Article:
Ayodeji, A., Di Buono, A., Pierce, I. et al. (2024) Wavy-attention network for real-time cyber-attack detection in a small modular pressurized water reactor digital control system. Nuclear Engineering and Design, 424. 113277. ISSN: 0029-5493

https://doi.org/10.1016/j.nucengdes.2024.113277

Nuclear Engineering and Design 424 (2024) 113277

Contents lists available at ScienceDirect

Nuclear Engineering and Design

journal homepage: www.elsevier.com/locate/nucengdes

Wavy-attention network for real-time cyber-attack detection in a small modular pressurized water reactor digital control system

Abiodun Ayodeji [a,*], Antonio Di Buono [b], Iestyn Pierce [c], Hafiz Ahmed [d]

[a] Brunel Innovation Centre, Brunel University London, Uxbridge UB8 3PH, UK
[b] National Nuclear Laboratory, Central Laboratory, Sellafield, Seascale CA20 1PG, UK
[c] School of Computer Science and Electronic Engineering, Bangor University, Bangor, Gwynedd LL57 1UT, UK
[d] Nuclear AMRC Midlands, University of Sheffield, Derby DE73 5SS, UK

ARTICLE INFO

Keywords:
Cybersecurity
Small modular reactor
Deep learning
Industrial control system
Intrusion detection system
Artificial Intelligence

ABSTRACT

Global interest in advanced reactors has been reignited by recent investments in small modular reactors and micro-reactor design. The use of digital devices is essential for meeting the size and modularity requirements of small modular reactor controls. By fully digitizing the small modular reactor control systems, critical information can be obtained to optimize control, reduce costs, and extend the reactor's lifetime. However, the potential for cyber-attacks on digital devices leaves digital control systems vulnerable. To address this risk, this study presents a novel wavy-attention network for sensor attack detection in nuclear plants. The wavy-attention network comprises stacks of batch-normalized, dilated, one-dimensional convolution neural networks, and sequential self-attention modules, superior to conventional single-layer networks on sequence classification tasks. To evaluate the proposed wavy-attention network architecture, the International Atomic Energy Agency's Asherah Nuclear Simulator and a false data injection toolbox found in the literature, both implemented in MATLAB/SIMULINK, are utilized. This approach leverages changes in process measurements to identify and classify cyber-attacks on priority signals using the proposed wavy-attention network. Three false data injection attacks are simulated on the simulator's pressure, temperature, and level sensors to obtain representative process measurements. The wavy-attention network is trained and validated with normal and compromised process variables obtained from the simulator. The performance of the wavy-attention network to discriminate between the reactor states using the test set shows 99% accuracy, as opposed to other baseline models such as vanilla convolution neural networks, long short-term memory networks, and bi-directional long short-term memory networks with 90%, 77%, and 91% accuracy, respectively. An ablation study is also conducted to test the contribution of each component of the proposed architecture. The theoretical framework of the proposed wavy-attention network and its implementation for nuclear reactor digital sensor attack detection are discussed in this paper.

List of Abbreviations

Abbreviation	Definition
SMR	Small Modular Reactor
PLC	Programmable Logic Controller
WAN	Wavy Attention Network
PWR	Pressurized Water Reactor
IAEA	International Atomic Energy Agency
ANS	Asherah Nuclear Simulator
FDI	False Data Injection
HFIA	High-Frequency Injection Attack
HSMIA	High Slope Measurement Injection Attack
RCA	Restart Communication Attack
CNN	Convolution Neural Network
LSTM	Long Short-Term Memory
Bi-LSTM	Bidirectional Long Short-Term Memory
vCNN	Vanilla Convolution Neural Network
ICS	Industrial Control System
OPC-UA	Open Platform Communications Unified Architecture

* Corresponding author.
E-mail address: ayod_abe@yahoo.com (A. Ayodeji).

https://doi.org/10.1016/j.nucengdes.2024.113277
Received 9 December 2023; Received in revised form 27 April 2024; Accepted 29 April 2024
Available online 3 May 2024

1. Introduction

Recent investments in the design and development of small and micro-scale reactors indicate a major nuclear renaissance could be on the horizon. These advanced reactor designs aim to address some of the key challenges associated with traditional large-scale nuclear power plants, such as high capital costs, long construction times, and concerns over safety and waste management. Small modular reactors (SMRs) and micro-reactors are designed to be more flexible, scalable, and potentially safer than their larger counterparts. They can be manufactured in a factory setting and transported to desired locations, reducing on-site construction work and costs. Additionally, their smaller size and modular design allow for incremental capacity additions, which could make them more attractive to utilities and investors.

The development of these smaller reactors is driven by the need for reliable, carbon-free baseload power to complement intermittent renewable energy sources. However, there are still significant technical and safety hurdles to overcome before these advanced reactor designs can be widely deployed. Firstly, while SMRs offer exciting potential benefits, their increased reliance on digital instrumentation and control systems and potential remote operation introduces new cyber security risks that must be carefully addressed. Unlike traditional analog systems, modern digital controls utilize software, networks, and remote connectivity that could potentially be exploited by adversaries. Malicious cyber attacks against SMR systems could have severe consequences, ranging from unplanned shutdowns and equipment damage to environmental releases that threaten public safety.

Secondly, the decentralized nature of SMR development and the proposed autonomous operation may also expose it to advanced persistent threats and other threat actors. Industrial control systems, including those used in nuclear reactors, face threats from cyberattacks aimed at compromising digital instruments. In fact, Stuxnet, the first major cyber weapon, was aimed at a nuclear facility's digital control system. As SMR designs move toward higher levels of automation and remote operation, proactively designing effective cybersecurity measures for these new reactor technologies from the outset will be crucial to ensuring their safe and secure operation.

Advanced reactors mostly rely on cyber-physical sensors that are vulnerable to cyber-attacks on system frequencies (Aamoth et al., 2022). One of the most potent potential cyber-attacks for SMRs is false data injection attacks, (FDI), where adversaries introduce subtle corruptions to sensor measurements or other signals (Sundaram et al., 2022). Researchers have analyzed two scenarios in which threat actors could leverage false data injection attacks against advanced reactors specifically (Li et al., 2018). The first scenario involves attackers falsifying sensor outputs to give the appearance of abnormal conditions, potentially triggering unsafe control actions by the system. This kind of attack could potentially impact reactor priority signals that initiate reactor trips when critical safety parameters, such as neutron flux, coolant temperature, or pressure, exceed predefined limits. In an FDI attack targeting SMRs, an adversary could compromise the digital instrumentation and control systems to inject false data into the priority signals. For example, they could manipulate the reactor trip signal, preventing it from being triggered even when critical safety parameters are exceeded, potentially leading to an uncontrolled or unsafe reactor state. Similarly, they could alter the emergency safety features actuation signal, disabling crucial safety systems from activating during an accident scenario.

The second scenario focuses on hackers manipulating the control logic itself to alter the system's response to actual events or falsify process data without activating the conventional anomaly detection techniques (Li et al., 2018). FDI attacks on power control signals could cause the reactor to operate at unsafe power levels, risking damage to the core or other components. Corrupted coolant system signals could mislead operators about the actual cooling conditions, potentially resulting in inadequate cooling and overheating. False radiation monitoring signals could mask the presence of radioactive leaks or conceal abnormal radiation levels, posing health and environmental risks. For the attack described in these scenarios to be successful, especially for an advanced persistent threat, there has to be an effective pre-attack stage referred to as a reconnaissance phase, where attackers gather valuable information about the target systems and their vulnerabilities. In the context of SMRs and advanced reactors, attackers would try to identify and map the various networks, systems, and components that make up the SMR's instrumentation and control infrastructure, including the communication protocols and data flows. The reconnaissance phase could also involve gathering information about the normal operating parameters, sensor readings, and control logic of the reactor systems. This knowledge would aid the attacker in crafting realistic false data that can bypass existing detection mechanisms.

The consequences of successful FDI attacks on priority signals in SMRs could be catastrophic, ranging from equipment damage and unplanned shutdowns to potential radiological releases and severe accidents. Detecting and mitigating such sophisticated cyber attacks is a significant challenge, as the corrupted data may appear legitimate and within normal operating ranges, making it difficult to discern from genuine sensor readings, especially during the reconnaissance phase. Therefore, it is crucial to develop effective real-time cyber-attack detection methods that can accurately identify and mitigate potential threats. Moreover, incorporating cybersecurity measures into the design of an SMR's control system is crucial to prevent expensive retrofits, avoid unnecessary delays, and minimize disruptions to plant operations.

Towards a resilient Industrial Control System (ICS), several data-driven network monitoring, intrusion detection, and protection systems have been proposed to address potential cyber-attacks on ICS (Ayodeji et al., 2020). Specifically, neural networks have been successful in identifying patterns and anomalies in complex and dynamic systems, such as nuclear plant abnormal event diagnosis (Lee et al., 2021), electric gate valves (Liu et al., 2020), and nuclear safety assessment (Ayodeji et al., 2022). Neural networks have also been used to detect attacks in critical industrial control systems (Nedeljkovic and Jakovljevic, 2021).

The state-of-the-art deep learning methods offer different comparative advantages. For instance, the simple and easy-to-implement feed-forward network is good for problems with fixed-size input and output, while Convolution Neural Networks (CNN) are effective for grid-like data such as images due to spatial hierarchies. Autoencoders can compress data into a lower-dimensional space, useful in anomaly detection and data denoising, while Long-Short Term Memory (LSTM) networks can handle sequences of variable lengths, beneficial in tasks such as time series prediction, natural language processing, and speech recognition. However, the conventional deep learning model architecture is sub-optimal in terms of performance when applied to detect subtle, yet potentially dangerous attacks captured in network logs or dynamic and time-varying sequences in process measurement.

In recent years, attention mechanisms have been used to improve the predictive performance of deep learning models. The attention layers, common in transformer architecture, allow for parallel processing of sequences and are effective in capturing long-range dependencies and context-aware representations. Unlike other approaches, the attention mechanism can model the dependencies between the target output and the input sequences. Attention modules enable deep learning models to handle variable-length sequences, which is useful in many time-series and sequential tasks (Ayodeji et al., 2109). However, many existing attention-based models are mostly designed for large language models, and the capability of attention-based models to learn attack patterns in nuclear plant control signals has not been extensively explored. This is largely a result of the lack of data that represents the behavior of process variables under cyber-attack (Ayodeji et al., 2020).

This work builds upon our previous conference paper (Ayodeji et al., 2023), where we initially proposed the wavy-attention network (WAN) method for cyber-attack detection task. In this paper, we discuss the

performance of the proposed novel WAN for real-time cyber-attack detection in a Pressurized Water Reactor (PWR) digital control system. The WAN is designed to capture the dynamic and periodic nature of the system by introducing a sequential self-attention layer, in a wavy architecture. This mechanism allows the model to extract important temporal and frequency features from the system's signals and use them to detect cyber-attacks. Because the real-world log data from a nuclear plant is not available, and the process measurement could indicate a cyber-attack (Ayodeji et al., 2020), the WAN model is trained using the process data obtained from the Asherah Nuclear Simulator (ANS). The ANS, implemented in MATLAB/SIMULINK, is used to simulate the operating conditions of the plant in a steady state and under attack. First, High-Frequency Injection Attack (HFIA), High Slope Measurement Attack (HSMIA), and Restart Communication Attack (RCA) are simulated in the pressurizer pressure signal, the pressurizer level signal, and the reactor coolant system mean coolant temperature sensor signals respectively. These sensors are critical to ensure the safe operation of the reactor control system, and compromising these safety–critical sensors may have dangerous consequences. The simulated cyber-attacks are adopted from the false data injection toolbox, developed by Potluri et al. (Potluri et al., 2019). The resulting process variables that reflect the effect of the cyber-attack on the reactor are used to train, validate, and test the WAN classifier.

This work significantly extends our previous conference publication (Ayodeji et al., 2023) by providing a comprehensive analysis and evaluation of the proposed WAN for cyber-attack detection in SMR digital control systems. The key additions in this work are as follows:

1. We present the complete graph and description of the proposed WAN model (in Fig. 2), enumerating all layer parameters to facilitate reproducible implementation. (Section 3.1).
2. We simulate and demonstrate the process signature that defines the onset of a reconnaissance attack on reactor priority signals.
3. We document the reactor signals used for the proposed WAN model evaluation and input representation, along with details of our data processing pipeline (Section 3.1.1).
4. We provide an in-depth discussion of model training and evaluation, covering hyperparameters, hardware specifications, and software versions (Section 4.1).
5. We conduct an extensive ablation study (Section 4.3) to isolate and analyze the impact of different model components on overall WAN performance.

The remaining sections of the paper are arranged as follows: Section 2 describes the state of the art in SMR cybersecurity, the Asherah Nuclear Simulator used for cybersecurity evaluation, and background on deep learning architecture. Section 3 discusses the proposed WAN architecture, section 4 presents the evaluation results, and the last section concludes the study.

2. Background

2.1. Cybersecurity of small modular reactor control systems

Several research works on attack detection techniques in nuclear plant instrumentation and control systems have been proposed for conventional nuclear power plants, which applies to other advanced reactor designs as well. However, SMRs, like all nuclear power systems, are inherently complex and have distinct safety and security requirements, resulting from their unique attributes such as remote and distributed siting, increased automation, modularity and design diversity. Moreover, the integration with smart technologies increases the complexity of security management, introducing vulnerabilities through increased connectivity (Rodriguez, 2017). Further, the proposed remote operation of SMRs presents novel cyber vulnerabilities for the nuclear industry to address (Aamoth et al., 2022).

Nevertheless, SMRs have unique advantages, which position them for better security architecture to be incorporated by design. These unique advantages also inform the recent drive for cybersecurity to be incorporated into the early phase of SMR-driven integrated energy systems projects (Eggers, 2023). The integration of cybersecurity by design, which involves embedding security features at the early stages of the architectural and design processes, prescribes the implementation of cybersecurity in instrumentation and control systems within SMRs to ensure that these systems are resilient from the ground up. Further, significant initiatives are currently underway to integrate cybersecurity measures into the entire systems engineering lifecycle. Approaches like cyber-informed engineering and security by design frameworks aim to identify and mitigate cybersecurity risks throughout the design, development, and implementation phases (Eggers, 2023).

While these methods are valuable in promoting the importance of considering cybersecurity from the early stages of design to create more secure systems, they may not comprehensively address the full spectrum of digital risks (Eggers, 2023). Moreover, the future of SMRs is likely intertwined with their integration into broader renewable energy systems, which poses unique cybersecurity challenges. For instance, the interaction between SMRs and smart grids would necessitate novel security protocols that can handle dynamic and distributed energy networks (Rodriguez, 2017).

Research is ongoing into developing more comprehensive cybersecurity frameworks that anticipate and mitigate potential breaches (Ayodeji et al., 2023). For instance, a recent study proposes a systematic mapping review that evaluates and validates new tools and methods for cybersecurity risk assessment specifically tailored for nuclear power contexts (De Brito and De Sousa, 2022). The criticality of the adoption of state-of-the-art cybersecurity technologies has also been emphasised by studies that review currently developing technologies, providing insights into both traditional cybersecurity measures and advanced digital solutions tailored for advanced reactors (Poresky et al., 2017).

A growing body of literature has investigated the cybersecurity of SMRs, with significant advancements in regulatory frameworks, design methodologies, and technological innovations. However, the pace of technological development and the sophistication of cyber threats necessitate continuous research and evaluation of cybersecurity measures. The ongoing commitment to integrating cybersecurity into all phases of SMR design and operation would be better supported by research that seeks to validate the security postures in every developmental phase. Further, there is a noted deficiency in empirical research that tests the theoretical models and frameworks proposed in the literature. This emphasises the importance of developing flexible, low-cost testbeds where more case studies could be performed to validate the proposed cybersecurity measures under real-world conditions. This also emphasizes the need to adopt and adapt tools like the Asherah Nuclear Simulator.

2.2. Asherah nuclear simulator for advanced reactors cybersecurity analysis

The Asherah Nuclear Simulator (ANS) is a full-scope, open-source, modularized simulator of a 2700MWt pressurized water reactor developed as part of an International Atomic Energy Agency (IAEA) Coordinated Research Project (CRP) (Silva et al., 2020). The ANS is a physics simulator that models the complete behaviour of a PWR system. Developed in the MATLAB/Simulink environment, the ANS is suitable for implementing a hardware-in-the-loop cybersecurity evaluation (Silva et al., 2020). It contains modules defining primary and secondary loop dynamics, the plant's primary and secondary loop control and communication systems, and the reactor protection system. The communication module is implemented with the Modbus and open platform communication-universal architecture (OPC-UA) protocol. ANS possesses several key features that are essential for modelling and simulating nuclear power plant digital control system cyber-attacks.

These include simulated control interfaces such as valves, pumps, and actuators, as well as the separation of the process simulation model from the control system model. Additionally, ANS incorporates a solver that supports external data injection, cyber-attack scenario simulation and data acquisition, as well as the evaluation of computer security measures (Maccarone et al., 2023) (Busquim et al., 2021). These capabilities are crucial as they enable the control system components to be decoupled from the rest of the simulator and replaced with external controller models, facilitating the study and analysis of control system vulnerabilities and attack scenarios in a hardware-in-the-loop setting (Maccarone et al., 2023).

The ANS's capability to provide a realistic and detailed simulation environment for cybersecurity evaluation has been explored by different studies. The tool has been previously used in testing a localised cyber-attack detection kit designed to enhance the resilience of critical reactor equipment against cyber-attacks (Zhang and Coble, 2020). This kit includes a cyber-attack detection model that identifies anomalies in key components like control system actuators and an inference model that can reconstruct compromised signals to maintain safe operations temporarily. The simulator has also been used to study manipulation attacks that alter process data presented to NPP operators (Lee et al., 2021). The simulation involved scenarios where safety-related state data were manipulated and an adaptive Kalman filter was successfully employed to verify the reliability of displayed data and facilitate the correct safety response actions. The condensate tank water level sensor is also validated using the ANS in research that studies the safety and security impact of cyber-attacks against the nuclear reactor digital interface (Shin et al., 2021).

Although the ANS is developed for PWR, it can be repurposed as a Hardware-in-the-loop cybersecurity testbed for advanced reactors and SMRs (Lee et al., 2019). The ANS has been repurposed to develop an Advanced Reactor Cyber Analysis and Development Environment (ARCADE) that was integrated with a Small Modular Advanced High-Temperature Reactor (SmAHTR) model for cybersecurity applications (Rowland et al., 2022). The ANS is also instrumental in the security by design analysis of advanced reactors, providing a simulated platform to assess and improve the security features of reactor digital systems during the design phase (Hahn et al., 2022). In another application, the simulator is used to train nuclear facility operators in detecting and responding to cyber incidents (Song et al., 2020; Allison et al., 2023). This training aspect is crucial for enhancing the cybersecurity posture of nuclear power plants. The simulator also supports system-level cybersecurity analysis, helping researchers understand the interactions between different system components under cyber-attack scenarios. The tool was also used to test and validate a host-based intrusion detection system designed for modern PLCs (Kiranyaz et al., 2020).

These ANS adaptations and applications underscore its role in preemptive security strategy development for SMRs and advanced reactors. Since SMR designs are relatively new and tools that mimic the industrial control system behavior are limited, using the ANS for SMR cybersecurity research has several justifiable reasons: SMRs are characterized by their scalability and modular design, which can be effectively simulated in environments like ANS that are adaptable for various reactor sizes and configurations. The modular aspect of SMRs often involves the replication of standard units, and ANS can simulate the interconnected nature of these units under cyber-attack scenarios, providing insights into cascading effects within a multi-unit SMR site. Further, the ANS is equipped with features that allow for detailed cyber-physical interaction analysis, which is critical for both advanced reactors and SMRs. The adaptability of the ANS to SMR cybersecurity research is supported by the fundamental similarities between the reactor types, the versatility of the simulation tool, and the critical need for advanced cybersecurity solutions as nuclear technologies evolve.

The ability to simulate complex cyber-attacks and their impact on digital interfaces and safety systems is crucial as SMRs will utilize more digital solutions than traditional reactors to accommodate size and modularity requirements, potentially increasing their vulnerability to cyber threats. Moreover, developing a new open-source simulator specific to each SMR design could be resource-intensive and costly. Leveraging the ANS would conserve resources and minimize development time, making cybersecurity research more efficient and effective. As SMRs are still relatively new in the nuclear energy field, the use of a tried and tested simulator like ANS allows researchers and engineers to push the boundaries of what's possible in SMR cybersecurity without the initial need for building a new simulation tool from scratch. This can accelerate the development of secure SMR designs by adapting existing tools to fit new contexts. In this work, the ANS is used to reproduce all relevant plant control network traffic, providing a representative process behaviour after a cyber-attack. This is then used to acquire and evaluate representative data used for the proposed deep learning model evaluation.

2.3. Deep learning architecture

2.3.1. 1-D convolution units

Convolution neural networks (CNN) extract features from input data with different levels of abstraction and are sensitive to spatial information. While 2D convolution is commonly used for image and video processing, it is not suitable for one-dimensional signal processing due to computational complexity. One-dimensional (1D) convolution is a promising deep learning approach for limited-size, one-dimensional signals as it extracts features layer-wise in a signal and is less complex, trains faster, and generalizes better on 1D signals. It is also suitable for real-time and low-cost applications and has demonstrated superior performance in tasks with limited data and high signal variation (Kiranyaz et al., 2019).

Consider an adaptive 1D convolution neural network with a kernel size of 3 and a subsampling factor of 2. The basic operation performed in the hidden layer of the k^{th} neuron can be described as the sum of the sequence of convolutions that passes through an activation function, denoted as f, used to extract the inherent features in the input. This is defined by the forward and backpropagation operations conducted from one layer to the another. For a 1D convolution neuron in layer l, with the previous layer and next layer defines as $l-1$ and$l+1$, respectively. The input of the k^{th} neuron in layer l can be expressed as (Bahdanau et al., 2016):

$$x_k^l = b_k^l + \sum_{i=1}^{N_{l-1}} \mathrm{conv1D}\left(w_{ik}^{l-1}, s_i^{l-1}\right), \tag{1}$$

where w_{ik}^{l-1} is the weight of the 1D kernel from the i^{th} neuron at layer $l-1$ to the k^{th} neuron at layer l, x_k^l is the input to the k^{th} neuron in layer l, N_{l-1} is the number of neuron in the previous layer, b_k^l is the bias term, and s_i^{l-1} is the output of the i^{th} neuron at layer $l-1$. For a network with input layer l, input vector p, output layer L, and the corresponding output vector $\left[y_1^L, \cdots, y_{N_L}^L\right]$, the objective is to minimize the output Mean-Square-Error (MSE) expressed as:

$$\mathrm{MSE} = \mathrm{E}\left(y_1^L, \cdots, y_{N_L}^L\right) = \sum_{i=1}^{N_L} \left(y_i^L - t_i\right)^2 \tag{2}$$

The MSE of the k^{th} neuron is therefore computed by finding the derivative of the error E in each input–output sequence with respect to the weight connected to the k^{th} neuron (w_{ik}^{l-1}) and the bias of the neuron b_k^l. A simplified representation of the 1D propagation computation occurring in each internal neuron of a 1D CNN, and the computation propagation path has been fully discussed (Kiranyaz et al., 2019). In the current work, a six-layer 1D architecture composed of 1D convolution and attention mechanism is utilized, developed into four major branches: the input, filter, gating, and the output. The theoretical background of the attention mechanism used in this work is described in the next section,

and a full description of the proposed WAN is provided in Section 3.

2.3.2. Attention mechanism

The attention mechanism is employed to enhance the learning of long-range sequential knowledge and is effective in addressing a key drawback of fixed-length encoding of context vectors in sequence-to-sequence recurrent networks – compression and loss of information. In natural language processing, an attention vector estimates word correlation and sums up their weighted values as an approximation of the target. However, its use for multivariate prediction implementation is rare, as the network forgets the beginning after processing the input. Hence, it is combined with a 1D convolution network in this work. A sequence-to-sequence learning tasks can be modeled using a recurrent encoder-decoder network architecture. The encoder processes a series of input vectors $x = (x_1, \cdots, x_n)$ sequentially, updating its hidden state at each timestep t. After encoding the full sequence, it derives a context vector $\widehat{c}$ to summarize the inputs. The decoder then utilizes this context vector $\widehat{c}$ to generate relevant outputs. The hidden state updates in the recurrent encoder at each t can be represented as (Ayodeji et al., 2109):

$$h_t = f(x_t, h_{t-1}) \tag{3}$$

The hidden state's context vector is given by:

$$\widehat{c} = q(\{h_1, \cdots, h_n\}) \tag{4}$$

where f and q are nonlinear functions. For the context vector and the previous sequence $\{y_1, \cdots, y_{t-1}\}$, the decoder predicts the next sequence y_t, by decomposing the joint probability, such that:

$$p(\mathbf{y}) = \prod_{t=1}^{T} p(y_t | \{y_1, \cdots, y_{t-1}\}, \widehat{c}) \tag{5}$$

where the decoder output vector $y = (y_1, \cdots, y_m)$. In the context of attention mechanism, each conditional probability expressed in equation (5) above is defined as:

$$p(y_1, \cdots, y_{i-1}, x) = g(y_{i-1}, a_i, \widehat{c}_i) \tag{6}$$

Where g is a nonlinear function, and s_t is the attention vector of the hidden state at time t, given as:

$$s_t = f(s_{t-1}, y_{t-1}, \widehat{c}_t) = f(\widehat{c}_t, h_t) \tag{7}$$

Hence, the context vector $\widehat{c}_i$ computed as a weighted sum of the sequence of annotations $h_t = (h_1, \cdots, h_n)$, is given by:

$$\widehat{\mathbf{c}}_t = \sum_{j=1}^{J} \alpha_{ij} h_j \tag{8}$$

In equation (8), α_{tj} is the attention weight from the t^{th} output to the j^{th} input, and the encoder state for the j^{th} input is h_j. The attention weights assigning importance across input tokens are derived by mapping the inputs to an alignment score. This attention score e_{tj} for each token j at decoding timestep t determines the subsequent relevance distribution in the attention layer, and it is given as:

$$\alpha_{t,j} = \frac{\exp(\text{score}(\mathbf{h}_t, \mathbf{h}_j))}{\sum_{j=1}^{J} \exp(\text{score}(\mathbf{h}_t, \mathbf{h}_j))} \tag{9}$$

The alignment model measures the distance between input positions around t and the output position, and it is defined as $e_{tj} = f(s_{t-1}, h_j)$. The function f scores the matching distance between input and output, and s_{t-1} represents the hidden state from the previous timestep. The attention annotations are described in detail (Dong et al., 2021), and this work uses the Keras implementation of the self-attention mechanism, which calculates the alignment of hidden scores $(h_{t,t'})$, the attention weight (a_t), the alignment model $(e_{t,t'})$, and the context vector (l_t) as:

$$h_{t,t'} = \tanh(x_t^T W_t + x_{t'}^T W_x + b_t) \tag{10}$$

$$e_{t,t'} = \sigma(W_a h_{t,t'} + b_a) \tag{11}$$

$$\alpha_t = \text{softmax}(e_t \tag{12}$$

$$l_t = \sum_{t'} \alpha_{t,t'} x_{t'} \tag{13}$$

*W*here W's and b's are weights and biases to be learned, x_t and $x_{t'}$ are the input vectors at time t and t' respectively, W_t and W_x are the weight matrices to be learned. In equation (11), W_a and b_a are the weight matrix and bias for calculating the attention score $e_{t,t'}$.

A recent analysis of deep self-attention networks comprised fully of stacked multi-head self-attention layers, uncovered an exponential convergence phenomenon. As more layers are added, the learned token representations rapidly approach identical, rank-1 matrices that map all tokens to the same representation (Van Den Oord et al., 2016). This trend suggests that extremely deep self-attention networks lose the capacity to distinguish input details. Further examination of this behavior could inform enhanced self-attention architectures.

3. Method

3.1. Wavy-attention model development and architecture

The wavy-attention network, inspired by the WaveNet architecture (El-Genk et al., 2021), is a specialized convolution architecture that systematically integrates dilated convolution layers. This design enables the network to learn temporal and spatial information inherent in datasets from complex engineering systems. The proposed wavy-attention network consists of a series of 1D causal convolution networks with an enlarged receptive field, parametrized skip connection, residual connection, and activation units enhanced with a self-attention mechanism, as depicted in Fig. 1.

This paper first adopts a padded convolution operation with a filter size of 16, and kernel = 1, based on experimental results, to map the inputs with the same feature dimension. This layer serves as the input preprocessing unit before channeling it to the dilated convolution layer, which is used as the network filter. The network filtering is accomplished by a dilated convolution with a filter size of 32 and a filter width of 2. A similar dilated convolution module is implemented as the network gating branch. Hence, a network of multi-layer dilated convolution is constructed, where the dilation factor (D_f) increases sequentially at each layer. To speed up training, a *LeakyRelu* and *BatchNormalization* module is introduced in each layer. To avoid model explosion and aggregate important information in each signal, a regularized self-attention module is added to the gating branch.

Dilated layers are typically incorporated into networks to aggregate multi-scale spatial information. By expanding the receptive field without increasing parameters, they allow models to jointly capture both local fine details and global context. To effectively increase the receptive field, a dilation rate is introduced. By increasing the dilation rate at each layer, the network achieves the desired exponential relationship between layer depth and receptive field size. Given a sequence of input $x \in R^n$, and filter $f : \{0,....k-1\} \rightarrow R$, dilation introduces a "hole" in the convolution without changing its weights (Kiranyaz et al., 2019).

Dilated convolution expands a kernel's receptive field without increasing the number of parameters. Dilated convolution achieves receptive field expansion by skipping input samples with a predetermined dilation rate α when calculating the convolution. Effectively, the spatial span encompassed by the original *kxk* kernel is enlarged to $\alpha(k-1)+1$, through this strided sampling. As a result, the receptive field is grown to cover a wider context region, enabling aggregation of multi-scale visual details.

The dilated convolution output in the WAN architecture splits into

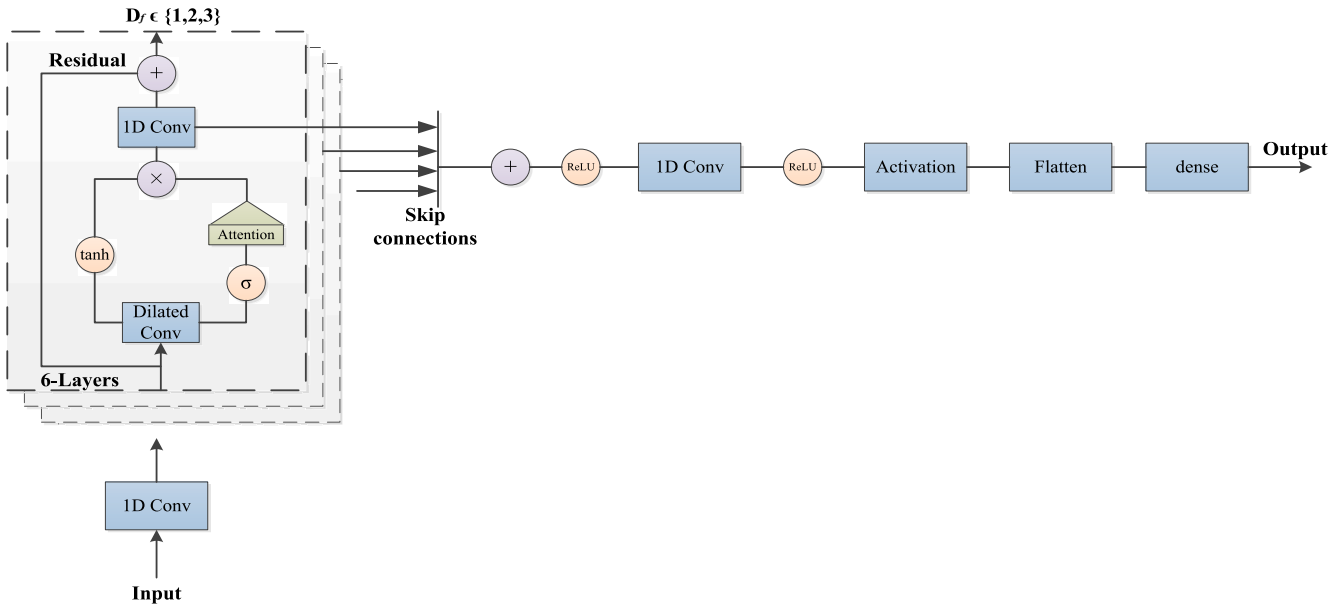

Fig. 1. The proposed WAN sub-structure.

two branches (filter and gating branch) and later recombines via element-wise multiplications. The skip connection is utilized to dynamically tune the layer numbers during training, benefiting model convergence and mitigating rank collapse. The skip connection enables the utilization of fewer layers and units while preserving the lower-level information from distortion. This also ensures that the network retains collections of feature output at all levels in the network hierarchy, as opposed to a singular set of complex feature outputs. Likewise, the residual connections facilitate the propagation and aggregation of representation output from each layer to enable enhanced collective processing in deeper layers. This is also used to address the layer degradation and vanishing/exploding gradients problems in deep learning architecture. For some function f, which represents the model's learned weight, the residual connection ensures the network output is mapped to the input as $x_{out} = f(x_{in}) + x_{in}$, as opposed to the traditional $x_{out} = f(x_{in})$. The gated activation unit is used to facilitate the recombination of the dilated convolution output in the wavy-attention architecture via element-wise multiplications. A similar gated activation unit in the conventional WaveNet is utilized in the wavy-attention network, mathematically expressed as:

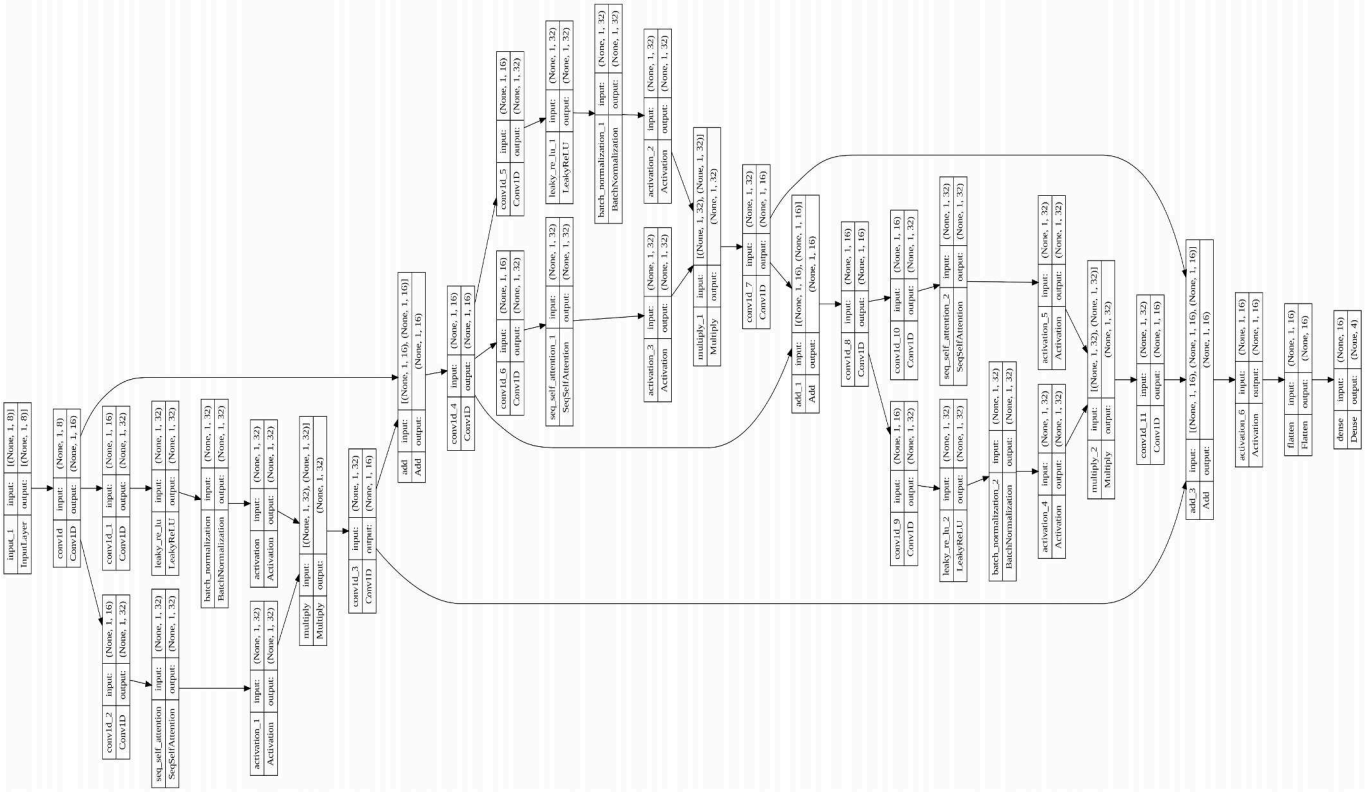

Fig. 2. The full graph of the proposed WAN.

$$z = \tanh\left(W_f * x\right) \odot \sigma\left(W_g * x\right) \tag{14}$$

Where *f* is the filter, *g* is the gate, * denotes the convolution operator, $\odot$ is the element-wise multiplicative operator, *W* is the learnable convolution filter, σ is the sigmoid function branch (representing the learned gate that regulates the information flow from the filter), and *x* is the input. With input dimensions spanning hundreds or thousands of layers, as commonly seen in computer vision and image recognition models, these architectural innovations help overcome optimization difficulties and accuracy limitations that arise with extreme depth. However, they are sub-optimal for multivariate prediction on structured data. Consequently, the output of the gating function is optimized with a regularized self-attention. Fig. 2 below illustrates the full architecture of the proposed WAN.

As seen in Fig. 2, the first layer defines the input layer of the model that accepts a 3D input tensor representing a sequence of feature vectors. Then the *dilation_rate* loop is constructed to iterate over different dilation rates and create multiple layers, each with a different field of view over the input sequence. This allows the network to integrate information from various time steps without increasing the number of parameters significantly. Before applying the dilated convolutions, the input is processed by a 1D convolutional layer with 16 filters and a kernel size of 1. This layer can be considered as a time-distributed dense layer, allowing for feature extraction and transformation. It processes each time step individually while keeping the sequence length constant. A dilated 1D convolutional layer with *n_filters* and *filter_width* is applied to the preprocessed input. This layer uses causal padding, ensuring that the output at each timestep depends only on past inputs.

The dilation rate is varied across iterations, allowing the model to capture patterns at different time scales. The filter applies a nonlinear transformation and uses dilated convolutions to capture long-range dependencies. The *LeakyReLU* and *BatchNormalization* are applied for non-linearity and to stabilize the learning. Another dilated 1D convolutional layer is applied to the preprocessed input, followed by a sequence self-attention layer. This branch acts as a gating mechanism which controls the flow of information while learning to weigh the importance of different input features and timesteps. The gate applies a self-attention mechanism that allows the model to focus on different parts of the input sequence, which is critical for capturing complex patterns. The two sets of layers act as filters and gates and are inspired by the gated mechanisms in LSTMs and GRUs.

Subsequently, the outputs of the filter and gating branches are combined using an element-wise multiplication. The filter output is passed through a *tanh* activation, while the gating output is passed through a *sigmoid* activation. This combination allows the model to selectively pass or block information based on the learned gating mechanism. The combined output is processed by another *1D* convolutional layer with 16 filters and a kernel size of 1, which can be considered as another time-distributed dense layer. The postprocessed output is added back to the original input, forming a residual connection. This technique helps mitigate the vanishing gradient problem and allows the model to learn residual mappings more effectively.

The output of each dilation rate iteration is collected as a skip connection. These skip connections will be combined later, enabling the model to leverage information from different dilation rates. All the collected skip connections are summed together, and a ReLU activation is applied. This step integrates the information from different dilation rates and captures patterns at multiple time scales. The combined skip connection outputs are flattened, and a dense layer with 4 units and a *softmax* activation is applied. This final layer produces the model's output, which could be used for tasks like classification or regression, depending on the problem.

The model is compiled with categorical cross-entropy loss and the Adam optimizer with a specified learning rate. The accuracy metric is also included for evaluation purposes. In summary, this architecture combines dilated convolutions, gating mechanisms, self-attention, residual connections, and skip connections, enabling the model to effectively process sequential data and capture long-range dependencies. The dilation rates allow the model to learn patterns at different time scales, while the gating and attention mechanisms help the model focus on relevant features and timesteps. The residual and skip connections facilitate better gradient flow and information propagation throughout the network.

3.1.1. False data injection attack simulation

The digital control system in a nuclear plant is vulnerable to False Data Injection Attack (FDIA) directed at critical process sensors [33]. The FDIA generate false instructions for controlling various components and could compromise the plant operation. To acquire representative data for sensor attacks, this work utilizes the IAEA's ANS simulator and the false data injection toolbox (Potluri et al., 2019).

The false data injection toolbox contains different function blocks that represent various types of industrial control system attacks (Ayodeji et al., 2023). The toolbox enables the injection of manipulated attack values into plant signals for fast modelling and simulation of various attacks. This ranges from naive malicious injection attacks, which can capture and alter the network packet from server to client, to Complex Malicious Response Injection (CMRI) attacks that could mask the physical response necessary for the feedback control loop. In this work, three different CMRI-type attacks on the sensor measurements are simulated. The attacks are executed on priority sensor signals critical to reactor control. The simulated attacks are:

i. High-frequency Measurement Injection: The HFMI attack obfuscate the authentic physical behavior of the targeted system, thereby exerting a negative influence on the control system that supervises the cyber-physical system behavior. Specifically, these attacks manipulate the frequency of the measured process, causing it to deviate from its standard rate and appear as if it is operating within normal system parameters. This attack is simulated on the pressurizer pressure sensor of the reactor coolant system.
ii. High Slope Measurement Injection: The HSMI is an attack method that involves sending identical process measurements repeatedly to conceal the actual state of the system. The measurements are fully recorded and then replayed to give the client the illusion that the system is functioning normally. The focus of this attack is on capturing and reproducing the sensor signal. The attack is simulated on the pressurizer level sensor signal of the reactor coolant temperature.
iii. Restart Communication Attack: Restart communication occurs when data is sent from server to client and the transmission is interrupted. This interruption may cause data to be temporarily lost, resulting in delays in communication. During this time, default values may be used, and no communication takes place. These interruptions can significantly disrupt closed-loop feedback control. The attack is simulated on the reactor mean coolant temperature signal, a priority signal for reactor control, pressurizer pressure control, pressurizer level control, and the reactor protection system.

These attacks could be introduced into small modular reactor controls via the supply chain or through a malicious insider. Implementing these attacks on sensor signals resulted in compromised signals without activating the reactor protection system. This makes the detection of the simulated attacks challenging for conventional intrusion detection systems. Table 1 summarizes the simulated attack, their classification, and the sensors impacted. Figs. 3-5 also show the critical sensor signals in steady state and under attack.

3.1.2. Dataset description

The proposed WAN architecture is evaluated on a sensor signal

Table 1
simulated attack, classification, and affected sensors signal.

S. No	Attack Name	Sensors Affected	Attack class
1	Normal	NIL	0
2	High Frequency Injection Attack (HFMI)	Pressurizer pressure	1
3	High Slope Measurement Attack (HSMI)	Pressurizer level	2
4	Restart communication Attack (RCA)	Reactor mean coolant temperature	3

facilitate reproducibility, annotated *jupyter* notebooks with code and trained models are provided in the first author's GitHub repository.[1]

To avoid overfitting, the *early_stopping* and *checkpoint* callback functions are implemented. The *early_stopping* function monitors the validation loss and stops model training once the loss does not improve, while the *checkpoint* saves the best model at the epoch with high validation accuracy. The *Adam* optimizer and a fixed learning rate of 3e-5 were used for all experiments. The metric used for model evaluation are the confusion matrix, precision, recall, and the f1_score. The model was trained with a *batch size* of 16, and the *epoch* was specified as 500. To demonstrate the performance of the proposed WAN, Fig. 6 illustrates the

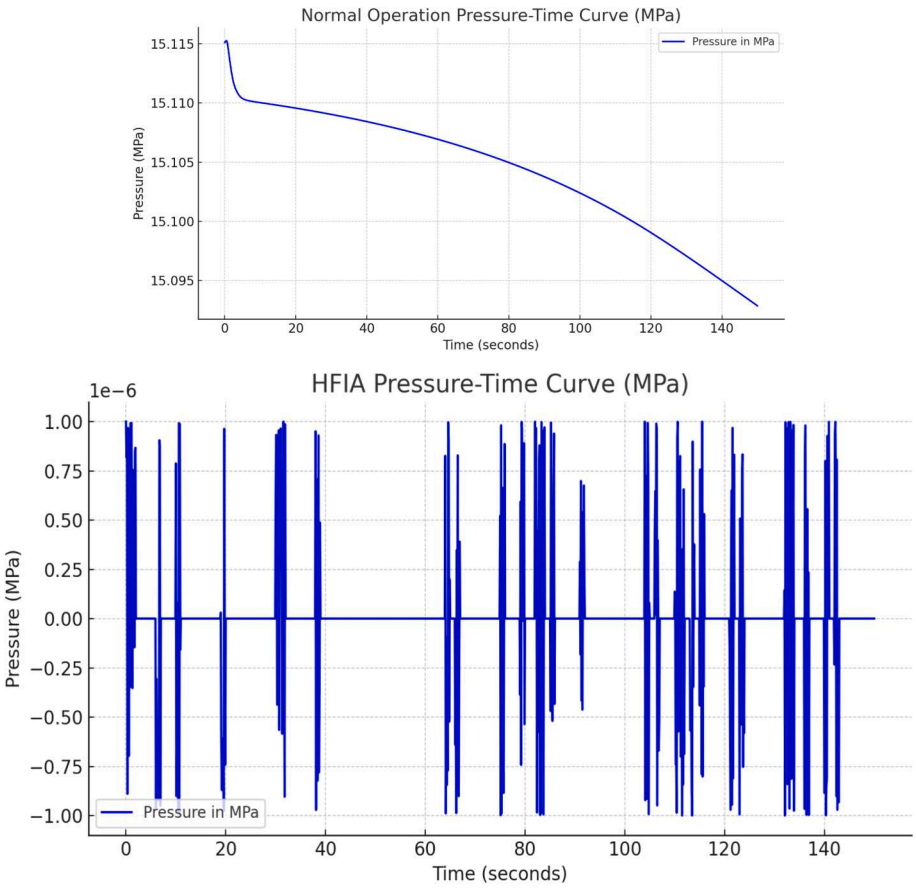

Fig. 3. (a) Normal pressurizer pressure −time curve. Fig. 3: (b) Pressurizer pressure–time curve under HFI attack.

attack dataset obtained from the IAEA's ANS simulator. The variables in the dataset define the dynamics of the reactor coolant system during normal operation and the stated attacks. The total data contains eight variables with 6005 timesteps of the reactor coolant system dynamics. For all logged variables, the input/output type is analog, and the MATLAB data type is single. The data is obtained by simulating each attack for 150 s. The training dataset comprises 70 % of the total data, with 15 % used as a validation set, and 15 % as the test set. A detailed description of the variables used as sensor attack indicators is presented in Table 2.

4. Result and discussion

4.1. Model training and evaluation

This study employs the *Keras* API on *Tensorflow* to develop the WAN model. Experiments are conducted on an Intel Core i-7 workstation with an RTX 2060 s GPU. Additionally, comparison experiments in Section 4.2 are carried out on the cloud-based platform, *Google Colab*. To

training and validation loss of the model. This figure shows the training and validation loss and accuracy curves for the proposed WAN model during the training process, demonstrating how the loss decreases and accuracy increases over epochs. Fig. 7 presents the graphical confusion matrix. This confusion matrix visualizes the performance of the WAN model on the test set, displaying the number of true positives, true negatives, false positives, and false negatives for each attack class and normal operation.

The confusion matrix shown in Fig. 7 is the performance of the WAN model in classifying different types of reactor conditions. The matrix displays the true labels on the rows and the predicted labels on the columns. The diagonal values represent the correctly classified instances, while the off-diagonal values represent the misclassified instances. The figure shows that the model correctly classifies all instances of Normal reactor operation with a value of 1.0 on the diagonal. For

[1] https://github.com/abiodun-ayodeji/Wavy-attention-network-for-cybersecurity.

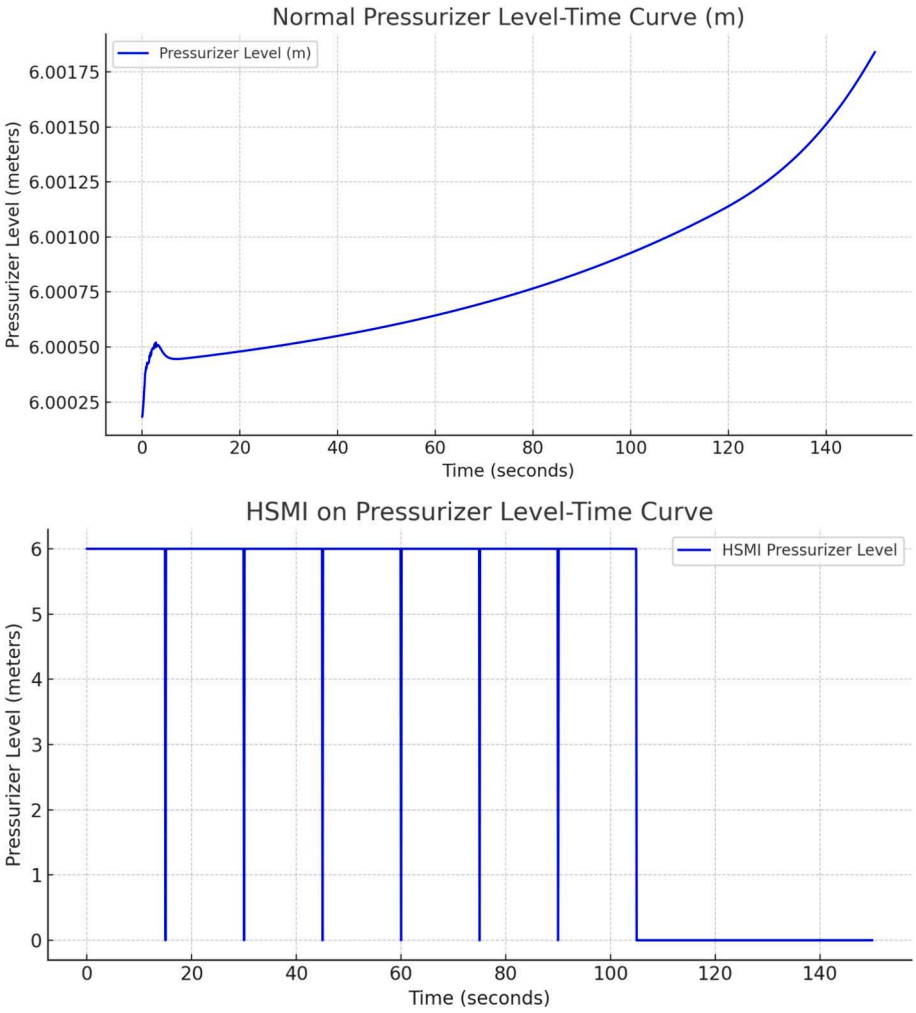

Fig. 4. (a) Normal pressurizer level − time curve. Fig. 4: (b) Pressurizer level − time curve under HSMI attack.

HFIA and HSMI on pressurizer pressure and level, the model also correctly classifies all instances of these classes with a value of 1.0 on the diagonal. For RCA on reactor mean coolant temperature, the model correctly classifies most instances of this class with a value of 0.97 on the diagonal. However, there are some misclassified instances (0.03) that are incorrectly classified as *Normal Operation.* Overall, the confusion matrix indicates that the WAN model performs exceptionally well in classifying the different event types, with a very high accuracy for most classes. The only notable misclassification occurs for a small fraction (0.03) of the *RCA_on_RX_MeanCoolTemp* class, which is misclassified as "Normal Operation". This confusion matrix suggests that the WAN model is highly reliable and accurate for classifying these types of events or conditions, with minimal confusion between the classes.

Table 3 showcases the predictive performance of the model when an out-of-sample test set is used to evaluate the model. It is observed that the model achieved 100 % accuracy during training. Furthermore, this high accuracy is maintained in its predictions on the test set as well. The confusion matrix reveals that the model performs at 99 % accuracy on the test data. To further evaluate the model's performance, Table 3 presents the precision, recall, and the f1-score of the model. The model demonstrates 100 % performance for all plant conditions evaluated, except the restart communication injection attack, where it achieves a 98 % accuracy.

Fig. 8 shows the Receiver Operating Characteristic (ROC) curve for WAN, which classifies instances into four classes (class 0, class 1, class 2, and class 3). The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 − specificity) for different classification thresholds, and, in general, the closer the ROC curve is to the top-left corner of the plot, the better the classification performance. The diagonal line represents a random classifier with no discriminative power. Based on the ROC curve, it is seen that the lines for all classes are close to the top left corner of the graph, indicating good performance for these classes. The high sensitivity and specificity values suggest that the model can effectively distinguish the classes from the other classes.

4.2. Comparison with baseline models

To properly account for the impressive performance achieved in this paper, Table 4 compares the WAN model with other baseline models that have been found to perform well on multivariate datasets. The compared models are the vanilla convolution neural network (vCNN), the long short-term memory network (LSTM), and the bidirectional LSTM (Bi-LSTM). All models are evaluated for 20 epochs. Table 4 displays the comparison result of the baseline models with the proposed WAN model.

Table 4 provides a comparison of four baseline models based on their average precision, average recall, average f1-score, and average test accuracy. The vCNN model has an average precision of 0.93, which shows that the model correctly identifies 93 % of the relevant instances among all the instances it identifies. The average recall of 0.90 indicates that the model correctly identifies 90 % of the relevant instances among all the actual instances. An f1-score of 0.90 demonstrates a balance between precision and recall, and the average test accuracy of 0.90 indicates that the model correctly predicts the class label on 90 % of the

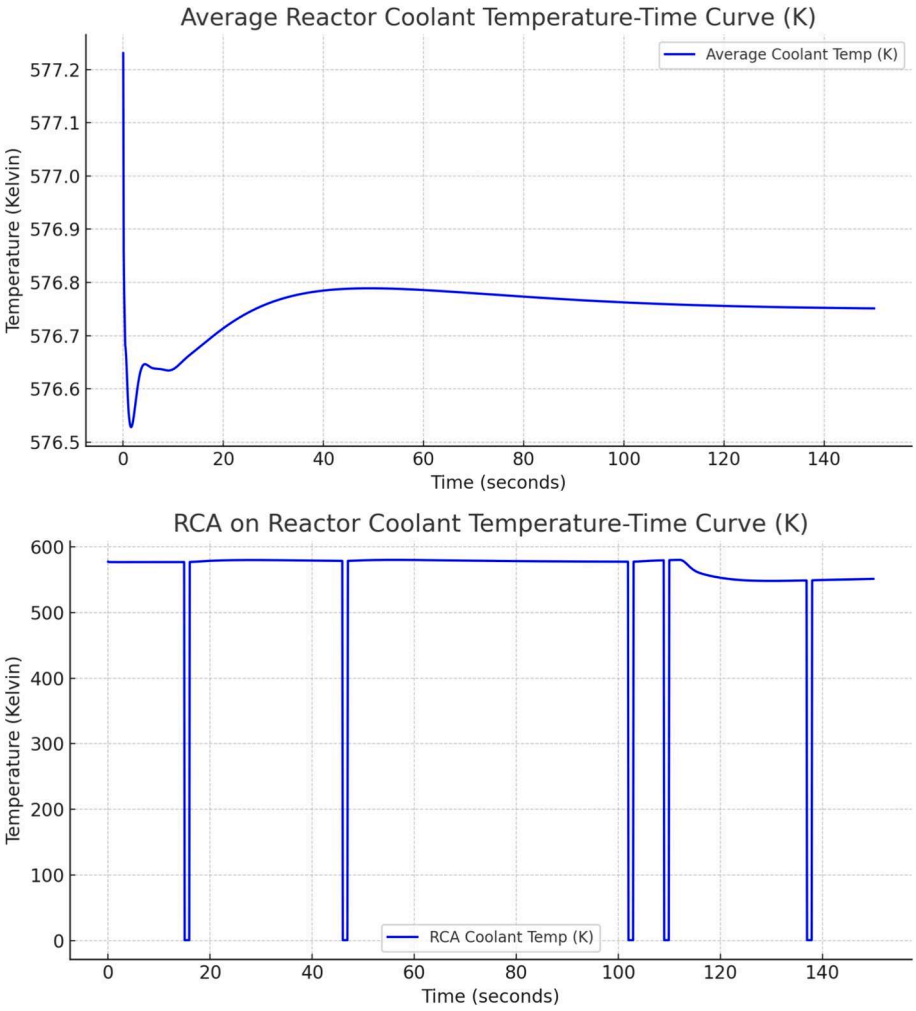

Fig. 5. (a) Normal reactor mean coolant temperature − time curve. Fig. 5: (b) Reactor mean coolant temperature–time curve under RCA attack.

Table 2
Parameters used for WAN model evaluation.

S. No	Tag name	Description	Unit
1	PZ_Press	Pressurizer pressure (reactor hot leg pressure)	Pa
2	PZ_Temp	Pressurizer temperature	K
3	PZ_Level	Pressurizer level	m
4	RC1_PumpSpeed	Reactor coolant pump 1 speed	100 %
5	RC1_PumpFlow	Reactor coolant pump 1 flow	Kg/s
6	RX_MeanCoolTemp	Reactor mean coolant temperature	K
7	RC2_PumpSpeed	Reactor coolant pump 2 speed	100 %
8	RC2_PumpFlow	Reactor coolant pump 2 flow	Kg/s

test instances.

The LSTM model has an average precision of 0.83, which is lower than vCNN, suggesting a higher chance of identifying false positives. The average recall of 0.77 indicates that the model correctly identifies only 77 % of the relevant instances, also lower than the vCNN. With an f1-score of 0.71, it is the lowest among all the models, highlighting an imbalance between precision and recall. Additionally, the average test accuracy of 0.77 is the lowest among all the models, signifying the highest error rate in predicting the class labels of the test instances.

The Bi-LSTM model has an average precision of 0.91, which is higher than LSTM but lower than WAN. With an average recall of 0.93, it surpasses vCNN and LSTM, indicating that the model can correctly identify most of the relevant instances. The f1-score of 0.93 is the highest among all the models, which shows a good balance between precision and recall. The average test accuracy of 0.91 is also the second highest, indicating that the model can predict the class labels of the test instances with high accuracy. The proposed WAN model outperforms all others with the highest average precision, recall, f1-score, and test accuracy. Its average precision of 0.99 means the model can correctly identify almost all the relevant instances among all the instances it identifies. An average recall of 0.99 indicates that the model can correctly identify almost all the relevant instances among all the actual instances. The f1-score of 0.99 is also the highest, which is a good balance between precision and recall. The average test accuracy of 0.99 is the highest among all models, demonstrating the model's ability to predict the class labels of the test instances with the utmost accuracy.

The above results demonstrate that the WAN model shows a significant improvement over all other baseline models across all performance metrics. The most substantial enhancement is observed in the average f1-score, where the WAN model improves by 39.44 % compared to LSTM. The improved performance of the WAN model is directly linked to its optimized architecture and the inclusion of a self-attention gate. The sequential self-attention layer enables the model to focus on relevant parts of the input sequence and learn their representations. This attention mechanism assigns weights to the input sequence elements based on their relevance to the target task. By attending to relevant parts of the sequence, the model can capture long-term dependencies and effectively represent the input sequence for classification.

The analysis above showcases the significant improvement enabled by the attention module of the proposed WAN. However, it is also

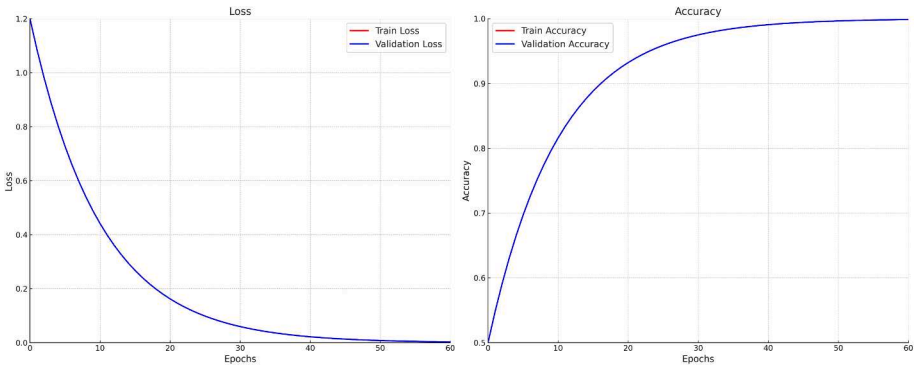

Fig. 6. The WAN training and validation loss and accuracy curve.

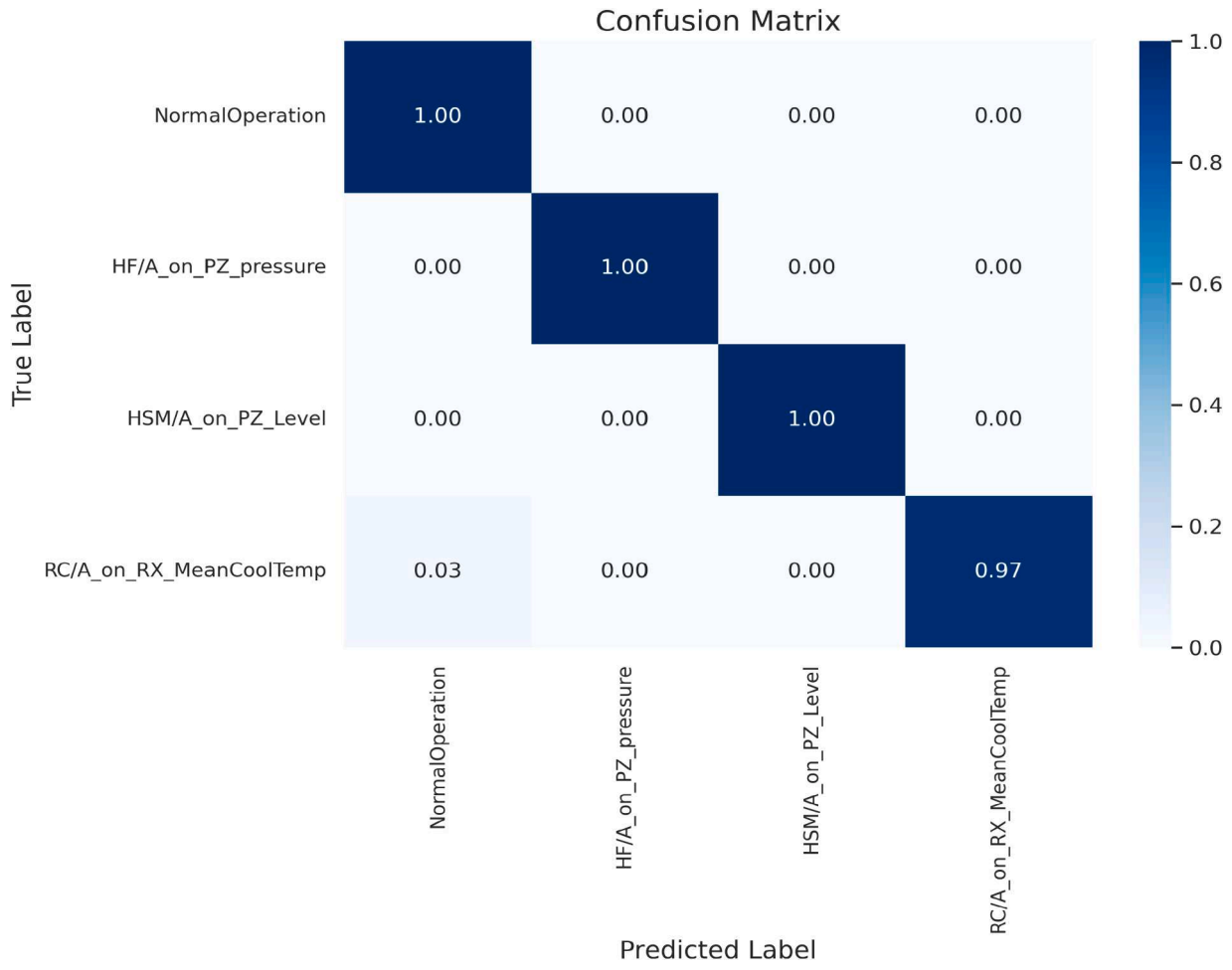

Fig. 7. WAN confusion matrix for attack prediction.

Table 3
Classification report for the proposed WAN network.

	Precision	Recall	F1-score	Support
0	0.98	1.00	0.99	447
1	1.00	1.00	1.00	460
2	1.00	1.00	1.00	453
3	1.00	0.98	0.99	442
accuracy			0.99	1802
macro avg	0.99	0.99	0.99	1802
weighted avg	0.99	0.99	0.99	1802

pertinent to mention a few limitations of this work. Firstly, the WAN model was trained exclusively using process measurements. In a real-world application, it is critical to fuse the process data with network packets for better decision-making. The authors are currently implementing a proof of concept, using the ANS. The concept involves programming and integrating a Siemens 1500 PLC into the simulator, in a hardware-in-the-loop configuration. The PLC reads tags from the OPC-UA server, implements the control logic, and writes in the server. Utilizing the OPC-UA communication module would enable network packets to be captured and analyzed using open-source tools such as Wireshark or tcpdump. Secondly, the ANS is a 2700MWth simulator, which is not an exact representation of a small modular reactor.

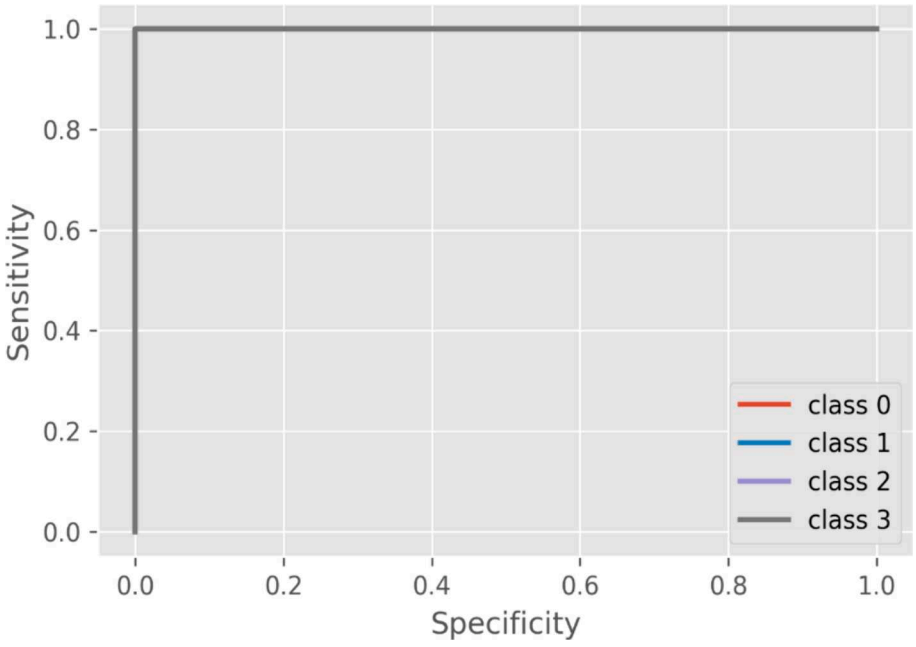

Fig. 8. ROC curve for the WAN.

Table 4
The comparison results of baseline models with the proposed WAN.

Model	Average precision	Average recall	Average f1-score	Average test accuracy
vCNN	0.93	0.90	0.90	0.90
LSTM	0.83	0.77	0.71	0.77
Bi-LSTM	0.91	0.93	0.93	0.91
WAN	0.99	0.99	0.99	0.99

However, the ANS could be modified to reflect the power rating of a typical SMR. In subsequent work, these limitations will be addressed.

4.3. Ablation study

To account for the contribution of individual layers in the proposed WAN, this section presents four ablated versions of the proposed architecture. The ablations aim to isolate components and simplify the architecture to test their contribution. In the first ablated version (filter-only WAN), the gating branch that weights the outputs of the filter convolutions before combining them is removed. In the full model, the gating helps learn which filter outputs are most relevant. The second ablated version (Single-dilation model) uses only a dilation rate of 1, removing the dilated convolutions. The full model uses multiple dilation rates to capture multi-scale temporal patterns. The third ablated version (plain CNN) removes attention and the WaveNet-style structure, using just plain CNN layers. The fourth ablated version removes the gating branch and also changes the convolutions to use 'same' padding to maintain sequence length instead of using causal padding. Table 5 shows the ablated versions and evaluation result.

Table 5
Performance evaluation of the ablated versions of WAN.

Model	Average precision	Average recall	Average f1-score	Average test accuracy
Plain-CNN	0.94	0.92	0.92	0.92
Same-padded WAN	0.06	0.24	0.09	0.25
Single Dilation WAN	0.99	0.99	0.99	0.98
Filter-only WAN	0.99	0.99	0.99	0.99

It is observed that simplifying the WaveNet-based architecture (WAN) to a plain convolution neural network (Plain-CNN) leads to a drop in performance across all evaluation metrics. Specifically, the Plain-CNN model achieves an average precision of 0.94, recall of 0.92, F1-score of 0.92, and accuracy of 0.92 on the tested dataset. This is likely because the Plain-CNN lacks components such as gated dilated convolution and residual connection which allow the WAN models to capture long-range dependencies in sequential data. Additionally, ablating the gating mechanism and replacing causal padding with *same* padding (Same-padded WAN) severely hampers model performance, resulting in very poor precision, recall, F1, and accuracy. This underscores the significance of both the gating component and causal padding in enabling the model to leverage useful context from the sequence history.

On the other hand, using a single dilation rate instead of multiple (Single Dilation WAN) has a negligible impact on metrics, with this model achieving performance on par with the un-ablated WAN architecture. This suggests that multi-scale feature extraction via dilated convolutions provides minimal benefits for the current dataset. Similarly, removing just the gating branches (Filter-only WAN) also barely impacts metrics, indicating they may be redundant in the presence of other structures that capture sequential dependencies.

In summary, plain CNN stacks fail to model temporal relationships as effectively as specialized architectures like WANs for sequential data. Moreover, gating and causal padding are critical components, while attention and multiple dilations appear less important for model performance on this particular dataset. The results show that mechanisms that enable access to useful contexts, such as gating and causal padding, remain vital for strong performance. Further evaluation is required on more complex datasets to properly analyze the effect of each component of the model.

5. Conclusion

This paper presents the simulation and detection of subtle false data injection attacks on priority signals in a reactor digital control system using a MATLAB/Simulink-based Asherah Nuclear Simulator and a false data injection toolbox. The attack detection is achieved using a novel

wavy-attention network (WAN) architecture that leverages spatio-temporal correlations in multivariate time-series data for real-time cyber-attack detection. The evaluation result of the WAN performance shows that the proposed WAN significantly outperforms conventional deep learning models on the challenging task of detecting subtle false data injection attacks on critical sensor signals in a reactor digital control system.

A key highlight of this work is the demonstration and detection of process signature that defines the onset of reconnaissance attack on priority reactor signals. Moreover, the WAN's innovative integration of dilated convolutions with a self-attention mechanism enables it to effectively capture long-range dependencies and learn robust representations of complex sequential patterns. The gating components and causal padding were also found to be crucial for strong performance through the ablation study. The implications of this work are far-reaching for enhancing cybersecurity resilience in safety–critical systems in advanced reactors and SMRs. By providing accurate and timely detection of stealthy cyber-attacks, the WAN can trigger appropriate mitigation responses before significant damage occurs. Its high accuracy of 99 % demonstrates the potential to reliably safeguard digital instrumentation and control systems.

Looking ahead, the WAN architecture holds significant promise for applications across the entire nuclear fuel cycle beyond just reactor operations. For instance, it could monitor sensor data streams during enrichment, fuel fabrication, spent fuel storage and reprocessing to identify cyber threats targeting these facilities. The WAN's generic sequence modelling capabilities make it readily adaptable to diverse data modalities. Moreover, the ability to detect anomalies in multivariate time-series data positions WAN as a powerful tool for predictive maintenance and condition-based monitoring across nuclear facilities. By learning normal operational patterns, it can promptly flag deviations indicative of emerging faults or degradation in critical equipment and processes. Integrating WAN into an attack-resilient control framework would be an interesting future research direction.

CRediT authorship contribution statement

Abiodun Ayodeji: Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Antonio Di Buono:** Writing – review & editing. **Iestyn Pierce:** Writing – review & editing, Validation. **Hafiz Ahmed:** Writing – review & editing, Validation, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The link to data and code is embedded in the paper

Acknowledgements

This work is partially supported by the High Value Manufacturing Catapult through NIN2394-FY2023 Developing Cyber Security Capability; by the CHIST-ERA funded TROCI Project (CHIST-ERA-22-SPiDDS-07) through the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/Y036344/1; and by the Sêr Cymru II 80761-BU-103 project by the Welsh European Funding Office under the European Regional Development Fund.

References

Aamoth, B., Lee, W.E., Ahmed, H., 2022. Net-Zero Through Small Modular Reactors - Cybersecurity Considerations. IECON 2022–48th Annual Conference of the IEEE Industrial Electronics Society.

Allison, D., Mclaughlin, K., Goosewolf, S.P., 2023. An Embedded Intrusion Detection System for Advanced Programmable Logic Controllers. Digital Threats. 4 (4), 1.

Ayodeji, A., Liu, Y., Chao, N., Yang, L., 2020. A new perspective towards the development of robust data-driven intrusion detection for industrial control systems. Nucl. Eng. Technol. 52 (12), 2687.

Ayodeji, A., Amidu, M.A., Olatubosun, S.A., Addad, Y., Ahmed, H., 2022. Deep learning for safety assessment of nuclear power reactors: Reliability, explainability, and research opportunities. Prog. Nucl. Energy 151.

Ayodeji A, Di Buono A, Pierce I, Mohamed M, Ahmed H. Wavy-Attention Network for Real-Time Cyber-Attack Detection in a Pressurized Water Reactor Digital Control System. 13th Nuclear Plant Instrumentation, Control & Human-Machine Interface Technologies (NPIC&HMIT 2023) 2023.

Ayodeji, Abiodun, Mokhtar Mohamed, Li Li, Antonio Di Buono, Iestyn Pierce, and Hafiz Ahmed. "Cyber security in the nuclear industry: A closer look at digital control systems, networks and human factors."Progr. Nucl. Energy 161 (2023): 104738.

Ayodeji, Abiodun, Wenhai Wang, Jianzhong Su, Jianquan Yuan, and Xinggao Liu. An empirical evaluation of attention-based multi-head models for improved turbofan engine remaining useful life prediction. arXiv preprint arXiv:2109.01761 (2021).

Bahdanau D, Cho K, Bengio Y. NEURAL MACHINE TRANSLATION BY JOINTLY LEARNING TO ALIGN AND TRANSLATE. 2016 -05-19.

Busquim, E., Silva, R.A., Piqueira, J.R.C., Cruz, J.J., Marques, R.P., 2021. Cybersecurity Assessment Framework for Digital Interface Between Safety and Security at Nuclear Power Plants. Int. J. Crit. Infrastruct. Prot. 34.

De Brito, I.B., De Sousa, R.T., 2022. Development of an Open-Source Testbed Based on the Modbus Protocol for Cybersecurity Analysis of Nuclear Power Plants. Appl. Sci. 12 (15).

Dong Y, Cordonnier JB, Loukas A. Attention is not all you need: Pure attention loses rank doubly exponentially with depth. InInternational Conference on Machine Learning 2021 Jul 1 (pp. 2793-2803). PMLR.

Eggers, Shannon Leigh, Robert Walker Youngblood III, Matthew Ryan Overlin, Ruixuan Li, Joseph C. Mahanes, and Katya L. Le Blanc. Digital Engineering and Cybersecurity Decision Analysis in Early Phases of SMR-Driven IES Projects. No. INL/RPT-23-74867-Rev000. Idaho National Laboratory (INL), Idaho Falls, ID (United States), 2023.

El-Genk, M.S., Altamimi, R., Schriener, T.M., 2021. Pressurizer dynamic model and emulated programmable logic controllers for nuclear power plants cybersecurity investigations. Ann. Nucl. Energy 154.

Hahn A, Rowland M, Trask D, Brown R, Tomazic S, Nickerson C, Spirito C. Performance Testing of Cyber Incident Response at Nuclear Power Plant Operators. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); 2022 Jun 1.

Kiranyaz, S., Gastli, A., Ben-Brahim, L., Al-Emadi, N., Gabbouj, M., 2019. Real-Time Fault Detection and Identification for MMC Using 1-D Convolutional Neural Networks. IEEE Trans. Ind. Electron. 66 (11), 8760.

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., Inman, D.J., 2020. 1D convolutional neural networks and applications: A survey. Mech. Syst. Sig. Process. 151.

Lee C, Lee CK, Choi JG, Seong PH. Development of a Demonstrable Nuclear Cyber Security Test-Bed and Application Plans. InKorean Nuclear Society Spring Meeting 2019 May 23 (pp. 23-24).23. Maccarone LT, Hahn AS, Valme R, Rowland MT, Kapuria A, Zhang Y, Cole DG. Advanced Reactor Cyber Analysis and Development Environment (ARCADE) for University Research. InTRTR-IGORR Joint Research Reactor Conference 2023.

Lee C, Song JG, Lee CK, Seong PH. Development of a method for securing the operator's situation awareness from manipulation attacks on NPP process data. Nucl. Eng. Technol. 2021 -12-10;54(6):2011.

Li, Yeni, Hany Abdel-Khalik, Elisa Bertino, and Arvind Sundaram. Development of Defenses against False Data Injection Attacks for Nuclear Power Plants. No. SAND-2018-12994. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2018.

Liu, Y., Zhou, W., Ayodeji, A., Zhou, X., Peng, M., Chao, N., 2020. A multi-layer approach to DN 50 electric valve fault diagnosis using shallow-deep intelligent models. Nuclear. Eng. Technol. 53 (1), 148.

Maccarone, Lee T., Hahn, Andrew S., Rowland, Michael T. System-Level Design Analysis for Advanced Reactor Cybersecurity 2023 October 12.

Nedeljkovic, D., Jakovljevic, Z., 2021. CNN based method for the development of cyber-attacks detection algorithms in industrial control systems. Comput. Secur. 114.

Poresky C, Andreades C, Kendrick J, Peterson P. Cyber security in nuclear power plants: Insights for advanced nuclear technologies. Department of Nuclear Engineering, University of California, Berkeley, Publication UCBTH-17-004. 2017 Sep.

Potluri, Sasanka, Christian Diedrich, Sai Ram Roy Nanduru, and Kishore Vasamshetty. "Development of injection attacks toolbox in MATLAB/Simulink for attacks simulation in industrial control system applications." In 2019 IEEE 17th International Conference on Industrial Informatics (INDIN), vol. 1, pp. 1192-1198. IEEE, 2019.

Rodriguez, Salvador B. Smart Grid Design Development and Cyber Security for Small Modular Reactors. No. SAND2017-4345C. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States), 2017.

Rowland M, Maldonado Rosado S, Hahn A, James J. Use of Modeling and Simulation technologies for Secure By Design (SeBD) Analysis of Advanced Reactors. Sandia National Lab.(SNL-NM), Albuquerque, NM (United States); 2022 Oct 1.

Shin, J., Choi, J., Lee, J., Lee, C., Song, J., Son, J., 2021. Application of STPA-SafeSec for a cyber-attack impact analysis of NPPs with a condensate water system test-bed. Nucl. Eng. Technol. 53 (10), 3319.

Silva RA, Shirvan K, Piqueira JR, Marques RP. Development of the Asherah nuclear power plant simulator for cyber security assessment. InProceedings of the International Conference on Nuclear Security, Vienna, Austria 2020 Feb (pp. 10-14).

Song JG, Lee JW, Lee CK, Lee DY, Choi JG. Preparation for Cyber Security Incident Response Training in Nuclear Power Plants. InIn Transactions of the 2020 Jul 9 (pp. 9-11). KNS Spring Meeting.

Sundaram, A., Li, Y., Abdel-Khalik, H., 2022. Denoising Algorithm for Subtle Anomaly Detection. Nucl. Technol. 208 (9), 1365.

Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior A, Kavukcuoglu K. Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499. 2016 Sep 19;12.

Zhang F, Coble JB. Robust localized cyber-attack detection for key equipment in nuclear power plants. Progress in Nuclear Energy. 2020 -08-28;128.