF state represented as 0.0. The dataset contains the corresponding decimal value of the sensors' binary state in the PLC register.

The instruments utilised in this study for experimentation to process and analyse the collected data, and train the ML models consisted of a Jupyter Notebook scikit-learn ML library [67] and a Hewlett-Packard Envy x360 x64-based Intel® Core™ i7-8565U CPU, 4 Cores 8 logical processors @ 1.8Ghz, 16GB Physical and 40GB of virtual memory.

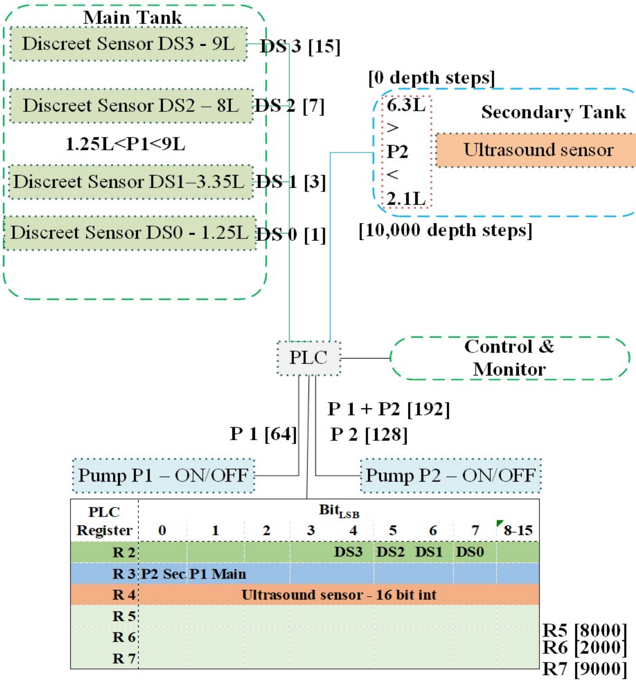

Fig. 8 The 'aNomalies' testbed schematic diagram and the structure of the data log registers

B. Introducing the Dataset

The data used in this experiment was produced from the 'aNomalies' testbed [66]. The dataset covered five operational scenarios: normal, accident, sabotage, breakdown cyber-attack, Table XI. The timestamp was presented in the format of dd/mm/yyyy hh:mm:ss.sss. A read request was sent to the PLC every 100ms. The bit segment of each PLC register according to the position of the LSB held the specific sensors' values. From the total of ten registers, three PLC registers corresponded with the values recorded in the dataset. Registers one, eight, nine and ten represented no values. Register two provided the binary state of the discreet sensors, using the first four bits of the PLC register. Register three provided the binary state for the pump using the last two bits of the PLC register. Register four recorded the value of the ultrasonic sensor as a 16-bit integer. Registers 5-7 record values but it was not clear from the dataset's description what values of these registers represented.

Table XI
Files that make up the temporal dataset [66].

File	Scenario - Type	Sensors affected	Duration [hh:mm:ss]
1	Normal	None	02:01:47
2	Plastic Bag	ultrasonic	00:33:20
3	Blocked measure 1	ultrasonic	00:00:25
4	Blocked measure 2	ultrasonic	00:00:17
5	2 floating objects in the main tan	ultrasonic	00:01:35
6	7 floating objects in the main tan	ultrasonic	00:01:22
7	Humidity	ultrasonic	00:00:18
8	Failure of a discreet sensor	Discreet 1	00:13:55
9	Failure of a discreet sensor	Discreet 2	00:03:40
10	Denial of Service attack	Network	00:01:37
11	Spoofing	Network	00:34:33
12	Wrong Connection	Network	00:15:33
13	Tank hit – with low intensity	The entire system	00:00:39
14	Tank hit – with medium intensity	The entire system	00:00:32
15	Tank hit – with high intensity	The entire system	00:00:33

C. Super Learner Ensemble's Performance Metrics

Before training the models, Spearman's correlation coefficient was used to produce a summary of the strength between the features in the combined dataset. Before training the model, highly correlated variables of at least 80% positive or negative correlation were removed from the dataset. Prior to model fitting, robust scaling standardisation was applied to tune the model. Robust scaling was a justified approach to avoid skewing the result due to the presence of instances of normal and anomalous classes in the dataset.

Despite the base-learners being trained on the same training dataset, the results were produced independently. Despite a lack of a widely accepted definition of diversity [32, 68] in classifier ensembles, ensemble base-learners are often complex making different assumptions about the prediction. A range of base-learner classifiers was used in forming the super learner including k-Neighbours (KNN), RF, Logistic Regression (LR), DT, Support Vector Classifier (SVC), AdaBoost Classifier (ABC), Extra Tree Classifier (ETC), Gaussian Naïve Bayes (GNB) and Bagging Classifier (BC). The base-learner algorithms were trained with the default parameters and with parameter optimisation, see Fig. 9. The individual base-learners do not produce a weak result, which would weaken the overall ensemble's performance. The base-learners' results vary, which is likely to improve the ensemble generalisation and produce high accuracy predictions. However, further optimisation is required, which could be an appropriate future direction to develop the model.

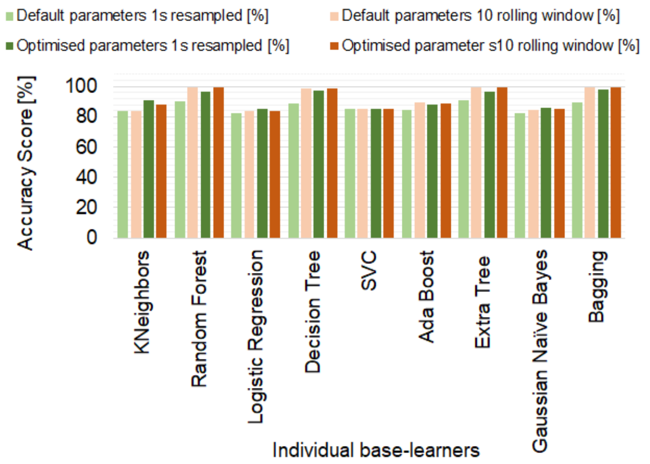

Fig. 9 Individual base-learners algorithms comparison at 1s intervals and 10s rolling window.

The overall performance of the models trained with 1s resampled interval and a 10s rolling window utilising the combined dataset is covered in Fig. 10. The details of the two best performing models are presented in Fig. 11. The difference in the performance between the two models is illustrated in Table XII. The optimisation improvement between the weakest and the best performing model is demonstrated in Table XIII. The optimisation achieved a consistent improvement in the overall F1-scores of 99.13%, an increase of 12.13% compared with the default 1s resampling rate. The most significant improvement was observed in the normal behaviour recall value by an increase of 23.46% and the anomalous behaviour precision value by an increase of 21.95% to achieve 100% in both cases. Fig. 11 covers the two best performing models. Firstly, the training time of the dataset utilising the 10s rolling window with the base-learner default values was 3m 43s. Next, the training time increased to 13m 41s with additional parameters optimisation applied to the base-learners. There are further notable differences in the testing time. Utilising the default base-learner parameters, testing time of 605ms, the attack prediction time of 9.65s were achieved. The attack prediction of individual attacks ranged from 694ms to 4.19s. Whereas following optimisation of the base-learner parameters, testing time of 703ms, the attack prediction time of 12.8s was achieved.

Further tuning was applied to the model for the individual attacks taking into consideration the imbalanced datasets. Therefore, resampling rates of 100ms, 300ms, 500ms and 1s, and 30% and 40% subset of the normal behaviour dataset in addition to the full normal behaviour dataset were applied. The performance details of the specific attacks trained with the best performing super learner are presented in Table XV and Table XVI. The supervised ML super learner's overall performance has been maintained consistently for a range of anomalies lasting between 17s and over 30m, Table XI and Table XV. The model using the 10s rolling window achieved an overall F1-score of 99.13% and in the specific anomaly cases, the model's overall F1-score remained above 97.92%. However, it was noted to be below 95%, therefore rate of >5% misclassification, in attacks 2, 3 and 6 covered in Table XI. The corresponding results are recorded in Table XV which present the values of the overall Area Under the Curve (AUC) of the Receiver Operator Curve (ROC) and in Table XVI which shows the corresponding values of the Anomaly Recall and F1-scores.

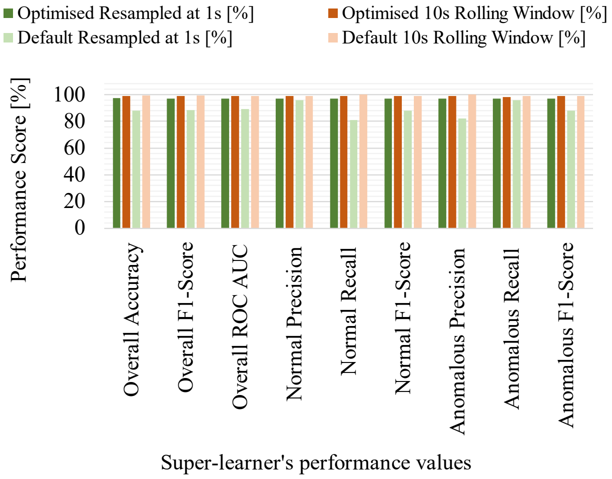

Fig. 10 Overall performance of the models trained with 1s resampling and 10s rolling window.

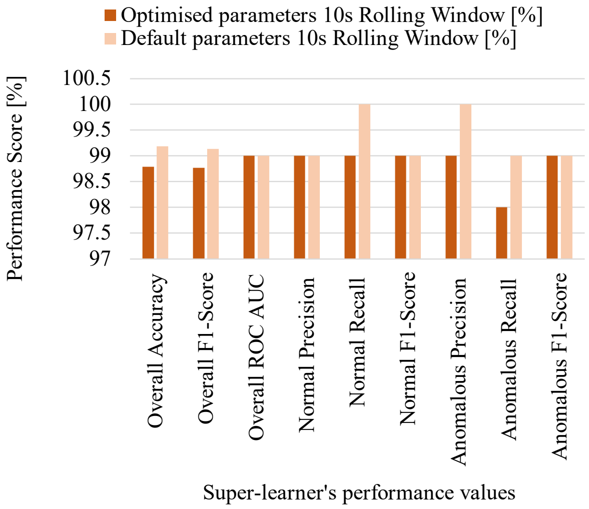

Fig. 11 The overall best performing super learner models.

Table XII
Overall performance details of the two best performing super learners and their percentage difference.

Super learner	Optimised parameters 10s Rolling Window [%]	Default parameters 10s Rolling Window [%]	Optimised to default parameters change [%]
Overall Performance [%]			
Accuracy	98.79	99.18	0.39
F1-score	98.77	99.13	0.36
ROC AUC	99.00	99.00	0.00
Normal behaviour performance [%]			
Precision	99.00	99.00	0.00
Recall	99.00	100.00	1.01
F1-score	99.00	99.00	0.00
Anomalous behaviour performance [%]			
Precision	99.00	100.00	1.01
Recall	98.00	99.00	1.02
F1-score	99.00	99.00	0.00

Table XIII
Overall performance details of the weakest and best performing super learners and their percentage difference.

Super learner	Default parameters resampled at 1s [%]	Default parameters 10s Rolling Window [%]	Optimised to default parameters change [%]
Overall Performance [%]			
Accuracy	88.14	99.18	12.53
F1-score	88.41	**99.13**	**12.13**
ROC AUC	89.00	99.00	11.24
Normal behaviour performance [%]			
Precision	96.00	99.00	3.13
Recall	81.00	**100.00**	**23.46**
F1-score	88.00	99.00	12.50
Anomalous behaviour performance [%]			
Precision	82.00	**100.00**	**21.95**
Recall	96.00	99.00	3.13
F1-score	88.00	99.00	12.50

D. Unsupervised Learners Performance Metrics

Comparatively, the unsupervised ML model was fitted using Python's scikit-learn library [67]. As part of the dataset preparation, Spearman's correlation coefficient was applied, and highly correlated features were removed. According to our framework, before fitting the model, the features were standardised utilising robust scaling applied to the 10s rolling window dataset. The comparison between the supervised and unsupervised models is based on the 10s rolling window. The one-class binary classification to detect outliers in the combined dataset and the individual attacks were trained on random subsets of the dataset. A stratified 5-fold CV was applied during the model training.

A novel IF unsupervised learning approach to outlier detection was utilised. IF detects anomalies by isolating instances and not by using distance or density measures [62]. A comparison was produced by applying a density-based approach utilising LOF [69] and a distance-based approach using SVM [70, 71]. The IF algorithm is based on the characteristics that anomalies are few and different from normal observations within datasets, therefore sensitive to isolating anomalies from the typical observations [62]. The authors [62] focused on unsupervised learning, continuous values in a non-parametric approach of multivariate data detection of anomalies only. Whereas in this study, IF was applied to parametric discreet data values of one-class binary classification for outlier detection. IF scales up to extremely large datasets with a high number of irrelevant features to solve high dimensional problems [62]. An important aspect to note about this dataset is the application of the IF learning algorithm to a dataset containing a few features and short periods of recorded anomalies shown in Table XI.

E. Deriving the CRVQ estimate

To demonstrate the CRVQ model, we define a set of input values expressed as:

Metric Group: Rating/Value,
Base (C_b:L/0.22, I_b:H/0.56,A_b:H/0.56);
Environmental (C_{er}:L/0.5, I_{er}:H/1, A_{er}:L/0.5, P_{er}:L/0.5, U_{er}:H/1, Au_{er}:H/1);
Base (AV_b:N/0.85, AC_b:L/0.77, Pr_b:L/0.62, S_{bi}:U/0.06, UI_b:R/0.62);
Temporal (RC_ABD_e:C/0.96, A_{cs}:C/0.96, CD_e:N/0.1).

To establish the ROUF, we derive values from the super learner ensemble performance metrics recorded in Table XIV, given the equations defined in III.H. If the RC_A_s and RC_ABD_s value exceed 0.96 the maximum value of 0.96 will be retained. If the RC_As and RC_ABDs values are 0.85 or below the value of 0.85 will be recorded. For example, we focus on the risk occurrence part of the model, to derive the ROUF score, utilising the ROUF normalised equations presented in III.H. The likelihood that the initially assessed cyber-risk changed based on the ML model's extracted performance metrics is 67%. increasing the cyber-risk score by 13% compared with the initial cyber-risk value.

Table XIV
Report Confidence metrics resampled to rolling 10s dataset individual learners accuracy score with repeated stratified 10fold 3 repeat CV 10s_R2-10s_R4

Metric Groups & Attributes	Rating	Update Factor Value
RC_A_s		
2	Plastic Bag	1.16
10	Denial of Service attack	1.01
11	Spoofing	1.24
RC_ABD_s		
2	Plastic Bag	0.87
10	Denial of Service attack	1.01
11	Spoofing	1.12

V. DISCUSSION

A. Comparison of the learners

As the outcomes are predicted based on input data, the ML models are dependent on the quality of the datasets. We compared the proposed ML supervised super learner method with other ML approaches proposed in the scientific literature [29, 31, 32, 58, 62]. While compared to data-driven approaches, model-based learning performs more effectively with lower computational overhead, particularly in larger datasets. However, in supervised learning, some of the efficiency is offset by the cost of the dataset preparation in features labelling. We demonstrate that supervised learning performs comparably in training and testing to the unsupervised ML algorithms in computational complexity and performance scores based on the same dataset [29].

The experiments produced promising results which are presented in Table XV and Table XVI. The performance metrics include the anomalous precision, recall, and F1-score values, and the Confusion Matrix for the True Positive (TP) (the correctly identified normal behaviour instances) and True Negative (TN) (the correctly identified anomalous instances) values for the combined and specific anomalies datasets [31, 32, 68]. The performance of each classifier is measured by using metrics that apply to multiple classifiers. The most commonly relied upon metrics are Precision measuring the likelihood of the classifier providing the correct result and Recall indicating the detection rate and F1-score [29, 68].

Variations were observed in results between the algorithms, including their performance consistency based on the level of anomalies in the datasets as shown in Fig. 12, Fig. 13 and in the algorithms' total running time which is presented in Fig. 14. Our analysis revealed that both models have a good anomaly detection ability. The supervised super learner achieved an overall F1-score of 99.13% and an anomalous recall score of 99% compared with the IF anomalous recall score of 98%. The IF anomalous recall score values achieved above 60% in the 8, 9, 13 datasets. SVM showed stronger performance where low levels of anomalous behaviour were present over a shorter period including datasets 3-7,10, 13-15 as labelled in Table XI. The respective results are presented in Table XVI and Fig. 12. The lower IF precision scores compared with the supervised ML super learner could be due to the behaviour during an attack being resemblant of normal operation hence resulting in a higher rate of FP behaviour during some of the analysed attacks.

Table XV
The Area Under ROC Curve of the individual attacks trained utilising the supervised ML super learner and the unsupervised ML algorithms with a 10s rolling window

Dataset		AUC			
Components	Anomaly [%]	Super learner	IF	SVM	LOF
All anomalies	89	0.99	0.59	0.54	0.5
Plastic_bag	27	0.88	0.58	0.54	0.5
Spoofing	28	0.96	0.52	0.54	0.5
High_blocked	3	0.98	0.92	0.91	0.5
Second_blocked	11	0.99	0.77	0.61	0.5
Bad_connection	13	0.96	0.57	0.60	0.5
DoS_attack	1	1.00	0.74	0.98	0.5
Hits_3	0.5	1.00	1.00	0.98	0.5
Wet_sensor	2	1.00	0.50	0.98	0.5
Poly_2	1.3	0.98	0.78	0.98	0.5
Poly_7	1.1	1.00	0.69	0.98	0.5
Hits_2	4	1.00	0.50	0.98	0.5
Hits_1	5	1.00	0.50	0.98	0.5
Blocked_1	4	1.00	0.50	0.98	0.5
Blocked_2	2	1.00	0.50	0.98	0.5

Table XVI
The anomalous behaviour performance metrics of the individual attacks for the supervised ML super learner and the unsupervised ML algorithms

Main & subset dataset	Algorithm	Anomaly			Confusion Matrix	
		Precision	Recall	F1	TP	TN
All anomalies	SVM:	0.92	0.1	0.18	0.99	0.1
	IF 5:	0.52	0.98	0.68	0.20	0.98
	Super:	0.99	0.99	0.99	0.99	0.99
Plastic bag	SVM:	0.50	0.12	0.19	0.97	0.12
	IF 25:	0.31	0.39	0.34	0.76	0.39
	Super:	0.94	0.77	0.85	0.99	0.77
Spoofing	SVM:	0.50	0.11	0.18	0.97	0.11
	IF 25:	0.24	0.31	0.27	0.72	0.31
	Super:	0.97	0.92	0.95	0.99	0.92
High blocked	SVM:	0.50	0.85	0.63	0.97	0.85
	IF 25:	0.81	0.84	0.82	0.99	0.84
	Super:	0.95	0.95	0.95	1.00	0.95
Second blocked	SVM:	0.50	0.24	0.33	0.97	0.24
	IF 25:	0.55	0.61	0.57	0.94	0.61
	Super:	0.99	0.99	0.99	1.00	0.99
Bad connection	IF 5:	0.22	0.25	0.24	0.89	0.25
	Super:	0.95	0.94	0.94	0.99	0.94
DoS attack	SVM:	0.26	1.00	0.42	0.96	1.00
	IF 45:	0.48	0.49	0.49	0.99	0.49
	Super:	1.00	1.00	1.00	1.00	1.00
Hits_3	SVM:	0.09	1.00	0.16	0.95	1.00
	IF 200	0.97	1.00	0.99	1.00	1.00
	Super:	1.00	1.00	1.00	1.00	1.00
Wet sensor	SVM:	0.05	1.00	0.09	0.95	1.00
	Super:	1.00	1.00	1.00	1.00	1.00
Poly_2	SVM:	0.26	1.00	0.41	0.96	1.00
	IF 35:	0.56	0.57	0.56	0.99	0.57
	Super:	0.93	0.97	0.95	1.00	0.97
Poly_7	SVM:	0.22	1.00	0.36	0.96	1.00
	IF 15:	0.46	0.39	0.42	0.99	0.39
	Super:	1.00	1.00	1.00	1.00	1.00
Hits_2	SVM:	0.09	1.00	0.16	0.95	1.00
	Super:	1.00	1.00	1.00	1.00	1.00
Hits_1	SVM:	0.11	1.00	0.19	0.95	1.00
	Super:	1.00	1.00	1.00	1.00	1.00
Blocked_1	SVM:	0.07	1.00	0.13	0.95	1.00
	Super:	1.00	1.00	1.00	1.00	1.00
Blocked_2	SVM:	0.05	1.00	0.09	0.95	1.00
	Super:	1.00	1.00	1.00	1.00	1.00

B. *Computational Complexity of the ML models*

The experiment results indicate that computational complexity and the cost of the supervised super learners is significantly higher in the combined dataset and in the individual datasets where the level of anomalies are above 25%. This complexity remains higher in datasets with anomalies' proportion of above 10% compared with the lower computational complexity and cost of the unsupervised multi-learners. The computational complexity was lower in supervised learning where fewer anomalies were prevalent and attacks lasted shorter. While the unsupervised multi-learners detected the attacks, IF did not detect all attacks, as illustrated Fig. 12 and Fig. 13, particularly where a very low occurrence of anomalies was prevalent and over a short period. In those cases, SVM produced a consistently better performance utilising the polynomial kernel which considers the input samples, their similarity, and combinations unlike IF. This could be explained by the behaviour in those datasets being similar to normal operations and was not detected as outliers without resulting in false positives.

It is important to note that no specific data sanitisation was applied [38] such as removing part of the anomalous dataset which could be considered as normal behaviour while flagged as anomalous. This could be typical during the post-attack recovery phase back to normal operation. We assert that the model should remain resilient to such behaviour and reflective of a typical operational pattern including a period of return to normal operation. Therefore, a future research direction could focus on the unsupervised model to further tune the hyperparameters constructing an unsupervised super learner. This could lead to simplified pre-processing, achieve lower model learning computational complexity and cost while consistently achieving performance at least similar to the super learner presented in our framework.

Our proposed approach produced encouraging attack prediction times ranging from 694ms to 4.19s for specific attacks and 9.65s for the combined dataset for the default base-learner parameters. The findings indicate that several factors influence the model performance. Such facets include the model structure, parameter tuning and the computational environment. However, another important challenge is the model's resilience when the data distribution evolves over time. Adapting to changes while maintaining the model efficacy in Near-Real-Time (NRT) utilising continuous data streams is critical in dealing with the time-critical nature of ICS. How to integrate effective NRT prediction and the trade-off in maintaining the model efficacy is a challenging problem that merits further research.

Addressing the first research question, it is noted in the analysis of the results, the models' resilience to detect anomalous behaviour in datasets increases by combining the learners. For example, as shown in Table XV and Table XVI, the unsupervised learning model's resilience do detect anomalies improved when multiple algorithms in addition to IF hyperparameter tuning were utilised. Similar behaviour was observed in the supervised super learner model. An individual learner within a model is not likely to outperform other learners in the stack. That said, the aim of utilising a super learner is to solve the problem of selecting a suboptimal classifier, increase the model's resilience of attack detection, improve predictive ability, and generalise the performance of the algorithm. Although the authors of the following study [7] assert a lack of agreed definition or performance metrics of the term "resilience"[72, 73], the importance of cyber resilience is acknowledged by the scientific community and governments. It is acknowledged that more must be done to improve the cyber resilience of the CNI, accepting that cyber resilience is a particular challenge in IoT [74-76]. Therefore, we argue that our approach of constructing a resilient model as a key part of the framework to detect anomalous behaviour in ICS CPS is security process-driven, the model's resilience improves the CPS' defence mechanism, security situational awareness and support for DFIR.

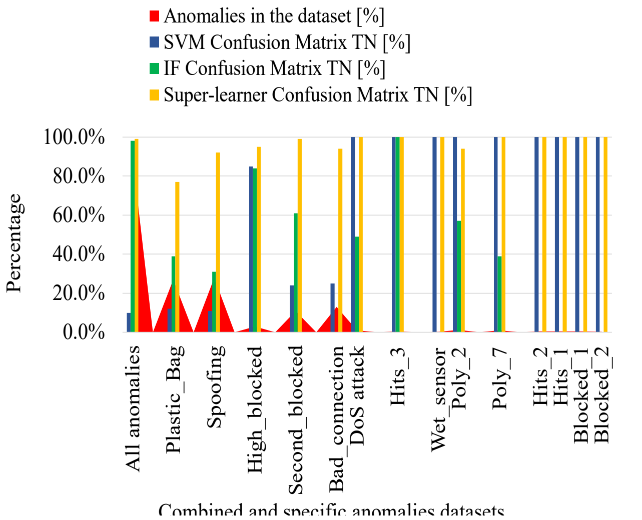

Fig. 12 Comparison of the learning algorithms' confusion matrices TN values and anomalies in the combined and specific anomalies datasets.

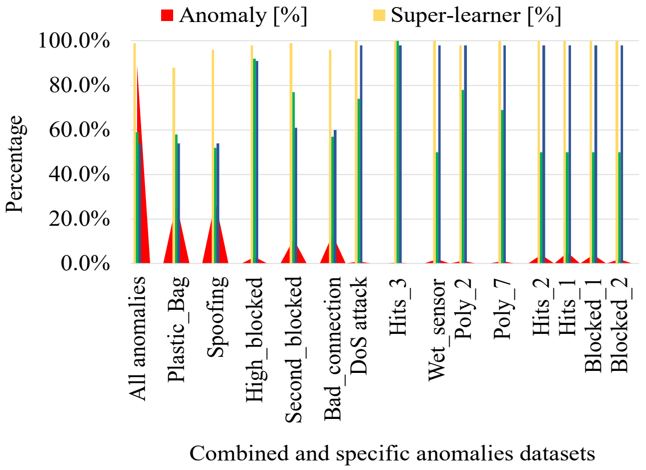

Fig. 13 AUC comparison of anomalies and algorithms in the combined and specific anomalies datasets.

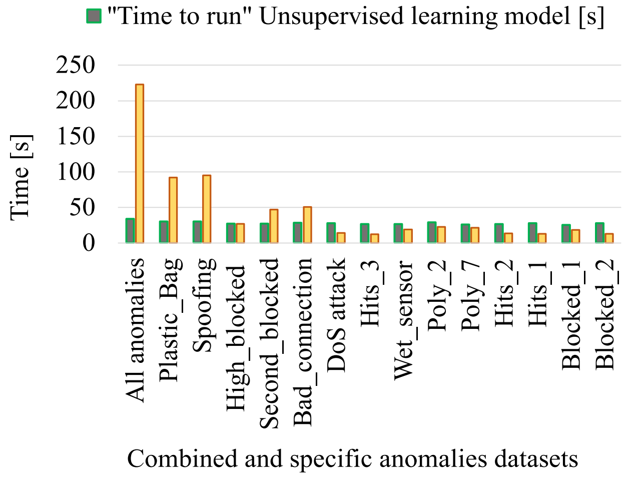

Fig. 14 Comparison of algorithms performance based on the total time to run.

C. SPEAR Framework's Application to the Cyber-risk Quantification.

We addressed the research question of how the SPEAR Framework can be utilised to quantify the cyber-risk in CPS. We have presented the CRVQ model that is an integral part of the SPEAR Framework. We outlined the structure of the CRVQ model, identified the algorithm and the performance metrics to objectively quantify the cyber-risk value in CPS. We demonstrated the model's applicability to quantifying the cyber-risk score and articulated the cyber-risk score change in the presence of anomalous behaviour. Existing risk assessment models are driven by the IS community and IT systems [77-80], whereas methodologies such as the qualitative Hazard and Operability study (HAZOP) tend to focus on risks to personnel and equipment, not cybersecurity [81]. While individual maturity exists, a disconnect remains between Information Technology (IT) and Operational Technology (OT), particularly in ICS. It is acknowledged that cybersecurity controls applicable in the IT realm are not necessarily applicable to the OT realm. Nevertheless, a holistic defence-in-depth security approach with layered protective controls which consider the converged realms is needed. We acknowledge that generalisation and scaling of the proposed CRVQ model at this stage is not possible. Further empirical studies are required to systematically investigate and report further findings from a wider pool of ICS assets to optimise the CRVQ model and its components.

That said, we assert that quantifying the cyber-risk forms an important part to enhance a robust defence-in-depth approach. We further assert that CPS sensor-generated data combined with ML anomaly detection techniques can contribute to the objective evaluation of the effectiveness of the assessed cyber-risk in CPS. Hence creating an opportunity for decision making for cybersecurity protective and corrective actions proactively. Such an approach aligns with the CREST principles of intelligence often referred to as CROSSCAT (Centralised, Responsive, Objective, Systematic, Sharing, Continuous review, Accessible, Timely) [82] to improve the cyber-threat intelligence capability creating opportunities to solve real-life problems.

D. *SPEAR Framework's Applicability to Support DFIR*

Finally, we addressed the research question concerning how the SPEAR Framework supports DFIR. Thus far we have outlined and discussed the correlation of raw data collection, information processing, generating knowledge and application of this knowledge to quantify the value of cyber-risk as part of defence-in-depth capability in CPS. To illustrate the applicability of the SPEAR Framework and its components to DFIR, we draw on comparisons with the generic Digital Forensic Research Workshop (DFRWS) and the ISO/IEC 27050:2016 standard data lifecycle phases. According to the following study [83], the stages can be broadly categorised into physical, logical and legal contexts. The physical context is concerned with capturing the data from seized physical media and maps to the identification and preservation stages, these are not the focus of this study. The logical context that is concerned with the data and information mapping to the collection, processing and analysis stages which are of particular interest to this research study, see Fig. 15. The initial risk assessment is a well-established field and is out of the scope of this research study. The SPEAR Framework's learning model introduces the capability of collecting and processing CPS datasets. Applying the learning algorithm to the data classify the anomalous behaviour from the sensor-generated datasets. Performance metrics are applied to the produced results. The results are further analysed, and the cyber-risk is quantified generating a score which can be applied to the effective evaluation of the cyber-risk impact.

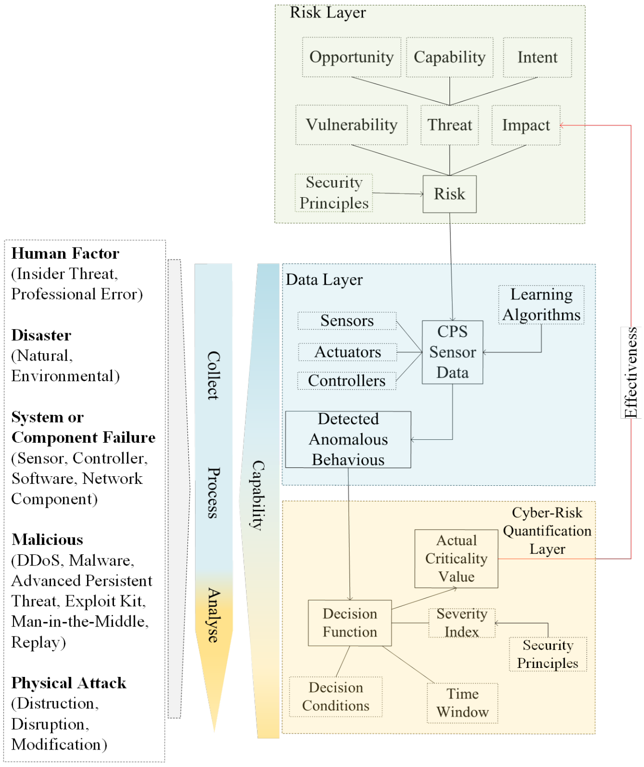

Fig. 15 The SPEAR Framework applicability to DFIR based on anomalous behaviour detection from CPS sensor data and Cyber-risk Quantification..

In ubiquitous CPS, particularly ICS and CNI seizing physical media is not always possible and innovative methods of gathering digital evidence are required. Such methods could include ML-based models, as presented in the SPEAR Framework. Collection of digital evidence could be achieved by continuously processing and analysing data from CPS in NRT, applying ML algorithms to detect anomalous behaviour as an early incident indicator. Therefore, cyber-threat intelligence and the knowledge produced from sensor data utilising the SPEAR Framework's learning algorithms could assist in the reconstruction of events and identification of prior patterns. However, consideration should be given to admissibility. For example, CPS objects can be modelled as "Digital Witnesses" (DW) to support DFIR [1, 83]. In such a case chain-of-custody need to be achieved utilising a suitable mechanism for admissibility in the Court of Law. Therefore, the SPEAR Framework has applicability to support the logical stages of the DFIR in CPS. However, it is recognised that more research is needed to understand the constraints and develop techniques that contribute to reducing the workload and cost of digital forensic investigations and generate admissible and trustworthy digital evidence.

VI. Conclusion

ICS automation and the interconnectedness of the IT and OT realms widen the attack surface. Data produced from sensors can be used to tackle anomalous behaviour detection in ICS. To address accidental and malicious activity, preventative measures are needed as part of a modern defence-in-depth approach.

We outlined and discussed the threat landscape and the evolving threat model. Next, the core components and the learning techniques that were fundamental to our framework were identified. Leveraging this knowledge, this study proposed a super learner ensemble and a hybrid unsupervised learning model for binary classification. Moreover, as part of the framework, we presented a model to objectively quantify the cyber-risk value. Utilising the super learner ensemble performance metrics, we have shown an increase in the cyber-risk score in the prevalence of detected anomalies. Additionally, we demonstrated the framework's resilience to detect anomalous behaviour in ICS datasets. This was achieved by utilising a small number of features coupled with anomalous behaviour prevalent in a short burst lasting as little as 17s and ranging from 0.5% up to 89% of the dataset. However, further prototyping and research are needed to optimise the models before the models can be standardised.

Although we demonstrated the framework within an ICS environment, the design concept of the SPEAR Framework extends to other CPS ecosystems. Our approach to the pre-processing stage and the learning algorithms were scrutinised to improve the predictive performance and the models' generalisation. Tuning of hyperparameters to optimise the algorithm was carried out. The algorithm optimisation improved the super learner's performance including the overall F1-score by 12.13% and the anomalous behaviour precision by 21.95%. The performance results of the models were analysed including the accuracy, precision, recall and F1-scores. An overall recall rate of 0.99 and 0.98 and F1-score of 0.99 and 0.68 in presence of 89% anomalies were achieved using supervised and unsupervised models, respectively. A recall rate of 1 in both cases and an F1-score of 1 and 0.99 were attained using supervised and unsupervised models respectively, where the anomalous behaviour rate was 0.5% of the total dataset. Nonetheless, we acknowledge that at this stage generalisation is not possible without further empirical research utilising broader datasets.

Furthermore, the framework's applicability to support DIFR as part of forensic readiness was scrutinised. The framework could be applied collaboratively and innovatively as part of a post-incident investigation, reconstruction of events and identification of prior patterns. The direction for future work in this area could focus on modelling the cyber-physical objects as DW to support DFIR. Finally, achieving a chain-of-custody should be considered as part of forensic readiness including the trade-off between usability and the cybersecurity principles relevant to the converged IT and OT realms in ICS.

REFERENCES

[1] G. Ahmadi-Assalemi, H. M. Al-Khateeb, G. Epiphaniou, J. Cosson, H. Jahankhani, and P. Pillai, "Federated Blockchain-Based Tracking and Liability Attribution Framework for Employees and Cyber-Physical Objects in a Smart Workplace", in 2019 IEEE 12th ICGS3. London, UK, pp. 1-9, 16-18 Jan. 2019, https://doi.org/10.1109/ICGS3.2019.8688297

[2] B. v. Lier, "The industrial internet of things and cyber security: An ecological and systemic perspective on security in digital industrial ecosystems", in 2017 21st ICSTCC. Sinaia, Romania, pp. 641-647, 19-21 Oct. 2017, https://doi.org/10.1109/ICSTCC.2017.8107108

[3] S. Schrecker, H. Soroush, J. Molina, J. LeBlanc, F. Hirsch, M. Buchheit, A. Ginter, R. Martin, H. Banavara, and S. Eswarahally, "Industrial internet of things volume G4: security framework", Ind. Internet Consort, pp. 1-173, 2016,[Online], Accessed: 28 Dec 2020, Available: https://www.iiconsortium.org/pdf/IIC_PUB_G4_V1.00_PB.pdf

[4] S. McLaughlin, C. Konstantinou, X. Wang, L. Davi, A. Sadeghi, M. Maniatakos, and R. Karri, "The Cybersecurity Landscape in Industrial Control Systems", Proceedings of the IEEE, vol. 104, no. 5, pp. 1039-1057, March 2016,[Online], https://doi.org/10.1109/JPROC.2015.2512235

[5] World Economic Forum, "The Global Risks Report 2020", 2020,[Online], Accessed: 29 Dec 2020, Available: http://www3.weforum.org/docs/WEF_Global_Risk_Report_2020.pdf

[6] F-Secure, "Attack Landscape H12019", 2019,[Online], Accessed: 29 Dec 2020, Available: https://blog-assets.f-secure.com/wp-content/uploads/2019/09/12093807/2019_attack_landscape_report.pdf

[7] G. Ahmadi-Assalemi, H. Al-Khateeb, G. Epiphaniou, and C. Maple, "Cyber Resilience and Incident Response in Smart Cities: A Systematic Literature Review", MDPI Smart Cities, vol. 3, no. 3, pp. 894-927, Aug 2020,[Online], https://doi.org/10.3390/smartcities3030046

[8] ENISA, "ENISA Threat Landscape Report 2018 15 Top Cyberthreats and Trends", 2019, Accessed: 20 Oct 2019, Available: https://www.enisa.europa.eu/publications/enisa-threat-landscape-report-2018

[9] C. Tankard, "Advanced Persistent threats and how to monitor and deter them", Network Security, vol. 2011, no. 8, pp. 16-19, Aug 2011,[Online], https://doi.org/10.1016/S1353-4858(11)70086-1

[10] F. Skopik, G. Settanni, and R. Fiedler, "A problem shared is a problem halved: A survey on the dimensions of collective cyber defense through security information sharing", Computers & Security, vol. 60, pp. 154-176, July 2016,[Online], https://doi.org/10.1016/j.cose.2016.04.003

[11] G. Settanni, Y. Shovgenya, F. Skopik, R. Graf, M. Wurzenberger, and R. Fiedler, "Correlating cyber incident information to establish situational awareness in Critical Infrastructures", in 2016 14th Annual Conf. on PST. Auckland, New Zealand, pp. 78-81, 14 Dec 2016, IEEE, https://doi.org/10.1109/PST.2016.7906940

[12] M. R. Asghar, Q. Hu, and S. Zeadally, "Cybersecurity in industrial control systems: Issues, technologies, and challenges", Computer Networks, vol. 165, pp. 106946, Dec 2019,[Online], https://doi.org/10.1016/j.comnet.2019.106946

[13] Q. Lin, S. Adepu, S. Verwer, and A. Mathur, "TABOR: A Graphical Model-based Approach for Anomaly Detection in Industrial Control Systems", in Proceedings of the 2018 on Asia Conference on Computer and Communications Security, Incheon, Republic of Korea, ACM, pp. 525-536, 2018, https://doi.org/10.1145/3196494.3196546

[14] D. F. Hsu, and D. Marinucci,"Advances in cyber security: technology, operations, and experiences", pp. 272, Oxford: Fordham University Press, ISBN: 9780823244584, 2012. [Online]. https://doi.org/10.5422/fordham/9780823244560.001.0001

[15] Q. Shafi, "Cyber Physical Systems Security: A Brief Survey", in 2012 12th ICCSA. Salvador, Brazil, pp. 146-150, 18-21 June 2012, IEEE, https://doi.org/10.1109/ICCSA.2012.36

[16] D. Gollmann, and M. Krotofil, "Cyber-Physical Systems Security", The New Codebreakers: Essays Dedicated to David Kahn on the Occasion of His 85th Birthday, pp. 195-204, Berlin, Heidelberg: Springer Berlin Heidelberg, Mar 2016, https://doi.org/10.1007/978-3-662-49301-4_14

[17] S. Berger, O. Bürger, and M. Röglinger, "Attacks on the Industrial Internet of Things – Development of a multi-layer Taxonomy", Computers & Security, vol. 93, pp. 101790, June 2020,[Online], https://doi.org/10.1016/j.cose.2020.101790

[18] S. A. Hildreth, “Congressional Research Service, Report for Congress. Cyberwarfare. ”, 19 June 2001,[Online], Accessed: 28 Dec 2020, Available: https://fas.org/sgp/crs/intel/RL30735.pdf

[19] R. Langner, “Stuxnet: Dissecting a Cyberwarfare Weapon”, IEEE Security & Privacy, vol. 9, no. 3, pp. 49-51, May 2011,[Online], https://doi.org/10.1109/MSP.2011.67

[20] Europol, “Internet Organised Crime Threat Assessment”, pp. 63, 2019,[Online], Accessed: 22 Oct 2019, Available: https://www.europol.europa.eu/activities-services/main-reports/internet-organised-crime-threat-assessment-iocta-2019

[21] SANS ICS, “Analysis of the cyber attack on the Ukrainian power grid”, Electricity Information Sharing and Analysis Center (E-ISAC), 2016,[Online], Accessed: 15 Oct 2019, Available: https://ics.sans.org/media/E-ISAC_SANS_Ukraine_DUC_5.pdf

[22] Verizon, "Data Breach Digest", 2016, Accessed: 22 Oct 2019, Available: https://enterprise.verizon.com/resources/reports/2016/data-breach-digest.pdf

[23] BBC, “Hacker tries to poison water supply of Florida city”, 2021,[Online], Accessed: 13 Feb 2021, Available: https://www.bbc.co.uk/news/world-us-canada-55989843

[24] K. Schwab,"The fourth industrial revolution", pp. 192, New York, USA, Crown Currency, ISBN: 1524758876, 2017. [Online]. Accessed: 30 Feb 2020, Available: https://books.google.co.uk/books?id=OetrDQAAQBAJ&printsec=frontcover

[25] W. Matsuda, M. Fujimoto, T. Aoyama, and T. Mitsunaga, "Cyber Security Risk Assessment on Industry 4.0 using ICS testbed with AI and Cloud", in 2019 IEEE AINS. Pulau Pinang, Malaysia, pp. 54-59, 19-21 Nov. 2019, IEEE, https://doi.org/10.1109/AINS47559.2019.8968698

[26] M. I. Jordan, and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects”, Science, vol. 349, no. 6245, pp. 255-260, July 2015,[Online], https://doi.org/10.1126/science.aaa8415

[27] Z. Zhang, and E. Sejdić, “Radiological images and machine learning: Trends, perspectives, and prospects”, Computers in Biology and Medicine, vol. 108, pp. 354-370, May 2019,[Online], https://doi.org/10.1016/j.compbiomed.2019.02.017

[28] N. A. Jalil, H. J. Hwang, and N. M. Dawi, "Machines Learning Trends, Perspectives and Prospects in Education Sector", in ICEMT 2019. Nagoya Japan, pp. 201-205, https://doi.org/10.1145/3345120.3345147

[29] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Marchetti, "On the effectiveness of machine and deep learning for cyber security", in 2018 10th CyCon. Tallinn, Estonia, pp. 371-390, 29 May-1 June 2018, IEEE, https://doi.org/10.23919/CYCON.2018.8405026

[30] J. M. Torres, C. I. Comesaña, and P. J. García-Nieto, “Machine learning techniques applied to cybersecurity”, International Journal of Machine Learning and Cybernetics, vol. 10, no. 10, pp. 2823-2836, Jan 2019,[Online], https://doi.org/10.1007/s13042-018-00906-1

[31] L. I. Kuncheva,"Combining pattern classifiers: methods and algorithms", pp. 384, Hoboken, New Jersey, USA, John Wiley & Sons, ISBN: 9781118914564, 2014. [Online]. https://doi.org/10.1002/9781118914564

[32] Z.-H. Zhou,"Ensemble methods: foundations and algorithms", pp. 236, New York, CRC press, ISBN: 9780429151095, 2012. [Online]. https://doi.org/10.1201/b12207

[33] T. G. Dietterich, “Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms”, Neural Computation, vol. 10, no. 7, pp. 1895-1923, Oct 1998,[Online], https://doi.org/10.1162/089976698300017197

[34] D. B. Araya, K. Grolinger, H. F. ElYamany, M. A. Capretz, and G. Bitsuamlak, “An ensemble learning framework for anomaly detection in building energy consumption”, Energy and Buildings, vol. 144, pp. 191-206, June 2017,[Online], https://doi.org/10.1016/j.enbuild.2017.02.058

[35] N. Moustafa, B. Turnbull, and K. R. Choo, “An Ensemble Intrusion Detection Technique based on proposed Statistical Flow Features for Protecting Network Traffic of Internet of Things”, IEEE Internet of Things Journal, pp. 1-1, Sep 2018,[Online], https://doi.org/10.1109/JIOT.2018.2871719

[36] L. H. N. Lorena, A. C. P. L. F. Carvalho, and A. C. Lorena, “Filter Feature Selection for One-Class Classification”, Journal of Intelligent & Robotic Systems, vol. 80, no. 1, pp. 227-243, Sep 2014,[Online], https://doi.org/10.1007/s10846-014-0101-2

[37] S. Ahmed, Y. Lee, S. Hyun, and I. Koo, “Unsupervised Machine Learning-Based Detection of Covert Data Integrity Assault in Smart Grid Networks Utilizing Isolation Forest”, IEEE Transactions on Information Forensics and Security, vol. 14, no. 10, pp. 2765-2777, March 2019,[Online], https://doi.org/10.1109/TIFS.2019.2902822

[38] M. Elnour, N. Meskin, K. Khan, and R. Jain, “A Dual-Isolation-Forests-Based Attack Detection Framework for Industrial Control Systems”, IEEE Access, vol. 8, pp. 36639-36651, Feb 2020,[Online], https://doi.org/10.1109/ACCESS.2020.2975066

[39] H. A. Boyes, R. Isbell, P. Norris, and T. Watson, "Enabling intelligent cities through cyber security of building information and building systems", in IET Conference on Future Intelligent Cities. London, UK, pp. 1-6, 4-5 Dec. 2014, https://doi.org/10.1049/ic.2014.0046

[40] D. B. Parker, "Toward a New Framework for Information Security?", Computer Security Handbook: John Wiley and Sons, 2015, https://doi.org/doi:10.1002/9781118851678.ch3

[41] L. Bass, R. Nord, W. Wood, and D. Zubrow, "Risk Themes Discovered through Architecture Evaluations", in 2007 WICSA'07, pp. 1-1, 6-9 Jan. 2007, IEEE, https://doi.org/10.1109/WICSA.2007.37

[42] Q. Zhang, C. Zhou, Y. Tian, N. Xiong, Y. Qin, and B. Hu, “A Fuzzy Probability Bayesian Network Approach for Dynamic Cybersecurity Risk Assessment in Industrial Control Systems”, IEEE Transactions on Industrial Informatics, vol. 14, no. 6, pp. 2497-2506, Nov 2017,[Online], https://doi.org/10.1109/TII.2017.2768998

[43] Q. Zhang, C. Zhou, N. Xiong, Y. Qin, X. Li, and S. Huang, “Multimodel-Based Incident Prediction and Risk Assessment in Dynamic Cybersecurity Protection for Industrial Control Systems”, IEEE Transactions on Systems, Man, and Cybernetics: Systems, vol. 46, no. 10, pp. 1429-1444, Dec 2015,[Online], https://doi.org/10.1109/TSMC.2015.2503399

[44] S. H. Houmb, and V. N. L. Franqueira, "Estimating ToE Risk Level Using CVSS", in 2009 ARES. Fukuoka, Japan, pp. 718-725, 16-19 March 2009, IEEE, https://doi.org/10.1109/ARES.2009.151

[45] J. Yan, M. Govindarasu, C. Liu, and U. Vaidya, "A PMU-based risk assessment framework for power control systems", in 2013 IEEE PESMG. Vancouver, BC, Canada, pp. 1-5, 21-25 July 2013, IEEE, https://doi.org/10.1109/PESMG.2013.6672731

[46] W. P. Mardyaningsih, P. H. Rusmin, and B. Rahardjo, "Anomaly Detection and Data Recovery on Mini Batch Distillation Column based Cyber Physical System", in 2019 6th EECSI. Bandung, Indonesia, pp. 454-458, 18-20 Sept. 2019, IEEE, https://doi.org/10.23919/EECSI48112.2019.8977070

[47] M. Ni, J. D. McCalley, V. Vittal, and T. Tayyib, “Online risk-based security assessment”, IEEE Transactions on Power Systems, vol. 18, no. 1, pp. 258-265, Feb 2003,[Online], https://doi.org/10.1109/TPWRS.2002.807091

[48] J. D. McCalley, A. Fouad, V. Vittal, A. Irizarry-Rivera, B. Agrawal, and R. G. Farmer, “A risk-based security index for determining operating limits in stability-limited electric power systems”, IEEE Transactions on Power Systems, vol. 12, no. 3, pp. 1210-1219, Aug 1997,[Online], https://doi.org/10.1109/59.630463

[49] I. Molloy, L. Dickens, C. Morisset, P.-C. Cheng, J. Lobo, and A. Russo, "Risk-based security decisions under uncertainty", in CODASPY '12. San Antonio Texas USA, pp. 157-168, Feb 2012, https://doi.org/10.1145/2133601.2133622

[50] H. Tsai, and Y. Huang, “An Analytic Hierarchy Process-Based Risk Assessment Method for Wireless Networks”, IEEE Transactions on Reliability, vol. 60, no. 4, pp. 801-816, Oct 2011,[Online], https://doi.org/10.1109/TR.2011.2170117

[51] W. Fu, S. Zhao, J. D. McCalley, V. Vittal, and N. Abi-Samra, “Risk assessment for special protection systems”, IEEE Transactions on Power Systems, vol. 17, no. 1, pp. 63-72, Aug 2002,[Online], https://doi.org/10.1109/59.982194

[52] K. Stine, K. Quill, and G. Witte, "Framework for improving critical infrastructure cybersecurity", National Institute of Standards and Technology, 2014, Accessed: 20/01/2021, Available: https://www.nist.gov/system/files/documents/cyberframework/cybersecurity-framework-021214.pdf

[53] N. Poolsappasit, R. Dewri, and I. Ray, “Dynamic Security Risk Management Using Bayesian Attack Graphs”, IEEE Transactions on Dependable and Secure Computing, vol. 9, no. 1, pp. 61-74, June 2011,[Online], https://doi.org/10.1109/TDSC.2011.34

[54] R. Gowland, “The accidental risk assessment methodology for industries (ARAMIS)/layer of protection analysis (LOPA) methodology: A step forward towards convergent practices in risk assessment?”, Journal of Hazardous Materials, vol. 130, no. 3, pp. 307-310, MArch 2006,[Online], https://doi.org/10.1016/j.jhazmat.2005.07.007

[55] J. Ren, I. Jenkinson, J. Wang, D. Xu, and J. Yang, “An offshore risk analysis method using fuzzy Bayesian network”, Journal of Offshore Mechanics and Arctic Engineering, vol. 131, no. 4, Sep 2009,[Online], https://doi.org/10.1115/1.3124123

[56] A. A. Cárdenas, S. Amin, Z.-S. Lin, Y.-L. Huang, C.-Y. Huang, and S. Sastry, "Attacks against process control systems: risk assessment, detection, and response", in ASIACCS '11. Hong Kong China, pp. 355-366, March 2011, https://doi.org/10.1145/1966913.1966959

[57] K. Wrona, and G. Hallingstad, “Real-time automated risk assessment in protected core networking”, Telecommunication Systems, vol. 45, no. 2-3, pp. 205-214, Jan 2010,[Online], https://doi.org/10.1007/s11235-009-9242-1

[58] M. J. van der Laan, E. C. Polley, and A. E. Hubbard, “Super Learner”, Statistical Applications in Genetics and Molecular Biology, vol. 6, no. 1, Sep 2007,[Online], https://doi.org/https://doi.org/10.2202/1544-6115.1309

[59] M. Kuhn, and K. Johnson,"Feature engineering and selection: A practical approach for predictive models", pp. 310, FL, USA, CRC Press, ISBN: 9781315108230, 2019. [Online]. https://doi.org/10.1201/9781315108230

[60] D. Kwiatkowski, P. C. Phillips, P. Schmidt, and Y. Shin, “Testing the null hypothesis of stationarity against the alternative of a unit root”, Journal of econometrics, vol. 54, no. 1-3, pp. 159-178, Dec 1992,[Online], https://doi.org/10.1016/0304-4076(92)90104-Y

[61] T. G. Dietterich, "Ensemble methods in machine learning", in Intl. workshop on MCS. Cagliari, Italy, pp. 1-15, 1 Dec 2000, Springer, https://doi.org/10.1007/3-540-45014-9_1

[62] F. T. Liu, K. M. Ting, and Z. Zhou, "Isolation Forest", in 2008 ICDM. Pisa, Italy, pp. 413-422, 15-19 Dec. 2008, IEEE, https://doi.org/10.1109/ICDM.2008.17

[63] P. Mell, K. Scarfone, and S. Romanosky, “Common Vulnerability Scoring System”, IEEE Security & Privacy, vol. 4, no. 6, pp. 85-89, Dec 2006,[Online], https://doi.org/10.1109/MSP.2006.145

[64] J. Andress,"The basics of information security: understanding the fundamentals of InfoSec in theory and practice", pp. 240, Syngress, Elsevier, ISBN: 978-0-12-800744-0, 2014. [Online]. https://doi.org/10.1016/C2013-0-18642-4

[65] T. D. Nielsen, and F. V. Jensen,"Bayesian networks and decision graphs", pp. 268, New york, USA, Springer Science & Business Media, ISBN: 978-1-4757-3502-4, 2001. [Online]. https://doi.org/10.1007/978-1-4757-3502-4

[66] P. M. Laso, D. Brosset, and J. Puentes, “Dataset of anomalies and malicious acts in a cyber-physical subsystem”, Data in Brief, vol. 14, pp. 186-191, Oct 2017,[Online], https://doi.org/10.1016/j.dib.2017.07.038

[67] G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, F. Pedregosa, and A. Mueller, “Scikit-learn: Machine Learning Without Learning the Machinery”, GetMobile: Mobile Comp. and Comm., vol. 19, no. 1, pp. 29–33, 2015,[Online], https://doi.org/10.1145/2786984.2786995

[68] M. Sokolova, and G. Lapalme, “A systematic analysis of performance measures for classification tasks”, Information processing & management, vol. 45, no. 4, pp. 427-437, July 2009,[Online], https://doi.org/10.1016/j.ipm.2009.03.002

[69] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “LOF: identifying density-based local outliers”, in Proceedings of the 2000 ACM SIGMOD Intl. Conf. on Mnmgt of data, Dallas, Texas, USA, Association for Computing Machinery, pp. 93–104, 2000, https://doi.org/10.1145/342009.335388

[70] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the Support of a High-Dimensional Distribution”, Neural Computation, vol. 13, no. 7, pp. 1443-1471, July 2001,[Online], https://doi.org/10.1162/089976601750264965

[71] D. M. J. Tax, and R. P. W. Duin, “Support Vector Data Description”, Machine Learning, vol. 54, no. 1, pp. 45-66, Jan 2004,[Online], https://doi.org/10.1023/B:MACH.0000008084.60811.49

[72] I. Friedberg, K. McLaughlin, P. Smith, and M. Wurzenberger, "Towards a Resilience Metric Framework for Cyber-Physical Systems", in 2016 ICS-CSR. Belfast, UK, 23 - 25 August 2016, https://doi.org/10.14236/ewic/ICS2016.3

[73] R. Arghandeh, A. Von Meier, L. Mehrmanesh, and L. Mili, “On the definition of cyber-physical resilience in power systems”, Renewable and Sustainable Energy Reviews, vol. 58, pp. 1060-1069, May 2016,[Online], https://doi.org/10.1016/j.rser.2015.12.193

[74] House of Lords House of Commons Joint Committee on the National Security Strategy, "Cyber Security of the UK’s Critical National Infrastructure", UK government, 2018, Accessed: 06 July 2019, Available: https://publications.parliament.uk/pa/jt201719/jtselect/jtnatsec/1708/1708.pdf

[75] Australian Cyber Security Growth Network, "Australia’s Cyber Security Sector Competitiveness Plan", Australia, 2018, Accessed: 06 July 2019, Available: https://www.austcyber.com/file-download/download/public/415

[76] HM Government, "National Cyber Security Strategy 2016-2021", 2016, Accessed: 30 Aug 2019, Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/567242/national_cyber_security_strategy_2016.pdf

[77] K. Sultan, A. En-Nouaary, and A. Hamou-Lhadj, "Catalog of Metrics for Assessing Security Risks of Software throughout the Software Development Life Cycle", in 2008 ISA. Busan, Korea (South), pp. 461-465, 24-26 April 2008, IEEE, https://doi.org/10.1109/ISA.2008.104

[78] B. Romero, M. Villegas, and M. Meza, "Simon's Intelligence Phase for Security Risk Assessment in Web Applications", in 2008 ITNG. Las Vegas, NV, USA, pp. 622-627, 7-9 April 2008, IEEE, https://doi.org/10.1109/ITNG.2008.163

[79] L. Xiao, Y. Qi, and Q. Li, "Information Security Risk Assessment Based on Analytic Hierarchy Process and Fuzzy Comprehensive", in 2008 ICRMEM. Beijing, China, pp. 404-409, 4-6 Nov. 2008, IEEE, https://doi.org/10.1109/ICRMEM.2008.71

[80] K. Clark, E. Singleton, S. Tyree, and J. Hale, “Strata-Gem: risk assessment through mission modeling”, in Proceedings of the 4th ACM workshop on Quality of protection, Alexandria, Virginia, USA, Association for Computing Machinery, pp. 51–58, 2008, https://doi.org/10.1145/1456362.1456374

[81] H. Pasman, and W. Rogers, “How can we improve HAZOP, our old work horse, and do more with its results? An overview of recent developments”, Chemical Engineering Transactions, vol. 48, pp. 829-834, 2016,[Online], https://doi.org/10.3303/CET1648139

[82] CREST, “What is Cyber Threat Intelligence and how is it used?”, 2019,[Online], Accessed: 15 Nov 2020, Available: https://www.crest-approved.org/wp-content/uploads/CREST-Cyber-Threat-Intelligence.pdf

[83] H. M. Al-Khateeb, G. Epiphaniou, and H. Daly, "Blockchain for Modern Digital Forensics: The Chain-of-Custody as a Distributed Ledger", Blockchain and Clinical Trial: Securing Patient Data, pp. 149-168, Cham: Springer International Publishing, 2019, https://doi.org/10.1007/978-3-030-11289-9_7