Review | Received 7 May 2024; Accepted 13 September 2024; Published date 23 September 2024
https://doi.org/10.55092/aias20240006

Cyberattack detection on SWaT plant industrial control systems using machine learning

Shadi Jaradat[1,*], Md Mostafizur Rahman Komol[2], Mohammed Elhenawy[2], and Naipeng Dong[1]

[1]School of Information Technology and Electrical Engineering, University of Queensland, Brisbane, Australia
[2]Centre for Accident Research and Road Safety, Queensland University of Technology, Brisbane, Australia

* Correspondence author; E-mail: shadi.jaradat@hdr.qut.edu.au.

Abstract: Detecting cyberattacks is critical for maintaining the security and integrity of industrial control systems (ICSs). This study introduces a machine learning approach for identifying cyberattacks on the Secure Water Treatment (SWaT) plant testbed. The dataset, sourced from the Singapore University of Technology and Design, includes data from 51 sensors and actuators. The research employs a Long Short-Term Memory (LSTM) network alongside traditional machine learning algorithms like Random Forest (R.F.), Support Vector Machine (SVM), and K-Nearest Neighbour (KNN) to classify cyberattacks. The LSTM model outperformed the other methods, achieving a test accuracy of 98.02% (cyberattack: 97.80%, non-attack: 98.30%). Given the imbalanced nature of the dataset, additional metrics such as precision, recall, and F1 score were evaluated, further confirming the LSTM model's robustness compared to traditional algorithms. This research demonstrates the LSTM network's effectiveness in enhancing cybersecurity for ICSs and underscores the need for proactive strategies in detecting and mitigating cyber threats.

Keywords: cyberattack detection; water treatment plant; machine learning; long short-term memory (LSTM)

1. Introduction

Cyberattacks are carried out by cybercriminals who use unauthorized access to internet-connected systems with the intent of stealing, damaging, or altering valuable information. The most prevalent threat actions include malware, ransomware, denial of service, DNS tunneling, backdoor trojan, SQL injection, zero-day exploit, and phishing. Cyberattacks often have illegal or political motives, such as espionage or intellectual property theft, targeting sectors

with high financial value and reputational stakes. Industries such as healthcare, finance, and government have been particularly affected, with significant financial repercussions. For instance, the average cost of a data breach in healthcare and financial services is $7.13 million and $5.86 million, respectively [1]. Cybercrime costs the global economy approximately $1 trillion annually, which is over 1% of global GDP [2].

Recent high-profile cyberattacks highlight the severity and range of threats. The Dark Side ransomware group disrupted the USA's Colonial Pipeline, demanding a $4.4 million ransom paid in bitcoin [3]. Similarly, a cyberattack on an Oldsmar, Florida water treatment facility temporarily raised sodium hydroxide levels to dangerous levels [4]. Another attack targeted Iran's Shahid Rajaei port, disrupting operations for three days [5].

Industrial Control Systems (ICSs) are particularly vulnerable due to their critical role in infrastructure. Protecting these systems requires rapid detection and response to cyber threats. Security Operations Centers (SOCs) and advanced technologies like Artificial Intelligence (A.I.) and machine learning are increasingly employed to enhance cybersecurity measures. A.I. techniques, especially machine learning, are pivotal in identifying and mitigating cyber threats efficiently. For instance, A.I. algorithms can analyze user and event behavior, improving anomaly detection in ICS [6].

The increasing frequency of cyberattacks on ICSs, such as water treatment plants, highlights the need for robust detection mechanisms. A notable example is the Stuxnet worm, which targeted Iran's nuclear facilities, underscoring the potential for cyberattacks to cause physical damage to critical infrastructure [7]. Similarly, a disgruntled contractor in Queensland, Australia, used radio commands to control systems at a waste plant, causing significant environmental damage [8]. These incidents demonstrate the urgent need for advanced cybersecurity measures to protect ICSs.

This study introduces a machine learning-driven method for identifying cyberattacks on the Secure Water Treatment (SWaT) testbed, an essential industrial control system (ICS). The research utilizes a Long Short-Term Memory (LSTM) network in combination with conventional machine learning techniques such as Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighbour (KNN) to categorize cyber threats. The dataset used for this analysis was obtained from the Singapore University of Technology and Design and comprises data from 51 different sensors and actuators. For time-series-based prediction, our operational data from the SWaT plant was set inherently sequential, with long-term dependencies that are vital for accurate anomaly detection. While traditional RNNs can handle sequential data as a time-series network, they often struggle with the vanishing gradient problem when working with long sequences [9]. This limitation affects their capacity to learn and retain information over extended periods, which is crucial for detecting complex and subtle cyberattack patterns that may unfold over longer durations. In contrast, LSTMs are specifically designed to tackle this issue. Their unique cell state and gating mechanisms enable them to effectively preserve information across lengthy sequences, making them better suited for capturing the temporal dependencies within the SWaT plant's operational data [10]. This capability is critical for accurately identifying and responding to potential cyberattacks, which might appear as subtle deviations from normal operational

behaviour over time. Therefore, LSTM networks is considered here as a robust RNN networks for their superior ability to capture long-term dependencies, their resistance to the vanishing gradient problem, and their proven effectiveness in time-series analysis, all of which are essential for robust cyberattack detection in Swat industrial control systems [11].

Our research contributions:

- We conducted a comprehensive classification of cyberattacks and normal conditions in water treatment plants. By employing both traditional machine learning algorithms and LSTM networks, and utilizing the SWaT dataset from the iTrust Centre for Research in Cyber Security, we were able to identify the most effective model for cyberattack classification in these critical systems.
- We significantly enhanced the usability and reliability of the SWaT dataset through extensive preprocessing. This included time zone conversion, automated labeling, handling of missing values, categorical data encoding, feature scaling, and addressing class imbalance. These improvements make the dataset more robust for machine learning applications in ICS security.
- Our feature importance analysis identified the most significant variables impacting cyberattack detection in water treatment plants. This contribution provides valuable insights for prioritizing critical sensors and actuators in real-world ICS security implementations, potentially improving the efficiency and effectiveness of cyberattack detection systems.

2. Literature review

Several studies have been conducted on a SWaT testbed dataset to identify the cyberattacks on ICSs. In the realm of cyberattack detection for ICSs like the SWaT plant, both traditional and machine learning methods have been explored. Traditional methods such as signature-based, anomaly-based, and rule-based detection systems have been widely used for threat and anomaly detection but face limitations in adaptability and effectiveness against novel threats [9].

Signature-based detection operates by matching network traffic or system behavior against predefined patterns or signatures of known threats. The system generates an alert when a match is found in the database. This approach is highly effective for identifying known threats but has significant limitations when it comes to detecting new or evolving attacks, particularly zero-day attacks, as it relies solely on previously identified patterns [12].

Anomaly-based detection systems focus on monitoring the behavior of systems or networks to detect deviations from established normal patterns. These systems typically use statistical models, machine learning algorithms, or heuristic methods to define “normal” behavior and identify anomalies. While they are effective at spotting unknown threats, they often have high false positive rates because unusual but harmless behavior may be mistakenly flagged as malicious [13,14].

Rule-based detection systems use predefined rules and heuristics, often developed from expert knowledge, to identify potentially malicious activities. These rules define specific conditions under which actions should be considered suspicious. However, the rigidity of

rule-based systems makes them less adaptable to new and evolving threats, and creating and maintaining these rules requires significant manual effort [15,16].

In contrast, machine learning approaches have shown significant promise in recent years. Supervised learning techniques have demonstrated high accuracy for both known and unknown attacks, though they require extensive labeled datasets [17]. Unsupervised learning methods excel in detecting novel attacks and offer better adaptability, but may need fine-tuning to balance false positives and negatives. Some researchers have proposed innovative approaches like the Graph Intrusion Detection (GID) framework for detecting cyberattacks on Industrial Internet-of-Things (IIoT) devices [17]. Hybrid approaches, such as the HybridRobustNet (HRN), which integrates various deep learning and machine learning algorithms, have shown improved detection accuracy and resilience against evolving attack patterns. The SWaT testbed, a scaled-down replica of a water treatment plant, has been instrumental in studying attack vectors and developing protection strategies for industrial facilities [18, 19]. Experiments using the SWaT dataset have been conducted to evaluate various detection models, including a one-class neural network supervised anomaly detection method [17]. Some studies have explored fusion approaches that combine design-based and data-driven methods to leverage the strengths of both traditional and machine learning techniques [20]. Despite the advantages of machine learning approaches, challenges remain in their implementation, particularly in terms of computational complexity and the need for large, representative datasets [19].

Recent research works have highlighted the advancements in machine learning for cyberattack detection, particularly within ICSs. Comprehensive studies have shown that machine learning techniques, such as anomaly detection and supervised learning, significantly enhance the identification and mitigation of cyber threats in ICS environments. For instance, a review by Koay *et al.* [21] discussed the current vulnerability landscape in ICSs and surveyed the advancements of machine learning-based methods, emphasizing their benefits and limitations in terms of detection accuracy and attack variety. Similarly, the study by Dehlaghi-Ghadim *et al.* [22] focused on the development of an anomaly detection dataset for ICSs, providing valuable insights into the effectiveness of machine learning techniques in identifying anomalies and improving cybersecurity measures in industrial settings. These studies underscore the importance of leveraging advanced machine learning methodologies to address the evolving landscape of cyber threats in industrial environments.

Recent studies have explored the use of automated deep learning methods to enhance intrusion detection in ICSs. For instance, synchronous optimization of parameters and architectures by genetic algorithms with convolutional neural network blocks has been proposed to secure the IIoT [23, 52]. This method optimizes both parameters and architectures, demonstrating significant improvements in detection accuracy and computational efficiency. Similarly, a differential evolution-based convolutional neural network has been designed for intrusion detection in ICSs, offering an automatic architecture design method that enhances the robustness and accuracy of intrusion detection [24]. Additionally, the use of slow-movement particle swarm optimization algorithms for scheduling security-critical tasks in resource-limited mobile edge computing has shown promising results, enabling the automatic

design of efficient and effective detection models [25]. These advancements highlight the potential of integrating genetic algorithms with deep learning techniques to address the evolving challenges in ICS security.

Kravchik and Shabtai designed a study to detect cyberattacks in ICSs using the convolutional neural network method [26]. Their study implies that dimensional convolutional neural networks are considerably smaller and faster to train; however, they can beat a more extensive coordination process when it regards anomaly detection in ICSs. Research has been developed by USA researchers to perceive the knowledge about water supply in the USA as critical infrastructure as well as the emerging technology solutions to improve the security of this vital infrastructure [27]. Improved detection rates, higher detection accuracy, and affordable processing and telecommunication costs can all be achieved through the adoption of innovative methods such as various machine learning algorithms. So, researchers have tried to build precise security mechanisms for the system's security using a machine learning-based intrusion detection system (IDS), which is software that monitors a computer system or network for contagious activity and policy violations [28]. A group of researchers from India suggested using a cloud-based machine learning technique to classify an attack into a cloud-based machine learning platform because conventional machine learning techniques are unable to support huge dataset processing efficiently. Using the NSL KDD Cup99 dataset, the authors describe an attack categorization methodology where the prototype is trained using Microsoft's Azure Machine Learning (Azure ML) platform and is focused on the Multiclass Decision Forest Machine Learning Algorithm [29]. Some researchers analyzed the way of adversarial learning to approach supervised models by producing competitive stimuli using the Jacobian-based Saliency Map attack and examining categorization characteristics, and they focused on how such inputs can allow controlled models with adversarial training to be more reliable. In their paper, they have used an authentic power system dataset to carry out the experiment [30]. Researchers from Turkey used eight machine learning techniques in their approaches to display cybercrimes in two separate models and estimate the effect of the specified parameters on the identification of the cyberattack technique and the offender. They also stated that findings from their experiment could demonstrate that the likelihood of a cyberattack lowers as the victim's education and wealth level rises, and their developed model will help to detect cyberattacks and make the struggle against them easier and more successful [31]. Researchers from South Korea experimented with the process of collecting data based on several incidents which have been accumulated from five small and medium companies. Moreover, for the categorization of events and prioritized-based activities, they constructed a model to utilize text mining approaches such as n-gram, bag-of-words, and machine learning algorithms [32]. A paper by UAE and Jordan researchers elaborated on the machine learning and deep learning techniques implementation in the cybersecurity area, as well as these two techniques' effects, drawbacks, and future in cybercrime-related areas [33]. Researchers from Tunisia and Saudi Arabia conducted a survey to assume cybersecurity attacks using deep learning techniques, where they provided a set of novel LSTM, RNN (Recurrent Neural Network), and MLP (Multilayer Perceptron) based forecasting models that use a lately accessible dataset called CTF to prevent the sort

of violence that could emerge. They also added in their paper that the model yielded promising results when CTF was used, particularly for the LSTM model, which had an F1 score larger than 93% [34]. Furthermore, Saudi Arabian scientists assessed several machine learning techniques, used a new dataset called Bot-IoT to identify IoT network assaults fast and efficiently, and then applied seven several machine learning algorithms in the implementation phase [35]. Several researchers from Japan analyzed various features to identify which feature would be the best to detect cyberattacks quickly; eventually, they identified ten major features which are the most prominent. Additionally, they mentioned two programs, SVM and PCA (Principal Component Analysis), which is employed to select the feature, and R.F., which helps to build the classifier [36]. USA and Canadian researchers published a paper where they suggested a stacked autoencoder (SAE) oriented deep learning system to produce machine-learned attributes against transmission SCADA threats to supplement more high-quality features for machine learning-based threat monitoring. In their paper, they have provided machine-learned features which significantly enabled more accurate detection against SCADA breaches in power transmission systems, according to simulations using data from an elevated smart grid testbed [37]. A group of South African researchers tried to answer specific research questions and illustrate current trends, difficulties, risks, and responsibilities in the water utility industry in terms of cybersecurity. In their paper, they featured the major technical guidelines presented by the Water Information Sharing and Analysis Centre (Water ISAC) to assist water services and vital infrastructure stakeholders in blocking cyberwarfare strikes to contribute significantly by displaying crucial features for attempting to influence effective information security deployment in the water utilities [38]. Researchers from Bangladesh and Australia advanced a brief about cybersecurity data science in their paper, where data were acquired from relevant cybersecurity resources, and they used algorithms to augment the newest statistics methods in order to provide further security management strategies. In order to analyze cybersecurity, they built a multi-layered machine learning-based framework [39]. A paper published by researchers from China, Saudi Arabia, Algeria and France explained the feasibility of a deep learning model as well as the functional relevance of the cyberattack detection phase by using the R.F. procedure, which enables quite computationally efficient models to identify suspicious invasions of privacy which have already overcome customary IDSs and to recognize phishing software affecting SCADA systems. They mentioned in their paper about using authentic data sources from standard laboratory SCADA systems, a two-line three-bus power transmission system, and a gas pipeline to assess the proposed method's efficiency, which surpassed standalone deep learning algorithms and state-of-the-art algorithms such as KNN, R.F., Naive Bayes, Adaboost, SVM, and OneR [32]. In research papers, some researchers compared and contrasted several machine learning algorithms as well as explained the practical consequences of implementing machine learning systems as a replacement to traditional power technical systems in their paper [40–46]. Researchers from the USA proposed using machine learning-based security mechanisms in order to detect the presence and define the characteristics of cyberattacks efficiently, and they suggested retrieval actions that should be implemented after detecting cyberattacks to minimize the implications

and proposed a machine learning-based provincial renovation methodology to provide approximate provincial monitoring system based on completely fabricated government assessments. They ensured in their paper that before appropriate measurement, if results are installed back online, then it will provide consistent functioning of the system [45,46]. Several researchers proposed employing four well-known machine learning techniques: C4.5, Bayesian Network (BayesNet), Decision Table (D.T.), and Naive Bayes to predict which host will be attacked based on historical data. BayesNet estimates the overall generalization ability of 91.68%, according to controlled research [47].

In this research, we classify cyberattacks in the SWaT plant using a time-series LSTM network and traditional machine learning algorithms KNN, SVM, and R.F. We have considered the cyberattack dataset of 51 sensors and actuator information for cyberattack classification. Automated tuning of LSTM model hyperparameters is performed in this research to achieve accurate model performance. Also, traditional machine learning models are tuned, and the result is compared with the LSTM model result to identify the best-performing model for real-time implementation. We also compare other performance metrics like precision, recall, and F1 score among different models to evaluate the performance of models for imbalanced dataset samples between attack and non-attack data.

3. Data summary and pre-processing

Data collection started at the SWaT testbed on 20 July 2019 at 12:33 PM (GMT+8) and lasted for four hours. The plant's normal run without attacks started at 12:35 PM to 02:50 PM. The plant faced six attacks that were launched between 03:08 PM and 04:16 PM. The plant's stop time was at 4:35 PM. The total number of samples collected is 14,997 for 51 sensors and actuators [48]. The dataset collected contains 14,997 samples; 1987 samples represent attack data, and 13,011 samples are normal data.

We cleaned the timestamp column in the original dataset and converted the time zone to GMT+8; we used a Python script to perform the task. The dataset is unlabeled, so we also labeled the dataset based on the start time and end time of the attacks. A new column was added to the dataset that represents the target (label), and we used a Python script to automate the labeling process. We also handled the missing values and encoded the categorical data using the one-hot encoding (OHE). Also, we scaled the input features into a (0, 1) ratio based on the mean and variance of the training data. Finally, we split the dataset into three subsets for training, evaluation, and testing. The dataset is imbalanced, so for more reliable classification accuracy results, we unsampled the attack data (minority class) using the scikit-learn resample library. We also performed a feature importance analysis to find the most significant features that impact the target variable.

4. Methodology

4.1. Conventional machine learning-based (SVM)

Our SVM model was implemented using the scikit-learn library; we used the SVM algorithms along with cross-validation to train the model; the input features were fed to the model in the traditional way, and all input data points (features), which equal to the number of sensors, at single time t were fed to the model. The values of the hyperparameters are shown in Table 1 shows. The highest accuracy achieved using this model was 82.8%. Table 2 shows different accuracy metrics.

Table 1. SVM hyperparameters.

Model	Hyperparameter	Value Range	Optimal Value
SVM	Kernel = RBF	RBF, Poly, Linear	RBF
	C	[0.1, 1, 10, 100, 1000]	100
	γ (kernel coefficient)	[1, 0.1, 0.01, 0.001, 0.0001]	0.001

Table 2. SVM performance.

Model	Accuracy	Precision	Recall	F1 score
SVM	82.8%	81%	86%	83%

4.2. Conventional machine learning-based (KNN)

The second approach we employed was the sliding window technique, as illustrated in Figure 1. Given the time-series nature of the data, we applied a sliding window approach with a specified length w. Input features were transformed into vectors of length w. Each sliding window was labeled based on a threshold value: if the number of log entries indicating an attack exceeded half of the entries, the window was labeled as an attack; otherwise, it was labeled as normal.

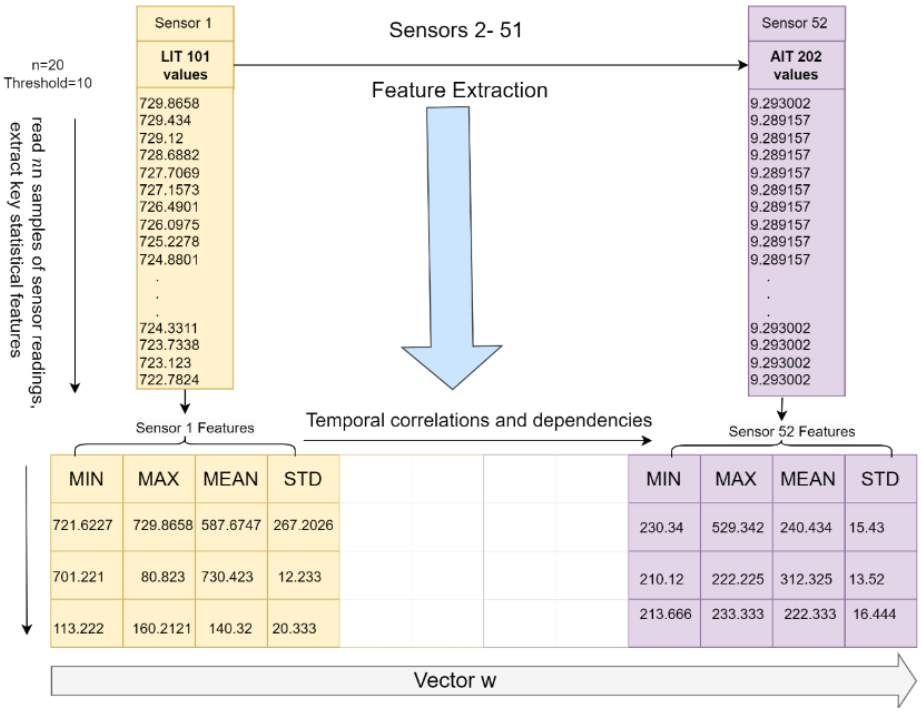

Figure 1. KNN rolling window model.

We used statistical metrics—minimum, maximum, mean, and standard deviation—for feature engineering and extraction. These new input features were then fed into the model for training. Hyperparameters k, window size w, and threshold t were optimized using the grid search approach, as shown in Figure 2. The tuned hyperparameters are summarized in Table 3. The KNN model achieved a 93% accuracy using the rolling window technique, which is significantly higher than the accuracy achieved with the SVM approach. Table 4 displays various performance metrics for the KNN model.

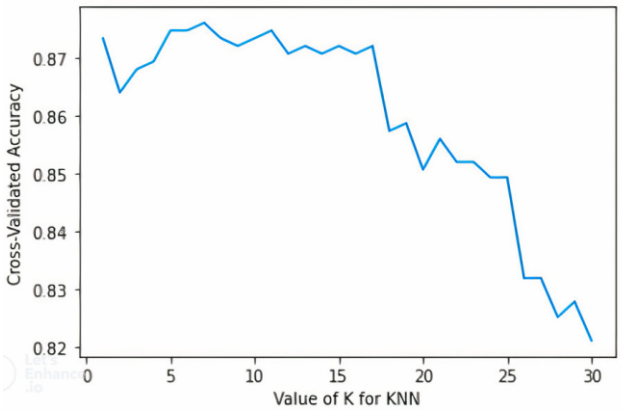

Figure 2. Hyperparameter k tuning.

During the hyperparameter tuning phase, we evaluated new hyperparameter values for the KNN model, specifically window size w and threshold t. For instance, setting the window size to 50 and the threshold to 25 means each window consists of 50 data points, and it is labeled as an attack if it contains 25 or more attack data points. Different values for window size and threshold resulted in varying levels of accuracy. Generally, reducing the www value

increased accuracy but lowered precision and F1 score, indicating overfitting. Therefore, we selected trade-off values for w and t that achieved the best balance between accuracy and F1 score.

The ROC (receiver operating characteristic) curve for both SVM and KNN indicates that the KNN Rolling Window approach achieved higher accuracy. This improvement is attributed to the consideration of the periodic nature of the data, similar to the LSTM approach. Figure 3 illustrates the ROC curve for both SVM and KNN models.

Table 3. KNN hyperparameters.

Model	Hyperparameter	Value Range	Best Value
KNN	k	[1, 2, 3, …, 28, 29, 30]	7
	w (window size)	[10, 20, 30, 40, 50, 60]	20
	t (threshold)	[5, 10, 15, 20, 25, 30]	10

Table 4. KNN Rolling Window performance

Model	Accuracy	Precision	Recall	F1 score
KNN Rolling Window	93%	69%	85%	76%

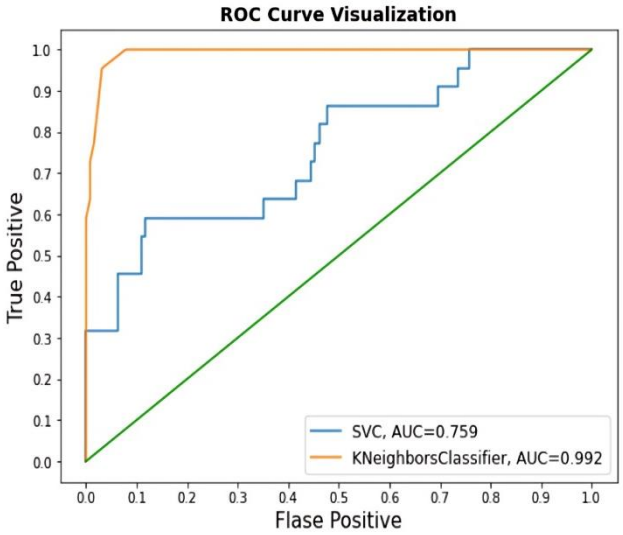

Figure 3. ROC curve.

4.3. LSTM-based

Our LSTM model was implemented using the Pytorch framework. We sequenced the data based on the time sequence window. It is challenging to guess accurate sequence dimensions, batch size, layer number, and other hyperparameters. Here, the random search hyperparameter optimization approach was used to adjust model hyperparameters. To optimize the random search hyperparameter, we use the Optuna Python package with our Pytorch-based coding framework [49]. The optimal hyperparameter values for the LSTM model were determined through a systematic random grid search approach. We defined a range of values for each hyperparameter, including hidden layer size (15–70), layer dimension (1–5), batch size (15–400), dropout (0.1–0.9), sequence length (10, 20, 40, 60),

number of epochs (10–70), and learning rate (0.001–0.01). Each trial's hyperparameter values were stored and assigned a trial identifier. To train each model, we set 100 random search trials with a 2-hour duration. The trial with the greatest validation accuracy indicates its hyperparameter values as the best-performing hyperparameter value combination once all trials have been completed. A random grid of all possible combinations of these hyperparameter values was created, and the LSTM model was trained and evaluated on the training and validation data for each configuration. The performance metric (e.g., accuracy, F1 score) was tracked for each configuration, and the combination that yielded the best performance was identified as the optimal hyperparameter values. This process was repeated for a specified number of iterations or until a stopping criterion was met, resulting in the best hyperparameter values presented in Table 5: hidden layer size of 30, layer dimension of 5, batch size of 47, dropout of 0.17044, sequence length of 20, 59 epochs, and a learning rate of 0.00676. This systematic approach to hyperparameter tuning allowed us to explore a wide range of configurations and identify the optimal settings for the LSTM model to achieve the highest performance on the given dataset. Hyperparameters are shown in Table 5. The architecture of the LSTM model, used for classifying cyberattacks in the SWaT dataset, is shown in Figure 4.

Table 5. Random grid of hyperparameter values for LSTM hyperparameter tuning.

Hyperparameter	Value Range	Best Value
Hidden layer size	15–70	30
Layer dimension	1–5	5
Batch size	15–400	47
Dropout	0.1–0.9	0.17044
Sequence	10, 20, 40, 60	20
n-epochs	10–70	59
Learning rate	0.001–0.01	0.00676

We further adjusted our learning rate by implementing a cyclical learning rate tuning method, which has recently proven to be highly effective [50]. This alternative approach to learning rate tuning explores a range of values between defined boundary rates. After initially setting the learning rate using random hyperparameter optimization, it is fine-tuned through the cyclical learning rate method. This process enhances precision in selecting the learning rate, a crucial hyperparameter for finding global optima in gradient descent. Additionally, determining the number of epochs is critical in neural network modeling, as too many epochs can push the loss function gradient beyond its global optima, leading to overfitting. Conversely, too few epochs can cause underfitting by not fully reaching the global optima. Both overfitting and underfitting hinder proper model optimization and, consequently, its performance on test data. We applied an early stopping criterion of 35 epochs [51] to fine-tune the number of epochs. If the model's validation accuracy does not improve over 35 consecutive epochs, training stops, and the epoch with the highest validation accuracy is selected as the optimal one. This approach allows the random grid search to run more

efficiently by quickly discarding trials with redundant epoch numbers and fine-tuning the epoch count more accurately.

The input features for training the LSTM models, using data from all 51 sensors and actuators in the SWaT dataset, are fed into a five-layer LSTM model with a batch size of 47. The first layer's input feature count is 51 (representing all sensor and actuator features), with 30 hidden neurons, which are then passed forward as input to the subsequent layers, each also using 30 hidden neurons. The dataset, recorded at 60 Hz, produces 60 data points per second. Hyperparameter tuning identified 20 as the optimal sequence length, meaning we used a dataset of 0.33 seconds as a single package. Consequently, the input tensor dimension is [20,47,51], where 47 represents the batch size, 20 is the sequence length, and 51 is the number of input features. The hidden dimension is [20,30,47], with 30 representing the number of hidden neurons.

The LSTM layers utilize the ReLU activation function, and the hidden neurons have a dropout rate of 0.17. The models are optimized using the Adam optimizer. The final layer is a fully connected (F.C.) layer, which outputs the values Y1 and Y2 corresponding to the two output classes: attack and non-attack. The probabilities of an event being an attack or not are calculated using the logarithm of the SoftMax function on the output values.

Table 6 shows details of the accuracy metrics of LSTM model. The overall accuracy of the LSTM model was 98.02%.

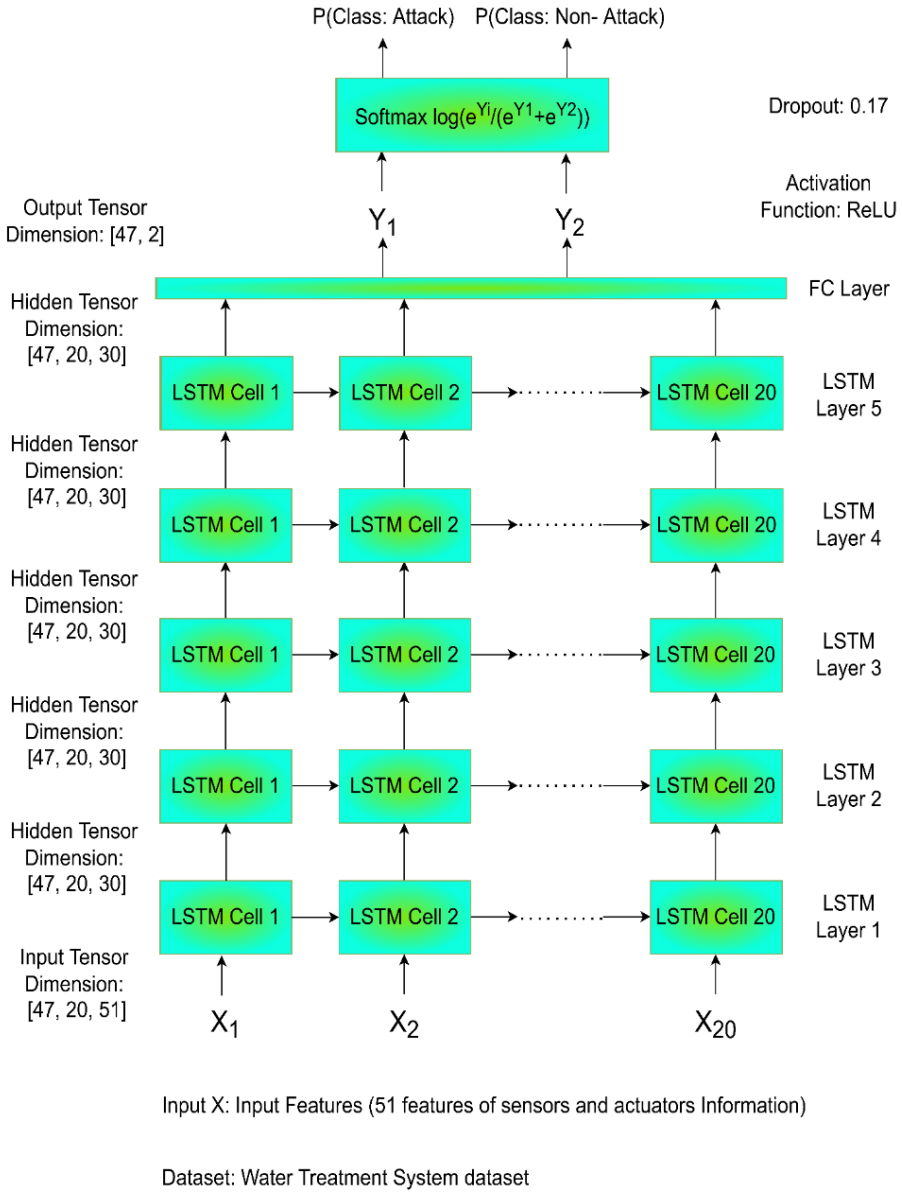

Figure 4. LSTM model architecture on classifying cyberattacks using SWaT dataset.

Table 6. LSTM accuracy by class.

Model	Overall Accuracy	Accuracy by Class	
LSTM	98.02%	Normal	Attack
		97.8%	98.3%

5. Comparison of model performances

To identify the best-performing algorithm, we need to use evaluation metrics. Since we have an imbalanced dataset, accuracy is not enough to identify the best model. Therefore, we used other metrics such as precision, recall, F1 score, and AUC values.

In this paper, we implemented machine learning and deep learning algorithms to detect cyberattacks on the SWaT testbed datasets. We used three powerful models, KNN, SVM, and LSTM, to train and build the classifiers. The hyperparameters were tuned as described

in the previous section, and the performance of LSTM outperformed SVM and KNN models. Figure 5 shows the LSTM model performance, and Figure 6 shows a comparison between the traditional machine learning models and LSTM model with respect to the accuracy, precision, recall, and F1 score.

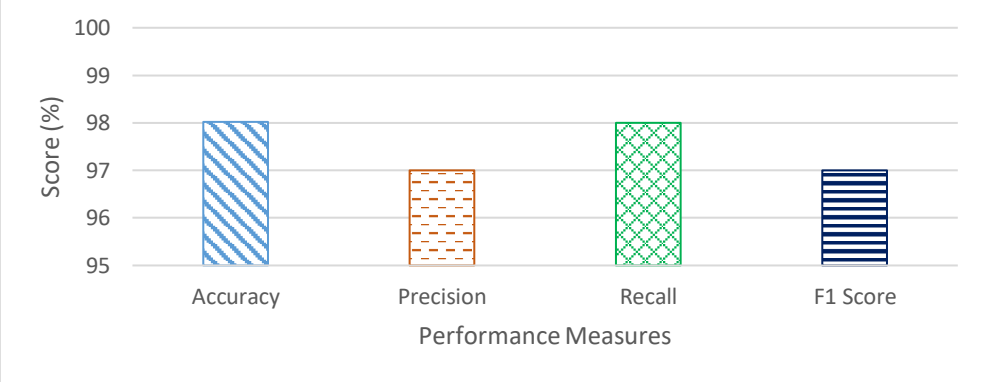

Figure 5. Performance measures of LSTM model.

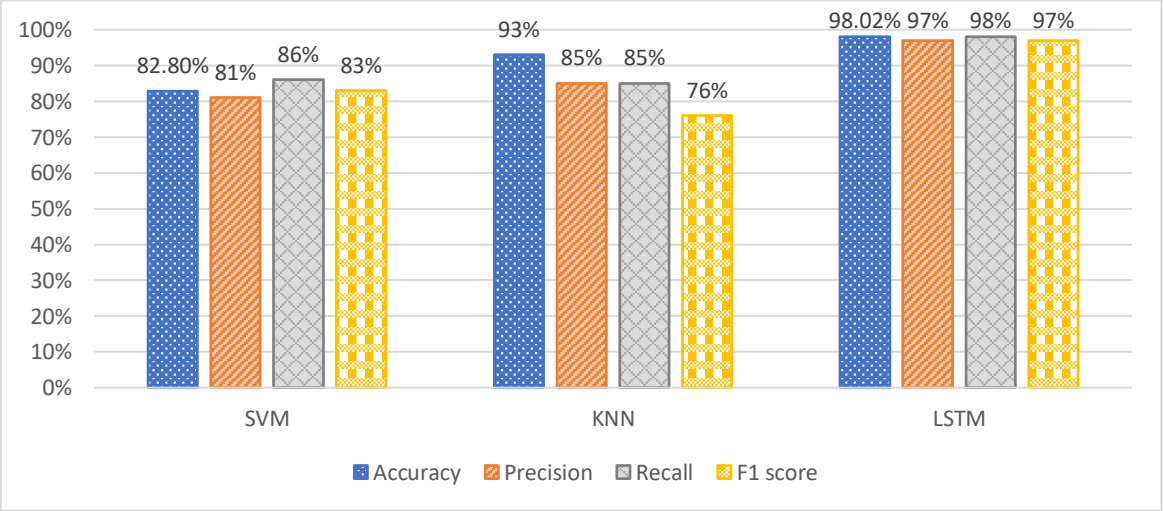

Figure 6. Performance comparison between traditional machine learning and LSTM model.

After analyzing Figure 6, we find that LSTM outperforms the KNN and SVM model's accuracy (82.80% and 93%, respectively) and achieved 98.02% accuracy. As the dataset is imbalanced, where there are much fewer attack data samples in comparison to regular data samples, we analyze other performance measures: precision, recall, and F1 score to evaluate models' performance with imbalanced samples. The precision score of LSTM (97%) is found to be higher than the precision score of SVM (81%) and KNN (85%). This indicates that the number of accurately classified attack samples in comparison to the total number of attack samples is higher with the LSTM model than with the other traditional models. Recall score, in another name, the sensitivity score of LSTM score (98%), is found to be higher than the recall score of SVM (86%) and KNN (85%). This shows that the number of accurately classified attack samples in comparison to the total number of samples classified as an attack is also higher with the LSTM model than with other traditional models. The F1 score of

LSTM score (97%) is also found to be higher than the precision score of SVM (83%) and KNN (76%).

Building upon the performance analysis of KNN, SVM, and LSTM models, it is crucial to consider additional factors that influence model selection for cyberattack detection in ICSs. While LSTM demonstrates superior performance across all metrics, we must also evaluate the models' computational requirements, interpretability, and adaptability to evolving threats.

LSTM's superior performance can be attributed to its ability to capture temporal dependencies in time-series data, which is particularly relevant for the sequential nature of ICS operations. This advantage allows LSTM to detect subtle anomalies that might indicate a cyberattack, even when the attack pattern evolves over time. However, LSTM models typically require more computational resources and longer training times compared to traditional machine learning algorithms like KNN and SVM.

On the other hand, KNN and SVM offer advantages in terms of interpretability and lower computational overhead. These characteristics can be crucial in resource-constrained environments or when rapid deployment is necessary. SVM, in particular, shows promising results with 93% accuracy, suggesting it could be a viable alternative in scenarios where the additional performance gain of LSTM does not justify the increased complexity.

Considering the critical nature of water treatment plants and the potential consequences of undetected cyberattacks, the significant performance improvement offered by LSTM (98.02% accuracy) over KNN (82.80%) and SVM (93%) makes it the recommended choice for this application. The higher precision, recall, and F1 scores of LSTM further solidify its position as the most suitable model, especially given the imbalanced nature of the dataset. However, it is important to note that model selection should always consider the specific constraints and requirements of the deployment environment. In scenarios where computational resources are limited or real-time performance is critical, a carefully tuned SVM model might provide a reasonable trade-off between accuracy and efficiency.

6. Discussion and future research

Our comparative analysis underscores the effectiveness of machine learning techniques, particularly LSTM networks, in detecting cyberattacks within industrial control systems (ICSs) like water treatment plants. The superior performance of LSTM, especially when compared to traditional algorithms such as KNN and SVM, highlights the importance of capturing temporal dependencies in time-series data for accurate anomaly detection.

LSTM's high accuracy, precision, recall, and F1 scores (98.02%, 97%, 98%, and 97%, respectively) demonstrate its capability to effectively distinguish between normal operations and attack scenarios, even when dealing with imbalanced datasets. This is a crucial advantage, as real-world ICSs often have skewed distributions of attack and normal data samples.

However, performance metrics alone should not dictate the choice of model. Considerations such as computational complexity, interpretability, and adaptability to evolving threats are also important. While LSTM delivers the best overall performance, its

higher computational demands and longer training times might be challenging in resource-limited environments or when rapid deployment is required.

On the other hand, KNN and SVM offer lower computational costs and greater interpretability, making them suitable alternatives in certain cases. SVM, in particular, achieved a commendable 93% accuracy, suggesting that it could be a viable option when the additional performance benefits of LSTM do not justify the increased complexity.

It's crucial to recognise that the effectiveness of these models can vary depending on the specific characteristics of the ICS, the nature of the cyberattacks, and the quality of the training data. Continuous monitoring and evaluation are essential to ensure the chosen model remains effective against evolving threats.

Future research should explore ensemble methods that leverage the strengths of different machine learning algorithms to strike a balance between performance, computational efficiency, and adaptability. Additionally, incorporating attack information, domain expertise, and feedback during model training could further improve the accuracy and interpretability of the cyberattack detection system. The study can be expanded in several ways: by implementing other machine learning and deep learning models to compare and evaluate accuracy, by further refining the accuracy of the applied models, by developing the models using real-life ICS datasets, and by extending AI. capabilities to identify which sensor has been attacked.

There are three primary limitations in this study: the lack of publicly available ICS datasets, the use of artificially injected attacks rather than real-life ones that might have different characteristics, and the limited coverage of attack types and scenarios in the available dataset.

In short, this study highlights the potential of machine learning techniques, particularly LSTM networks, in tackling the growing threat of cyberattacks on ICSs. While LSTM stands out as the recommended choice for water treatment plants based on our findings, the final decision should be guided by the specific requirements and constraints of the deployment environment. Continuous improvement and adaptation will be vital to maintain the effectiveness of these systems against evolving cyber threats.

7. Conclusion

ICSs are essential to a wide range of industries, including nuclear power, energy, and water treatment. In an era of increasing cyberattacks, securing these systems has become a critical challenge. This research demonstrated that machine learning, particularly LSTM networks, can substantially improve cyberattack detection compared to traditional methods like KNN and SVM. Our LSTM models, trained on real-time SWaT testbed data, showed superior performance with an accuracy of 98.02%, precision of 97%, recall of 98%, and an F1 score of 97%. These results highlight the model's effectiveness in distinguishing between attack and non-attack conditions, providing a robust and reliable approach for ICS cybersecurity.

Looking ahead, integrating advanced models like Transformer networks could further improve cyberattack classification accuracy by capturing complex temporal dependencies,

enhancing performance in real-world ICS environments. Additionally, applying transfer learning could allow models trained in one domain to adapt to various industrial sectors with minimal retraining, boosting both efficiency and scalability. Future research should focus on validating model performance across diverse cyberattack datasets that capture evolving and recent attack patterns from various industries, ensuring robust and adaptable solutions for ICS applications. Furthermore, implementing proactive cybersecurity strategies such as real-time anomaly detection and predictive analytics could help anticipate vulnerabilities before they are exploited. Hybrid models combining machine learning with domain-specific knowledge may also offer more effective and interpretable solutions, ensuring ICSs remain resilient against evolving threats.

Acknowledgments

We would like to extend our sincere gratitude to the Singapore University of Technology and Design for providing the dataset used in this study.

Conflicts of interests

The authors declare no conflict of interests.

Authors' contribution

Conceptualization, S.J. and N.D.; methodology, S.J.; software, S.J.; validation, S.J., N.D., and M.E.; formal analysis, S.J.; investigation, S.J.; resources, N.D.; data curation, S.J.; writing—original draft preparation, S.J.; writing—review and editing, N.D.; visualization, S.J.; supervision, N.D.; project administration, N.D.; funding acquisition, N.D. All authors have read and agreed to the published version of the manuscript.

References

[1] Parachute. 2022 Cyber attack statistics, data, and trends. Available: https://parachutetechs.com/2022-cyber-attack-statistics-data-and-trends/ (accessed on 7 May 2024).
[2] Kaspersky. What is cyber security? Available: https://www.kaspersky.com.au/resource-center/definitions/what-is-cyber-security (accessed on 7 May 2024).
[3] Touro College Illinois. The 10 biggest ransomware attacks of 2021: Recent Cyber Attacks Hit Infrastructure and Critical Facilities Across the U.S. Available: https://illinois.touro.edu/news/the-10-biggest-ransomware-attacks-of-2021.php (accessed on 7 May 2024).
[4] Kardon S. Florida water treatment plant hit with cyber attack. Available: https://www.industrialdefender.com/florida-water-treatment-plant-cyber-attack/ (accessed on 7 May 2024).
[5] USIP. Israel-Iran cyber war, gas station attack. Available: https://iranprimer.usip.org/blog/2021/nov/02/israel-iran-cyber-war-gas-station-attack (accessed on 7 May 2024).
[6] Segal E. A.I. applications in cybersecurity with real-life examples. Available: https://www.altexsoft.com/blog/ai-cybersecurity/ (accessed on 7 May 2024).
[7] Chabin T. How to protect nuclear power plants against cyber terrorist attacks? Available: https://teodorchabin.com/2019/01/12/nuclear-cyber-security/ (accessed on 7 May 2024).

[8] Dawda S, MacColl J. Water plant suffers cyber attack through the front door. Available: https://rusi.org/explore-our-research/publications/commentary/us-water-plant-suffers-cyber-attack-through-front-door (accessed on 7 May 2024).
[9] Ghojogh B, Ghodsi A. Recurrent neural networks and long short-term memory networks: Tutorial and survey. *arXiv* 2023, arXiv:2304.11461.
[10] Shiri FM, Perumal T, Mustapha N, Mohamed R. A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU. *arXiv* 2023, arXiv:2305.17473.
[11] Cahuantzi R, Chen X, Güttel S. A comparison of LSTM and GRU networks for learning symbolic sequences. In *Science and Information Conference*, London, United Kingdom, 13–14 July 2023, pp. 771–785.
[12] Modi C, Patel D, Borisaniya B, Patel H, Patel A, *et al.* A survey of intrusion detection techniques in cloud computing environment. *J. Network Comput. Appl.* 2013, 36(1):42–57.
[13] García-Teodoro P, Díaz-Verdejo J, Maciá-Fernández G, Vázquez E. Anomaly-based network intrusion detection: Techniques, systems and challenges. *Comput. Secur.* 2009, 28(1–2):18–28.
[14] Iturbe Araya JI, Rifà-Pous H. Anomaly-based cyberattacks detection for smart homes: A systematic literature review. *Internet Things* 2023, 22:100792.
[15] Liu Q, Hagenmeyer V, Keller HB. A review of rule learning-based intrusion detection systems and their prospects in smart grids. *IEEE Access* 2021, 9:29641-29660.
[16] Behzadi A, Sadrizadeh S. A rule-based energy management strategy for a low-temperature solar/wind-driven heating system optimized by the machine learning-assisted grey wolf approach. *Energy Convers. Manage.* 2023, 277:116590.
[17] Karacayılmaz G, Artuner H. A novel approach detection for IIoT attacks via artificial intelligence. *Cluster Comput.* 2024, 27:10467–10485.
[18] Kapersky. SWaT Testbed. Available: https://mlad.kaspersky.com/swat-testbed/ (accessed on 7 May 2024).
[19] MR GR, Ahmed CM, Mathur A. Machine learning for intrusion detection in industrial control systems: challenges and lessons from experimental evaluation. *Cybersecur.* 2021, 4:27.
[20] MR GR, Mathur A. Fusing design and machine learning for anomaly detection in water treatment plants. *Electronics* 2024, 13(12):2267.
[21] Koay AMY, Ko RKL, Hettema H, Radke K. Machine learning in industrial control system (ICS) security: Current landscape, opportunities and challenges. *J. Intell. Inf. Syst.* 2023, 60:377–405.
[22] Dehlaghi-Ghadim A, Moghadam MH, Balador A, Hansson H. Anomaly detection dataset for industrial control systems. *IEEE Access* 2023, 11:107982–107996.
[23] Huang JC, Zeng GQ, Geng GG, Weng J, Lu KD. 2023. SOPA-GA-CNN: Synchronous optimisation of parameters and architectures by genetic algorithms with convolutional neural network blocks for securing industrial Internet-of-Things. *IET Cyber-Sys. Robot.* 2023, 5(1):e12085.
[24] Huang JC, Zeng GQ, Geng GG, Weng J, Lu KD, *et al.* Differential evolution-based convolutional neural networks: An automatic architecture design method for intrusion detection in industrial control systems. *Comput. Secur.* 2023, 132:103310.
[25] Zhang Y, Liu Y, Zhou J, Sun J, Li K. Slow-movement particle swarm optimization algorithms for scheduling security-critical tasks in resource-limited mobile edge computing. *Future Gener. Comput. Syst.* 2020, 112:148–161.
[26] Kravchik M, Shabtai A. Detecting cyber attacks in industrial control systems using convolutional neural networks. In *Proceedings of the 2018 Workshop on Cyber-Physical Systems Security and PrivaCy*, Toronto, Canada, 15–19 October 2018, pp. 72–83.

[27] Clark RM, Panguluri S, Nelson TD, Wyman RP. Protecting drinking water utilities from cyberthreats. *J. Am. Water Works Assn.* 2017, 109(2):50–58.
[28] Mironeanu C, Archip A, Amarandei CM, Craus M. Experimental cyber attack detection framework. *Electronics* 2021, 10(14):1682.
[29] Ge M, Syed NF, Fu X, Baig Z, Robles-Kelly A. Towards a deep learning-driven intrusion detection approach for Internet of Things. *Comput. Networks* 2021, 186:107784.
[30] Anthi E, Williams L, Rhode M, Burnap P, Wedgbury A. Adversarial attacks on machine learning cybersecurity defences in Industrial Control Systems. *J. Inf. Secur. Appl.* 2021, 58:102717.
[31] Bilen A, Özer AB. Cyber-attack method and perpetrator prediction using machine learning algorithms. *PeerJ Comput. Sci.* 2021, 7:e475.
[32] Mohasseb A, Aziz B, Jung J, Lee J. Predicting cybersecurity incidents using machine learning algorithms: A case study of Korean SMEs. In *Proceedings of the 5th International Conference on Information Systems Security and Privacy (ICISSP)*, Prague, Czech Republic, 23–25 February 2019, pp. 230–237.
[33] Salloum SA, Alshurideh M, Elnagar A, Shaalan K. Machine learning and deep learning techniques for cybersecurity: A review. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)*, Cairo, Egypt, 8–10 April 2020, pp. 50–57.
[34] Ben Fredj O, Mihoub A, Krichen M, Cheikhrouhou O, Derhab A. Cybersecurity attack prediction: A deeplLearning approach. In *13th International Conference on Security of Information and Networks*, Merkez, Turkey, 4–7 November 2020, pp. 1–6.
[35] Alsamiri J, K. Alsubhi K. Internet of things cyber attacks setection using machine learning. *Int. J. Adv. Comput. Sci. Appl.* 2019, 10(12):627–634.
[36] Feng Y, Akiyama H, Lu L, Sakurai K. Feature selection for machine learning-based early detection of distributed cyber attacks. In *2018 IEEE 16th International Conference on Dependable, Autonomic and Secure Computing, 16th International Conference on Pervasive Intelligence and Computing, 4th International Conference on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech)*, Athens, Greece, 12–15 August 2018, pp. 173–180.
[37] Wilson D, Tang Y, Yan J, Lu Z. Deep learning-aided cyber-attack detection in power transmission systems. In *2018 IEEE Power & Energy Society General Meeting (PESGM)*, Portland, OR, USA, 5–10 August 2018, pp. 1–5.
[38] Alabi M, Telukdarie A, van Rensburg NJ. Cybersecurity and water utilities: Factors for influencing effective cybersecurity implementation in water sector. In *ASEM 41st International Annual Conference Proceedings*, 2020.
[39] Sarker IH, Kayes ASM, Badsha S, Alqahtani H, Watters P, *et al*. Cybersecurity data science: an overview from machine learning perspective. *J. Big Data* 2020, 7(1):41.
[40] Wang W, Harrou F, Bouyeddou B, Senouci SM, Sun Y. A stacked deep learning approach to cyber-attacks detection in industrial systems: Application to power system and gas pipeline systems. *Cluster Comput.* 2021, 25:561–578.
[41] Jaradat S, Nayak R, Paz A, Ashqar HI, Elhenawy M. Multitask learning for crash analysis: A Fine-Tuned LLM framework using twitter data. *Smart Cities* 2024, 7(5):2422–2465.
[42] Koay AM, Ko RK, Hettema H, Radke K. Machine learning in Industrial Control System (ICS) security: Current landscape opportunities and challenges. *J. Intell. Inf. Syst.* 2023, 60(2):377–405.
[43] Hink RCB, Beaver JM, Buckner MA, Morris T, Adhikari U, *et al.* Machine learning for power system disturbance and cyber-attack discrimination. In *2014 7th International*

Symposium on Resilient Control Systems (ISRCS), Denver, CO, USA, 19–21 August 2014, pp. 1–8.
[44] Elhenawy M, Komol MMR, Masoud M, Liu SQ, Ashqar HI, *et al.* A novel crowdsourcing model for micro-mobility ride-sharing systems. *Sensors* 2021, 21(14):4636.
[45] Jaradat S, Nayak R, Paz A, Elhenawy M. Ensemble learning with Pre-Trained transformers for crash sseverity classification: A deep NLP approach. *Algorithms* 2024, 17(7):284.
[46] Wu Z, Chen S, Rincon D, Christofides PD. Post cyber-attack state reconstruction for nonlinear processes using machine learning. *Chem. Eng. Res. Des.* 2020, 159:248–261.
[47] Nanda S, Zafari F, DeCusatis C, Wedaa E, Yang B. Predicting network attack patterns in SDN using machine learning approach. In *2016 IEEE Conference on Network Function Virtualization and Software Defined Networks (NFV-SDN)*, Palo Alto, CA, USA, 7–10 November 2016, pp. 167–172.
[48] Goh J, Adepu S, Junejo KN, Mathur A. A dataset to support research in the design of secure water treatment systems. In *Critical Information Infrastructures Security*, Paris, France, 10–12 October 2016, pp. 88–99.
[49] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage, AK, USA, 4–8 August 2019, pp. 2623–2631.
[50] Smith LN. Cyclical learning rates for training neural networks. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Santa Rosa, CA, USA, 24–31 March 2017, pp. 464–472.
[51] Prechelt L. Early stopping — But when? In *Neural Networks: Tricks of the Trade*. 2nd ed. Berlin: Springer, 2012. pp. 53–67.