mathematics

Article

Detecting and Handling Cyber-Attacks in Model Predictive Control of Chemical Processes

Zhe Wu [1], Fahad Albalawi [2], Junfeng Zhang [1], Zhihao Zhang [1] and Helen Durand [3] and Panagiotis D. Christofides [1,4,*]

[1] Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA 90095-1592, USA; zhewu2008@gmail.com (Z.W.); jf-zhang13@mails.tsinghua.edu.cn (J.Z.); zhihaozhang@ucla.edu (Z.Z.)

[2] Department of Electrical and Computer Engineering, Taif University, Taif 21974, Saudi Arabia; eng.fahad19@gmail.com

[3] Department of Chemical Engineering and Materials Science, Wayne State University, Detroit, MI 48202, USA; helen.durand@wayne.edu

[4] Department of Electrical and Computer Engineering, University of California, Los Angeles, CA 90095-1592, USA

* Correspondence: pdc@seas.ucla.edu

Received: 24 August 2018; Accepted: 21 September 2018; Published: 25 September 2018

Abstract: Since industrial control systems are usually integrated with numerous physical devices, the security of control systems plays an important role in safe operation of industrial chemical processes. However, due to the use of a large number of control actuators and measurement sensors and the increasing use of wireless communication, control systems are becoming increasingly vulnerable to cyber-attacks, which may spread rapidly and may cause severe industrial incidents. To mitigate the impact of cyber-attacks in chemical processes, this work integrates a neural network (NN)-based detection method and a Lyapunov-based model predictive controller for a class of nonlinear systems. A chemical process example is used to illustrate the application of the proposed NN-based detection and LMPC methods to handle cyber-attacks.

Keywords: industrial cyber-physical systems; cyber-attacks; neural network; model predictive control; nonlinear chemical processes

1. Introduction

Recently, the security of process control systems has become crucially important since control systems are vulnerable to cyber-attacks, which are a series of computer actions to compromise the security of control systems (e.g., integrity, stability and safety) [1,2]. Since cyber-physical systems (CPS) or supervisory control and data acquisition (SCADA) systems are usually large-scale, geographically dispersed and life-critical systems where embedded sensors and actuators are connected into a network to sense and control the physical devices [3], the failure of cybersecurity can lead to unsafe process operation, and potentially to catastrophic consequences in the chemical process industries, causing environmental damage, capital loss and human injuries. Among cyber-attacks, targeted attacks are severe threats for control systems because of their specific designs with the aim of modifying the control actions applied to a chemical process (for example, the Stuxnet worm aims to modify the data sent to a Programmable Logic Controller [4]). Additionally, targeted attacks are usually stealthy and difficult to detect using classical detection methods since they are designed based on some known information of control systems (e.g., the process state measurement). Therefore, designing an advanced detection system (e.g., machine learning-based detection methods [5,6]) and a suitable optimal control scheme for nonlinear processes in the presence of targeted cyber-attacks is an important open issue.

Due to the rapid development of computer networks of CPS in the past two to three decades, the components (e.g., sensors, actuators, and controllers) in a large-scale process control system are now connected through wired/wireless networks, which makes these systems more vulnerable to cyber-attacks that can damage the operation of physical layers besides cyber layers. Additionally, since the development of most of the existing detection methods still depends partly on human analysis, the increased use of data and the designs of stealthy cyber-attacks pose challenges to the development of timely detection methods with high detection accuracy. In this direction, the design of cyber-attacks, the anomaly detection methods focusing on physical layers, and the corresponding resilient control methods have received a lot of attention. A typical method of detection [4] is using a model of the process and comparing the model output predictions with the actual measured outputs. In [7], a dynamic watermarking method was proposed to detect cyber-attacks via a technique of injecting private excitation into the system. Moreover, four representative detection methods were summarized in [3] as Bayesian detection with binary hypothesis, weighted least squares, χ^2-detector based on Kalman filters and quasi-fault detection and isolation methods.

Besides the detection of cyber-attacks, the design of resilient control schemes also plays an important role in operating a chemical process reliably under cyber-attacks. To guarantee the process performance (e.g., robustness, stability, safety, etc.) and mitigate the impact of cyber-attacks, resilient state estimation and resilient control strategies have attracted considerable research interest. In [2,8], resilient estimators were designed to reconstruct the system states accurately. An event-triggered control system was proposed in [9] to tolerate Denial-of-service (DoS) attacks without jeopardizing the stability of the closed-loop system.

On the other hand, as a widely-used advanced control methodology in industrial chemical plants, model predictive control (MPC) achieves optimal performance of multiple-input multiple-output processes while accounting for state and input constraints [10]. Based on Lyapunov methods (e.g., a Lyapunov-based control law), the Lyapunov-based model predictive control (LMPC) method was developed to ensure stability and feasibility in an explicitly-defined subset of the region of attraction of the closed-loop system [11,12]. Additionally, process operational safety can also be guaranteed via control Lyapunov-barrier function-based constraints in the framework of LMPC [13]. At this stage, however, the potential safety/stability problem in MPC caused by cyber-attacks has not been studied with the exception of a recent work that provides a quantitative framework for the evaluation of resilience of control systems with respect to various types of cyber-attacks [14].

Motivated by this, we develop an integrated data-based cyber-attack detection and model predictive control method for nonlinear systems subject to cyber-attacks. Specifically, a cyber-attack (e.g., a min-max cyber-attack) that aims to destabilize the closed-loop system via a sensor tamper is considered and applied to the closed-loop process. Under such a cyber-attack, the closed-loop system under the MPC without accounting for the cyber-attack cannot ensure closed-loop stability. To detect potential cyber-attacks, we take advantage of machine learning methods, which are widely-used in clustering, regression, and other applications such as model order reduction [15–17], to build a neural network (NN)-based detection system. First, the NN training dataset was obtained for three conditions: (1) The system without disturbances and cyber-attacks (i.e., nominal system); (2) The system with only process disturbances considered; (3) The system with only cyber-attacks considered. Then, a NN detection method is trained off-line to derive a model that can be used on-line to predict cyber-attacks. In addition, considering the classification accuracy of the NN, a sliding detection window is employed to reduce false cyber-attack alarms. Finally, a Lyapunov-based model predictive control (LMPC) method that utilizes the state measurement from secure, redundant sensors is developed to reduce the impact of cyber-attacks and re-stabilize the closed-loop system in finite time.

The rest of the paper is organized as follows: in Section 2, the class of nonlinear systems considered and the stabilizability assumptions are given. In Section 3, we introduce the min-max cyber-attack, develop a NN-based detection system and a Lyapunov-based model predictive controller (LMPC) that guarantees recursive feasibility and closed-loop stability under sample-and-hold implementation within an explicitly characterized set of initial conditions. In Section 4, a nonlinear chemical process example is used to demonstrate the applicability of the proposed cyber-attack detection and control method.

2. Preliminaries

2.1. Notation

Throughout the paper, the notation $|\cdot|$ is used to denote the Euclidean norm of a vector, the notation $|\cdot|_Q$ denotes a weighted Euclidean norm of a vector (i.e., $|x|_Q^2 = x^T Q x$ where Q is a positive definite matrix). x^T denotes the transpose of x. $\mathbf{R}_+$ denotes the set $[0, \infty)$. The notation $L_f V(x)$ denotes the standard Lie derivative $L_f V(x) := \frac{\partial V(x)}{\partial x} f(x)$. For given positive real numbers β and ϵ, $\mathcal{B}_\beta(\epsilon) := \{x \in \mathbf{R}^n \mid |x - \epsilon| < \beta\}$ is an open ball around ϵ with a radius of β. Set subtraction is denoted by "\", i.e., $A \backslash B := \{x \in \mathbf{R}^n \mid x \in A, x \notin B\}$. $\lceil x \rceil$ maps x to the least integer greater than or equal to x and $\lfloor x \rfloor$ maps x to the greatest integer less than or equal to x. The function $f(\cdot)$ is of class $\mathcal{C}^1$ if it is continuously differentiable in its domain. A continuous function $\alpha : [0, a) \to [0, \infty)$ is said to belong to class $\mathcal{K}$ if it is strictly increasing and is zero only when evaluated at zero.

2.2. Class of Systems

The class of continuous-time nonlinear systems considered is described by the following state-space form:

$$\dot{x} = f(x) + g(x)u + d(x)w, \; x(t_0) = x_0 \tag{1}$$

where $x \in \mathbf{R}^n$ is the state vector, $u \in \mathbf{R}^m$ is the manipulated input vector, and $w \in W$ is the disturbance vector, where $W := \{w \in \mathbf{R}^q \mid |w| \leq \theta, \; \theta \geq 0\}$. The control action constraint is defined by $u \in U = \{u_{min} \leq u \leq u_{max}\} \subset \mathbf{R}^m$, where u_{min} and u_{max} represent the minimum and the maximum value vectors of inputs allowed, respectively. $f(\cdot)$, $g(\cdot)$ and $d(\cdot)$ are sufficiently smooth vector and matrix functions of dimensions $n \times 1$, $n \times m$ and $n \times q$, respectively. Without loss of generality, the initial time t_0 is taken to be zero ($t_0 = 0$), and it is assumed that $f(0) = 0$, and thus, the origin is a steady-state of the system of Equation (1) with $w(t) \equiv 0$, (i.e., $(x_s^*, u_s^*) = (0, 0)$). In the manuscript, we assume that every measured state is measured by multiple sensors that are isolated from one another such that if one sensor measurement is tampered by cyber-attacks, a secure network or some secure way can still be used to send the correct sensor measurements of $x(t)$ to the controller. This can also be viewed as secure, redundant sensors or just having an alternative, secure network to send the sensor measurements to the controller. However, if this assumption does not hold, i.e., no secure sensors are available, then the system has to be shut down after the detection of cyber-attacks, or to be operated in an open-loop manner thereafter with an accurate process model.

2.3. Stabilizability Assumptions and Lyapunov-Based Control

Consider the nominal system of Equation (1) with $w(t) \equiv 0$. We first assume that there exists a stabilizing feedback control law $u = \Phi(x) \in U$ such that the origin of the nominal system of Equation (1) can be rendered asymptotically stable for all $x \in D_1 \subset \mathbf{R}^n$, where D_1 is an open neighborhood of the origin, in the sense that there exists a positive definite $\mathcal{C}^1$ control Lyapunov function V that satisfies the small control property and the following inequalities:

$$\alpha_1(|x|) \leq V(x) \leq \alpha_2(|x|), \tag{2a}$$

$$\frac{\partial V(x)}{\partial x} F(x, \Phi(x), 0) \leq -\alpha_3(|x|), \tag{2b}$$

$$\left| \frac{\partial V(x)}{\partial x} \right| \leq \alpha_4(|x|) \tag{2c}$$

where $\alpha_j(\cdot), j = 1, 2, 3, 4$ are class $\mathcal{K}$ functions. $F(x, u, w)$ is used to represent the system of Equation (1) (i.e., $F(x, u, w) = f(x) + g(x)u + d(x)w$).

An example of a feedback control law that is continuous for all x in a neighborhood of the origin and renders the origin asymptotically stable is the following control law [18]:

$$\varphi_i(x) = \begin{cases} -\dfrac{p + \sqrt{p^2 + |q|^4}}{|q|^2} q, & \text{if} \quad q \neq 0 \\ 0, & \text{if} \quad q = 0 \end{cases} \tag{3a}$$

$$\Phi_i(x) = \begin{cases} u_i^{min}, & \text{if} \quad \varphi_i(x) < u_i^{min} \\ \varphi_i(x), & \text{if} \quad u_i^{min} \leq \varphi_i(x) \leq u_i^{max} \\ u_i^{max}, & \text{if} \quad \varphi_i(x) > u_i^{max} \end{cases} \tag{3b}$$

where p denotes $L_f V(x)$ and q denotes $(L_g V(x))^T = [L_{g_1} V(x) \cdots L_{g_m} V(x)]^T$. $\varphi_i(x)$ of Equation (3a) represents the ith component of the control law $\Phi(x)$ before considering saturation of the control action at the input bounds. $\Phi_i(x)$ of Equation (3b) represents the ith component of the saturated control law $\Phi(x)$ that accounts for the input constraints $u \in U$. Based on the controller $\Phi(x)$ that satisfies Equation (2), the set of initial conditions from which the controller $\Phi(x)$ can stabilize the origin of the input-constrained system of Equation (1) is characterized as: $\phi_n = \{x \in \mathbf{R}^n \mid \dot{V} + \kappa V(x) \leq 0,\ u = \Phi(x) \in U, \kappa > 0\}$. Additionally, we define a level set of $V(x)$ inside ϕ_n as $\Omega_\rho := \{x \in \phi_n \mid V(x) \leq \rho\}$, which represents a stability region of the closed-loop system of Equation (1).

3. Cyber-Attack and Detection Methodology

From the perspective of process control systems, cyber-attacks are malicious signals that can compromise actuators, sensors or their communication networks. Specifically, among sensor cyber-attacks, DoS attacks, replay attacks and deception attacks are the three most common and easily implementable ones by attackers [5]. On the other hand, since stealthy cyber-attacks are designed to damage the performance of CPS (e.g., stability and safety), developing more reliable detection and control methods that can detect, locate and mitigate cyber-attacks in a timely fashion and control the damage within a tolerable limit is imperative.

In this section, the min-max cyber-attack designed to damage closed-loop stability of the system of Equation (1) is first introduced. Subsequently, a general model-based detection method [4] and the corresponding stealthy cyber-attacks that can evade such detection are presented. Therefore, to better detect different types of cyber-attacks, the data-based detection scheme that utilizes machine learning methods is finally developed with a sliding detection window.

3.1. Min-Max Cyber-Attack

In this subsection, we first consider a deception sensor cyber-attack, in which the minimum or maximum allowable sensor measurement values are fed into process control systems (e.g., a Lyapunov-based control system with a stability region Ω_ρ defined by a level set of Lyapunov function $V(x)$) to drive the closed-loop states away from their expected values and finally ruin the stability of the closed-loop system. Since $\forall x \in \Omega_\rho$, there exists a feasible control action $u = \Phi(x)$ such that $\dot{V} < 0$, closed-loop stability is maintained within the stability region Ω_ρ under $\Phi(x)$. Assuming that attackers know the stability region of the system of Equation (1) in advance and have access to some of the sensors (but not all), to remain undetectable by a simple stability region-based detection method (i.e., the cyber-attack is detected if the state is out of the stability region), the min-max cyber-attack is designed with the following form such that the fake sensor measurements are still inside Ω_ρ:

$$\bar{x} = \arg\max_{x \in \mathbf{R}} \{V(x) \leq \rho\} \tag{4}$$

where $\bar{x}$ is the tampered sensor measurement. Since the controller needs to get access to true state measurements to maintain closed-loop stability in a state feedback control system, wrong state measurements under cyber-attacks can affect control actions and eventually drive the state away from its set-point. In the section "Application to a chemical process example", it is shown that if attackers apply a min-max cyber-attack to safety-critical sensors (e.g., temperature or pressure sensors in a chemical reactor) in process control systems, closed-loop stability may not be maintained (i.e., the closed-loop state goes out of Ω_ρ) and the system may have to be shut down.

3.2. Model-Based Detection and Stealthy Cyber-Attack

Based on the known process model of Equation (1), a cumulative sum (CUSUM) statistic detection method [4] can be developed to minimize the detection time when a cyber-attack occurs. Specifically, the CUSUM statistic method detects cyber-attacks by calculating the cumulative sum of the deviation between expected and measured states. The method is developed by the following equations:

$$S(k) = (S(k-1) + z(k))^+, \; S(0) = 0 \tag{5a}$$

$$D(S(k)) = \begin{cases} 1, & \text{if } S(k) > S_{TH} \\ 0, & \text{otherwise} \end{cases} \tag{5b}$$

where $S(k)$ is the nonparametric CUSUM statistic and S_{TH} is the threshold of the detection of cyber-attacks. $(S)^+ = S$, if $S \geq 0$ and $(S)^+ = 0$ otherwise. D is the detection indicator where $D = 1$ indicates that the cyber-attack is confirmed or there is no cyber-attack if $D = 0$. $z(k)$ is the deviation between expected states $\tilde{x}(t_k)$ and measured states $x(t_k)$ at time $t = t_k$: $z(k) := |\tilde{x}(t_k) - x(t_k)| - b$ where $\tilde{x}(t_k)$ is derived using the known process model, the state and the control action at $t = t_{k-1}$, and b is a small positive constant to reduce the false alarm rate due to disturbances.

With a carefully selected S_{TH}, the model-based detection method can detect many sensor cyber-attacks efficiently. However, the above model-based method may be evaded and becomes invalid for stealthy cyber-attacks if attackers know more about the system (e.g., the system model and the principles of the detection method). For example, three advanced stealthy cyber-attacks were proposed in [4] to damage the system without triggering the threshold of the model-based detection method. Specifically, a surge cyber-attack is designed to maximize the damage for the first few steps (similar to min-max cyber-attacks) and switch to cyber-attacks with small perturbations for the rest of time when $S(k)$ reaches S_{TH}. The form of a surge cyber-attack is given by the following equations:

$$x(t_k) = \begin{cases} x(t_k)^{min}, & \text{if } S(k) \leq S_{TH} \\ \tilde{x}(t_k) - |S_{TH} + b - S(k-1)|, & \text{otherwise} \end{cases} \tag{6}$$

The above surge cyber-attack is able to maintain $S(k)$ within its threshold and therefore is undetectable by the above detection method. In this case, the defenders should either develop more advanced detection methods for stealthy cyber-attacks (i.e., it becomes an interactive decision-making process between an attacker and a defender [19]), or develop a detection method from another perspective, for example, a data-based method. Since the purpose of any type of stealthy cyber-attack is to change the normal operation and destroy the performance of the system of Equation (1), the dynamic operation of the system of Equation (1) (e.g., dynamic trajectories in state-space) under cyber-attacks becomes different from that of the nominal system of Equation (1). The deviation of the data can be regarded as an intrinsic indicator for detection of cyber-attacks. In this direction, a data-based detection system is developed via machine learning methods in the next subsection.

3.3. Detection via Machine Learning Techniques

Machine learning has a wide range of applications in classification, regression, and clustering problems. To detect cyber-attacks, classification methods can be utilized to determine whether there is a cyber-attack on the system of Equation (1) or not. The data-based learning problems are usually categorized into unsupervised learning and supervised learning.

Unsupervised learning (e.g., k-means clustering) uses unlabeled data to derive a model that can split the data into different categories. On the other hand, supervised learning aims to develop a function that maps an input to an output based on labeled dataset (input-output pairs). There are two types of supervised learning tools, (1) classification tools (e.g., k-nearest neighbor (k-NN), support vector machine (SVM), random forest, neural networks) are used to develop a function based on labeled training datasets to predict the class of a new set of data that was not used in the training stage; (2) regression tools (e.g., linear regression, support vector regression, etc.) aim to predict the outcome of an event based on the relationship between variables obtained from the training datasets (labeled input-output pairs) [20]. Since supervised learning concerns labeled training data, we utilize a neural network (NN) algorithm to predict whether the system of Equation (1) is nominally operating, under disturbances or under cyber-attacks. Subsequently, a Lyapunov-based model predictive controller is proposed to stabilize the closed-loop system during the absence and presence of cyber-attacks.

3.4. NN-Based Detection System

Since the evolution of the closed-loop state from the initial condition $x(0) = x_0 \in \Omega_\rho$ is determined by both the nonlinear system model of Equation (1) and the design of process control systems, it is difficult to distinguish normal operation from the operation under cyber-attacks. Moreover, even if a detection method is developed for a specific cyber-attack (e.g., min-max cyber-attack), the detection strategy is not guaranteed to identify a different type of cyber-attack. Motivated by these concerns, this work proposes a data-based detection system for different types of cyber-attacks by using machine learning methods.

As a widely-used machine learning method, neural networks build a general class of nonlinear functions from input variables to output variables. The basic structure of a feed-forward multiple-input-single-output neural network with one hidden layer is given in Figure 1, where N_{uj}, $j = 1, 2, \ldots, n$ denotes the input variables in the input layer, θ_{1i}, $i = 1, 2, \ldots, h$ denotes the neurons in the hidden layer and N_y denotes the output in the output layer. Specifically, the hidden neurons θ_{1i} and the output N_y (i.e., the classification result) are obtained by the following equations, respectively [21]:

$$\theta_{1i} = \sigma_1\Big(\sum_{j=1}^{n} N_{wij}^{(1)} N_{uj} + N_{wi0}^{(1)}\Big) \tag{7}$$

$$N_y = \sigma_2\Big(\sum_{j=1}^{h} N_{wj}^{(2)} \theta_{1j} + N_{w0}^{(2)}\Big) \tag{8}$$

where σ_1, σ_2 are nonlinear activation functions, $N_{wij}^{(1)}$ and $N_{wj}^{(2)}$ are weights, and $N_{wi0}^{(1)}$, $N_{w0}^{(2)}$ are biases. For simplicity, the input vector $\mathbf{N_u}$ will be used to denote all the inputs N_{uj}, and the weight matrix $\mathbf{N_w}$ will be used to represent all the weights and biases in Equations (7) and (8). The neurons in the hidden layer receive the weighted sum of inputs and use activation functions σ_1 (e.g., ReLu function $\sigma(x) = \max(0, x)$ or sigmoid function $\sigma(x) = 1/(1 + e^{-x})$) to bring in the nonlinearity such that the NN is not a simple linear combination of the inputs. The output neuron generates the class label via a linear combination of hidden neurons and an activation function σ_2 (e.g., sigmoid function for two classes or softmax function $\sigma_i(x) = e^{x_i} / \sum_{k=1}^{K} e^{x_k}$ for multiple classes where K is the number of classes).

Given a set of training data including the input vectors $\mathbf{N_u^i}$, $i = 1, 2, \ldots N_T$ and the corresponding classified labels (i.e., target vectors $\mathbf{N_t^i}$), the NN model is trained by minimizing the following error function (i.e., loss function):

$$\mathbf{E}(\mathbf{N_w}) = \frac{1}{2} \sum_{i=1}^{N_T} |\mathbf{N_y^i}(\mathbf{N_u^i}, \mathbf{N_w}) - \mathbf{N_t^i}|^2 \tag{9}$$

where $\mathbf{N_y^i}(\mathbf{N_u^i}, \mathbf{N_w})$ is the predicted class for the input $\mathbf{N_u^i}$ under $\mathbf{N_w}$. The above nonlinear optimization problem is solved using the stochastic gradient descent (SGD) method, in which the backpropagation method is utilized to calculate the gradient of $\mathbf{E}(\mathbf{N_w})$. Meanwhile, the weight matrix $\mathbf{N_w}$ is updated by the following equation:

$$\mathbf{N_w} := \mathbf{N_w} - \eta \nabla \mathbf{E}(\mathbf{N_w}) \tag{10}$$

where η is the learning rate to control the speed of convergence. Additionally, to avoid over-fitting during the training process, k-fold cross-validation is employed to randomly partition the original dataset into $k - 1$ subsets of training data and 1 subset of validation data, and early-stopping is activated once the error on the validation set stops decreasing.

Finally, the classification accuracy of the validation dataset is utilized to demonstrate the performance of the neural network since the validation dataset is independent of the training dataset and is not used in training the NN model. Specifically, the classification accuracy (i.e., the test accuracy) of the trained NN model is obtained by the following equation:

$$N_{acc} = \frac{n_c}{n_{val}} \tag{11}$$

where n_c is the number of data samples with correct predicted classes, and n_{val} is the total number of data samples in the validation dataset. In general, the NN performance depends on many factors, e.g., the size of dataset, the number of hidden layers and nodes, and the intensity and the amount of disturbance applied [22–24]. In Remark 1, the method of determining the number of layers and nodes is introduced.

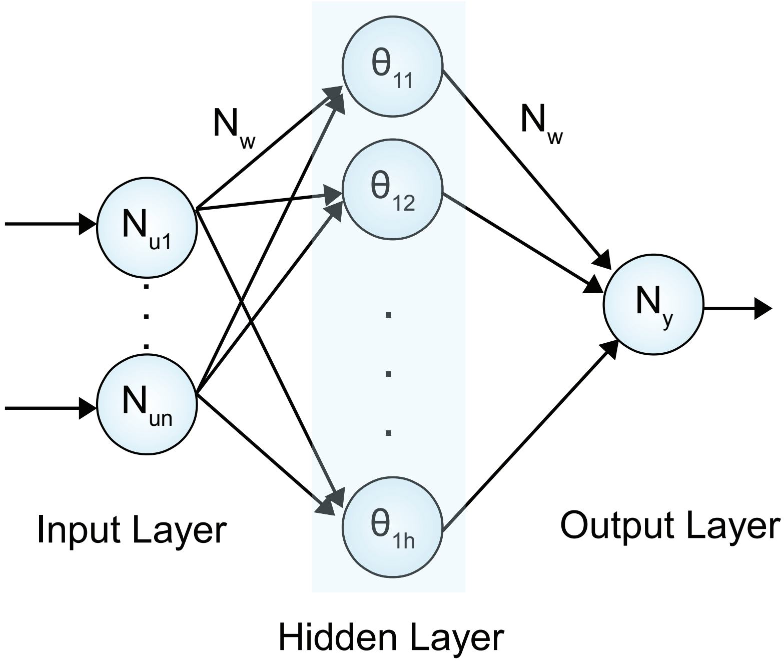

Figure 1. Basic structure of a feed-forward neural network used for cyber-attack detection.

In this paper, the NN is developed to derive a model M to classify three classes: the nominal closed-loop system, the closed-loop system with disturbances, and the closed-loop system under cyber-attacks. A large dataset of time-varying states for various initial conditions (i.e., dynamic trajectories) of the above three cases is used as the input to the neural network. The output of the neural network is the classified class. Since the feed-forward NN is a static model with a fixed input dimension (i.e., fixed time length) but the detection method should be applied during the dynamic operation of the system of Equation (1), multiple NN models with various sizes of input datasets (i.e., various time lengths) are used for the detection of cyber-attacks in real time until the time length corresponding to the available data since the beginning of the time of operation becomes equal to the time length that is preferred to be utilized for the remainder of the operating time. Specifically, given a training dataset of time-series state vectors (i.e., closed-loop trajectories): $N_u \in \mathbf{R}^{n \times T}$ where n is the number of states and T is the number of sampling steps of each trajectory, the NN model is obtained and applied as follows: (1) the NN is trained with data corresponding to time lengths from the initial time to T sampling steps in intervals of N_a sampling steps, i.e., the ith NN model M_i is trained using data from $t = 0$ to $t = iN_a$, where $i = 1, 2, \ldots, T/N_a$ and T is a multiple integer of N_a; (2) when incorporating the NN-based detection system in MPC, real-time state measurement data can be readily utilized in the corresponding NN model M_i to check if there is a cyber-attack so far.

Remark 1. *With an appropriate structure (i.e., number of layers and hidden neurons) of the neural network, the weight matrix $\mathbf{N_w}$ is calculated by Equation (10) and will be utilized to derive the classification accuracy of Equation (11). However, in general, there is no systematic method to determine the structure of a neural network since it highly depends on the number of training data samples and also the complexity of the model needed for classification. Therefore, in practice, the neural network is initiated with one hidden layer with a few hidden neurons. If the classification result is unsatisfactory, we increase the hidden neurons number and further layers with appropriate regularization are added to improve the performance.*

Remark 2. *It is noted that the above classification accuracy of the NN model represents the ratio of the number of correct predictions to the total number of predictions for all classes. If we only consider the case of binary classification (i.e., whether the system is under cyber-attacks or not), sensitivity (also called recall or true positive rate) and specificity (also called true negative rate) are also useful measures. Specifically, sensitivity measures the proportion of actual cyber-attacks that are correctly identified as such, while specificity measures the proportion of actual non-cyber-attacks that are correctly identified as such. Therefore, in the presence of multiple types of cyber-attacks or disturbances, it becomes straightforward to learn the performance of the NN-based method to detect true cyber-attacks via sensitivity and specificity.*

3.5. Sliding Detection Window

Since the classification accuracy of a NN is not perfect, false alarms may be triggered based on a one-time detection (i.e., non-cyber-attack case may be identified as cyber-attack). In order to reduce the false alarm rates, a detection indicator D_i generated by each sub-model M_i and a sliding detection window with length N_s are proposed as follows:

$$D_i = \begin{cases} 1, & \text{if attack is detected by } M_i \\ 0, & \text{if no attack is detected by } M_i \end{cases} \tag{12}$$

Based on the detection indicator D_i at every N_a sampling steps, the weighted sum of detection indicators within the sliding detection window D_I shown in Figure 2 at $t = t_k = k\Delta$ is calculated as follows:

$$D_I = \sum_{j=\lceil (k-N_s+1)/N_a \rceil}^{\lfloor k/N_a \rfloor} \gamma^{\lfloor \frac{k}{N_a} \rfloor - j} D_j \tag{13}$$

where γ is a detection factor that gives more weight to recent detections within the sliding window because the classification accuracy of the NN increases as more data is used for training. If $D_I \geq D_{TH}$, where D_{TH} is a threshold that indicates a real cyber-attack in the closed-loop system, then the cyber-attack is confirmed and reported by the NN-based detection system; otherwise, the detection system remains silent and the sliding window will be rolled one sampling time. To balance false alarms and missed detections, the threshold D_{TH} is determined via extensive closed-loop simulations under cyber-attacks to derive a desired detection rate.

Additionally, since there is no guaranteed feasible control action that can drive the state back towards the origin once the state of the system of Equation (1) is outside the stability region Ω_ρ due to the way of characterizing ϕ_n and Ω_ρ, it is also necessary to check whether the state is in Ω_ρ, especially when cyber-attacks occur but have not been detected yet. Therefore, to prevent the system state from entering a region in state-space where closed-loop stability is not guaranteed, the boundedness of the state vector within the stability region is also checked using the state measurement from redundant, secure sensors at the time when $D_i = 1$. If the state x has already left Ω_ρ, closed-loop stability is no longer guaranteed and in this case further safety system components (e.g., physical safety devices) need to be activated to avoid dangerous operations [25]. However, if $x \in \Omega_\rho$, the state measurement will be read from redundant, secure sensors instead of the original sensors to avoid deterioration of stability under the potential cyber-attack indicated by $D_i = 1$.

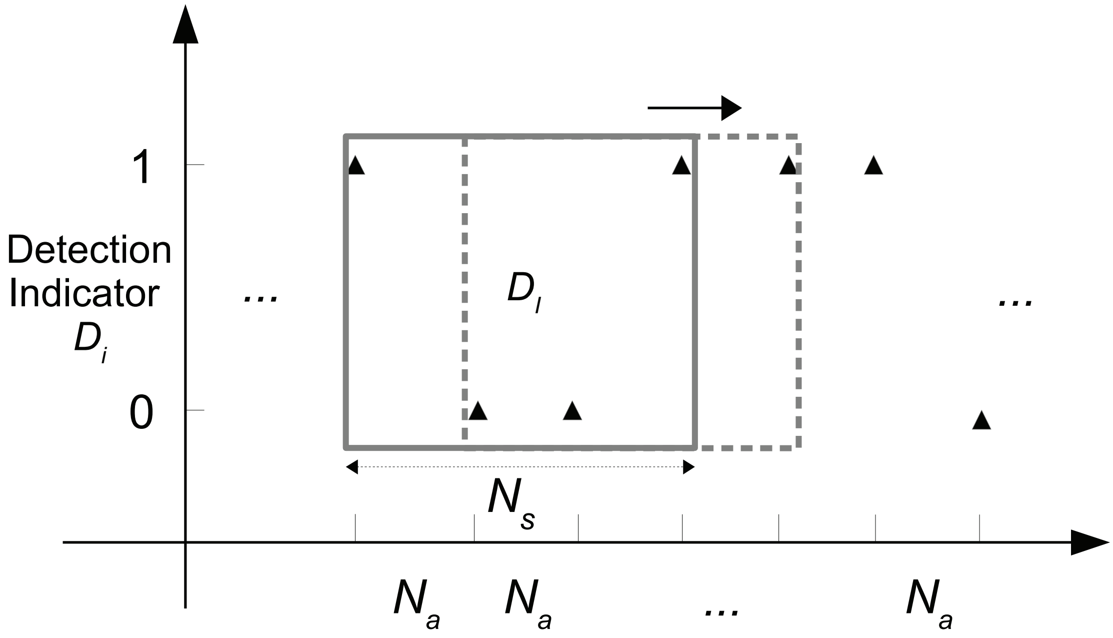

Figure 2. The sliding detection window with detection activated every N_a sampling steps, where triangles represent the detection indicator D_i and the box with length N_s represents the sliding detection window.

Remark 3. *The sliding window with length N_s is employed to reduce false alarm rates. Considering that the classification accuracy derived is not perfect, the idea behind the sliding detection window is that a cyber-attack is confirmed only if it has been detected for a few times continuously instead of a one-time detection. The length of sliding window N_s will balance the efficiency of detection and false alarm rates. Specifically, a larger N_s and a higher detection threshold D_{TH} ($D_I \geq D_{TH}$ within the sliding detection window represents the confirmation of a cyber-attack) lead to longer detection time but a lower false alarm rate, while a smaller N_s and a lower D_{TH} have the opposite effect. Therefore, N_s and D_{TH} should be determined well to achieve a balanced performance between detection efficiency and false alarm rate.*

Remark 4. *The above supervised learning-based cyber-attack detection method is able to distinguish the normal operation of the system of Equation (1) from the abnormal operation under cyber-attacks, provided that there is a large amount of labeled data available for training. However, for those unknown cyber-attacks which are never used for training, the detection is not guaranteed. Specifically, if there exists an unknown cyber-attack that is distinct from the trained cyber-attacks, the NN-based detection method may not be able to identify it as a cyber-attack. In this case, an unsupervised learning-based detection method may achieve better performance by clustering unknown cyber-attack data into a new class. However, if the unknown cyber-attack shares similar* *properties (e.g., similar attack mechanism) with a trained cyber-attack, the NN method may still be able to detect it and classify it as one of the available classes. For example, it is demonstrated in the section "Application to a chemical process example" that the unknown surge cyber-attack can still be detected by the NN-based detection system that is trained for min-max cyber-attacks because of the similarity between these two cyber-attacks.*

Remark 5. *Since different types of cyber-attacks may have various purposes, targeted sensors and attack duration, the dynamic behavior of a closed-loop system varies with different cyber-attacks, which can be eventually reflected by the data of states. Besides the detection of cyber-attacks, the above NN-based detection method is also able to recognize the types of cyber-attacks by training the NN model with data of various types of cyber-attacks labeled as different classes. As a result, the NN model can not only detect the occurrence of cyber-attacks, but also can identify the type of a cyber-attack if the data of that particular cyber-attack has been utilized for training.*

4. Lyapunov-Based MPC (LMPC)

To cope with the threats of the above sensor cyber-attacks, a feedback control method that accounts for the corruption of some sensor measurements should be designed by defenders to mitigate the impact of cyber-attacks and still stabilize the system of Equation (1) at its steady-state. Based on the assumption of the existence of a Lyapunov function $V(x)$ and a controller $u = \Phi(x)$ that satisfy Equation (2), the LMPC that utilizes the accurate measurement from redundant, secure sensors is proposed as the following optimization problem:

$$\mathcal{J} = \min_{u \in S(\Delta)} \int_{t_k}^{t_{k+N}} L_t(\tilde{x}(t), u(t))dt \tag{14a}$$

$$\text{s.t} \quad \dot{\tilde{x}}(t) = f(\tilde{x}(t)) + g(\tilde{x}(t))u(t) \tag{14b}$$

$$\tilde{x}(t_k) = x(t_k) \tag{14c}$$

$$u(t) \in U, \; \forall\, t \in [t_k, t_{k+N}) \tag{14d}$$

$$\dot{V}(x(t_k), u(t_k)) \leq \dot{V}(x(t_k), \Phi(x(t_k))), \quad \text{if } V(x(t_k)) > \rho_{min}, \tag{14e}$$

$$V(\tilde{x}(t)) \leq \rho_{min}, \; \forall\, t \in [t_k, t_{k+N}) \quad \text{if } V(x(t_k)) \leq \rho_{min} \tag{14f}$$

where $\tilde{x}(t)$ is the predicted state trajectory, $S(\Delta)$ is the set of piecewise constant functions with period Δ, and N is the number of sampling periods in the prediction horizon. $\dot{V}(x(t_k), u(t_k))$ represents the time derivative of $V(x)$, i.e., $\frac{\partial V}{\partial x}(f(\tilde{x}(t)) + g(\tilde{x}(t))u(t))$. We assume that the states of the closed-loop system are measured at each sampling time instance, and will be used as the initial condition in the optimization problem of LMPC in the next sampling step. Specifically, based on the measured state $x(t_k)$ at $t = t_k$, the above optimization problem is solved to obtain the optimal solution $u^*(t)$ over the prediction horizon $t \in [t_k, t_{k+N})$. The first control action of $u^*(t)$, i.e., $u^*(t_k)$, is sent to the control actuators to be applied over the next sampling period. Then, at the next sampling time $t_{k+1} := t_k + \Delta$, the optimization problem is solved again, and the horizon will be rolled one sampling time.

In the optimization problem of Equation (14), the objective function of Equation (14a) that is minimized is the integral of $L_t(\tilde{x}(t), u(t))$ over the prediction horizon, where the function $L_t(x, u)$ is usually in a quadratic form (i.e., $L_t(x, u) = x^T Rx + u^T Qu$, where R and Q are positive definite matrices). The constraint of Equation (14b) is the nominal system of Equation (1) (i.e., $w(t) \equiv 0$) to predict the evolution of the closed-loop state. Equation (14c) defines the initial condition of the nominal process system of Equation (14b,14d) defines the input constraints over the prediction horizon. The constraint of Equation (14e) requires that $V(\tilde{x})$ for the system decreases at least at the rate under $\Phi(x)$ at t_k when $V(x(t_k)) > \rho_{min}$. However, if $x(t_k)$ enters a small neighborhood around the origin $\Omega_{\rho_{min}} := \{x \in \phi_n \mid V(x) \leq \rho_{min}\}$, in which $\dot{V}$ is not required to be negative due to the sample-and-hold implementation of the LMPC, the constraint of Equation (14f) is activated to maintain the state inside $\Omega_{\rho_{min}}$ afterwards.

When the cyber-attack is detected by $D_i = 1$ but not confirmed by $D_I \geq D_{TH}$ yet, the optimization problem of the LMPC of Equation (14) uses the state measurement from redundant, secure sensors instead of the original sensors as the initial condition $x(t_k)$ for the optimization problem of Equation (14) until the next instance of detection. However, if the cyber-attack is finally confirmed by $D_I \geq D_{TH}$, the misbehaving sensor will be isolated, and the optimization problem of the LMPC of Equation (14) starts to use the state measurement from secure sensors instead of the compromised state measurement as the initial condition $x(t_k)$ for the optimization problem of Equation (14) for the remaining time of process operation. The structure of the entire cyber-attack-detection-control system is shown in Figure 3.

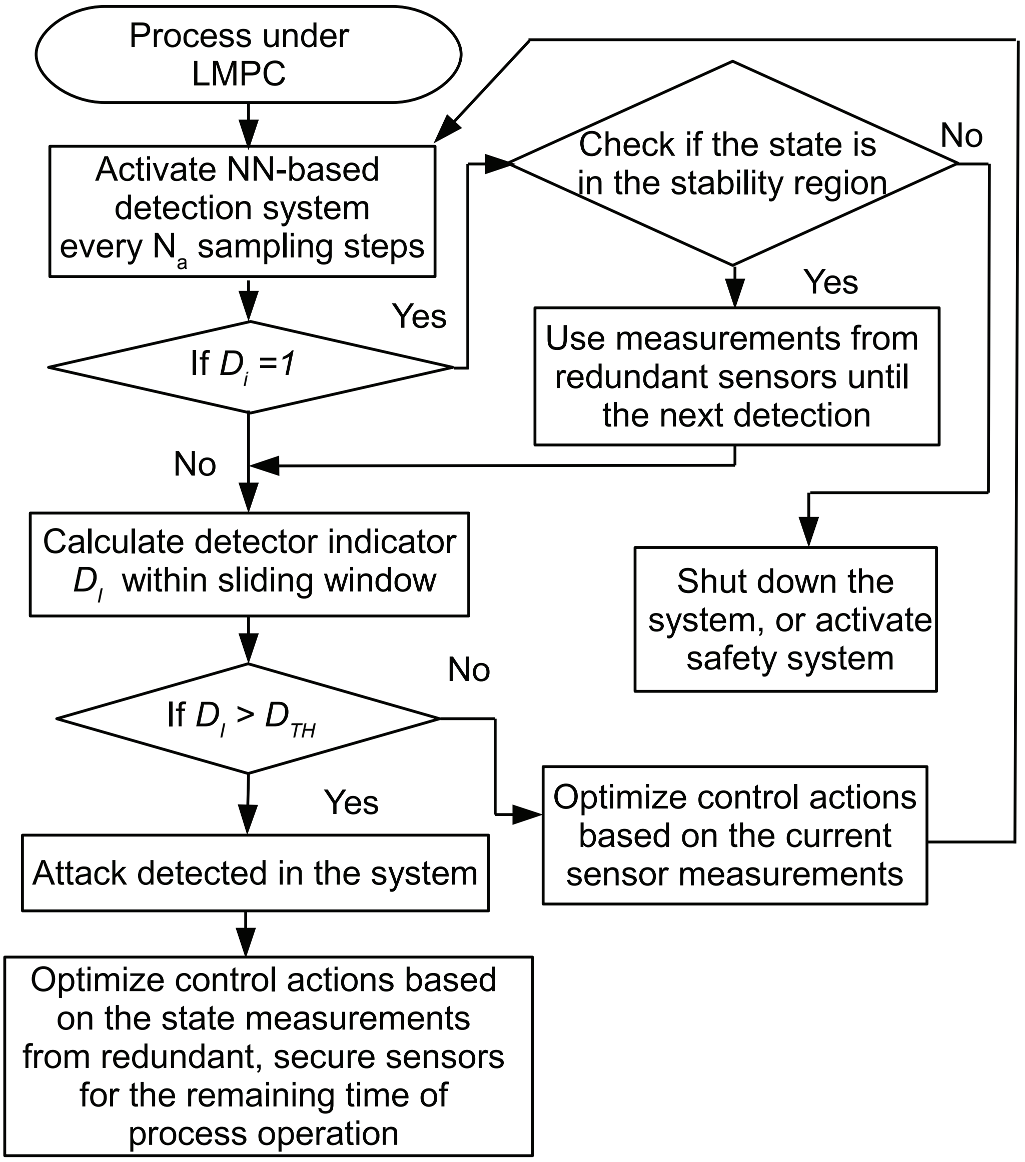
Process under LMPC
Activate NN-based detection system every N_a sampling steps
If D_i =1
Yes
No
Check if the state is in the stability region
No
Yes
Use measurements from redundant sensors until the next detection
Calculate detector indicator D_I within sliding window
Shut down the system, or activate safety system
If D_I > D_TH
No
Yes
Optimize control actions based on the current sensor measurements
Attack detected in the system
Optimize control actions based on the state measurements from redundant, secure sensors for the remaining time of process operation

Figure 3. Basic structure of the proposed integrated NN-based detection and LMPC control method.

If the cyber-attack is detected and confirmed before the closed-loop state is driven out of the stability region, it follows that the closed-loop state is always bounded in the stability region Ω_ρ thereafter and ultimately converges to a small neighborhood $\Omega_{\rho_{min}}$ around the origin for any $x_0 \in \Omega_\rho$

under the LMPC of Equation (14). The detailed proof can be found in [11]. An example trajectory is shown in Figure 4.

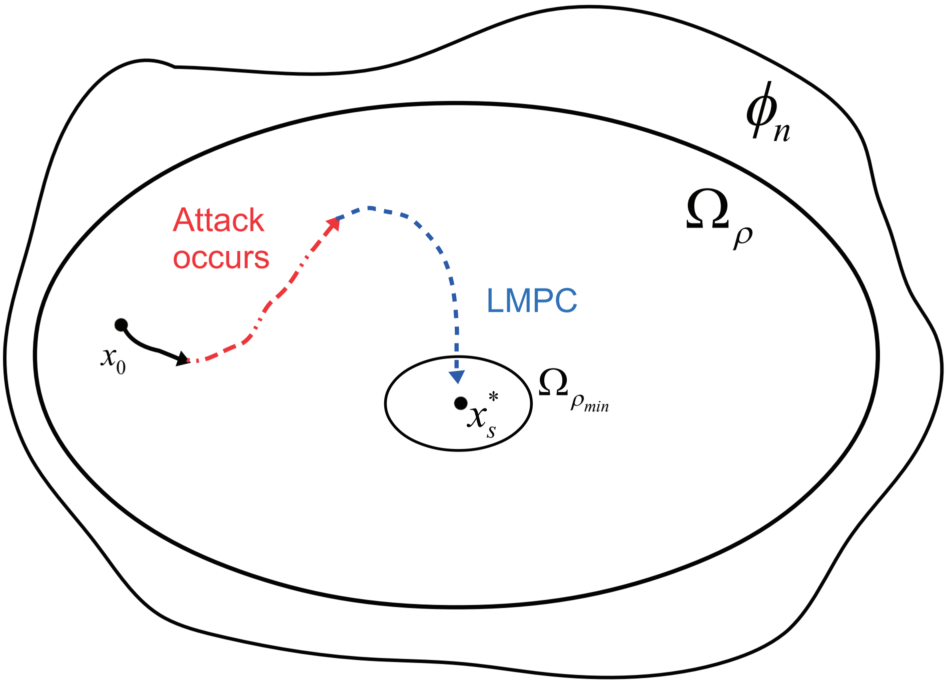

Figure 4. A schematic representing the stability region Ω_ρ and the small neighborhood $\Omega_{\rho_{min}}$ around the origin. The trajectory first moves away from the origin due to the cyber-attack and finally re-converges to $\Omega_{\rho_{min}}$ under the LMPC of Equation (14) after the detection of the cyber-attack by the proposed detection scheme.

Remark 6. *It is noted that the speed of detection (which depends heavily on the size of the input data to the NN, the number of hidden layers and the type of activation functions) plays an important role in stabilizing the closed-loop system of Equation (1) since the operation of the closed-loop system under the LMPC of Equation (14) becomes unreliable after cyber-attacks occur. In other words, if we can detect cyber-attacks in a short time, the LMPC can switch to redundant, secure sensors and still be able to stabilize the system at the origin before it leaves the stability region Ω_ρ. Additionally, the probability of closed-loop stability can be derived based on the classification accuracy of the NN-based detection method and its activation frequency N_a. Specifically, given the classification accuracy $p_{nn} \in [0,1]$, if the NN-based detection system is activated every $N_a = 1$ sampling step, the probability of the cyber-attack being detected at each sampling step (i.e., $D_i = 1$) is equal to p_{nn}, which implies that the probability of closed-loop stability $\forall x_0 \in \Omega_\rho$ is no less than p_{nn}. Moreover, for safety reasons, the region of initial conditions can be chosen as a conservative sub-region (i.e., $\Omega_{\rho_e} := \{x \in \phi_n \mid V(x) \leq \rho_e\}$, where $\rho_e < \rho$) inside the stability region to avoid the rapid divergence of states under cyber-attacks and improve closed-loop stability. For example, let $\rho_e = max\{V(x(t)) \mid V(x(t+\Delta)) \leq \rho, u \in U\}$ such that $\forall x(t_k) \in \Omega_{\rho_e}$, $x(t_{k+1})$ still stays in Ω_ρ despite a miss of detection of cyber-attacks. Therefore, the probability of closed-loop stability $\forall x_0 \in \Omega_{\rho_e}$ under the LMPC of Equation (14) reaches $1-(1-p_{nn})^2$ (i.e., the probability of cyber-attacks being detected within two sampling periods).*

Remark 7. *It is demonstrated in [11] that in the presence of sufficiently small bounded disturbances (i.e., $|w(t)| \leq \theta$), closed-loop stability is still guaranteed for the system of Equation (1) under the sample-and-hold implementation of the LMPC of Equation (14) with a sufficiently small sampling period Δ. In this case, it is undesirable to treat the disturbance as a cyber-attack and trigger the false alarm. Therefore, the detection system should account for the disturbance case and have the capability to distinguish cyber-attacks from disturbances (i.e., the system with disturbances should be classified as a distinct class or treated as the nominal system).*

5. Application to a Chemical Process Example

In this section, we utilize a chemical process example to illustrate the application of the proposed detection and control methods for potential cyber-attacks. Consider a well-mixed, non-isothermal continuous stirred tank reactor (CSTR) where an irreversible first-order exothermic reaction takes place. The reaction converts the reactant A to the product B via the chemical reaction $A \rightarrow B$. A heating jacket that supplies or removes heat from the reactor is used. The CSTR dynamic model derived from material and energy balances is given below:

$$\frac{dC_A}{dt} = \frac{F}{V_L}(C_{A0} - C_A) - k_0 e^{-E/RT} C_A \tag{15a}$$

$$\frac{dT}{dt} = \frac{F}{V_L}(T_0 - T) - \frac{\Delta H k_0}{\rho C_p} e^{-E/RT} C_A + \frac{Q}{\rho C_p V_L} \tag{15b}$$

where C_A is the concentration of reactant A in the reactor, T is the temperature of the reactor, Q denotes the heat supply/removal rate, and V_L is the volume of the reacting liquid in the reactor. The feed to the reactor contains the reactant A at a concentration C_{A0}, temperature T_0, and volumetric flow rate F. The liquid has a constant density of ρ and a heat capacity of C_p. k_0, E and ΔH are the reaction pre-exponential factor, activation energy and the enthalpy of the reaction, respectively. Process parameter values are listed in Table 1. The control objective is to operate the CSTR at the equilibrium point $(C_{As},\ T_s) = (0.57\ \text{kmol/m}^3,\ 395.3\ \text{K})$ by manipulating the heat input rate $\Delta Q = Q - Q_s$, and the inlet concentration of species A, $\Delta C_{A0} = C_{A0} - C_{A0_s}$. The input constraints for ΔQ and ΔC_{A0} are $|\Delta Q| \leq 0.0167$ kJ/min and $|\Delta C_{A0}| \leq 1$ kmol/m^3, respectively.

Table 1. Parameter values of the CSTR.

$T_0 = 310$ K	$F = 100 \times 10^{-3}$ m^3/min
$V_L = 0.1$ m^3	$E = 8.314 \times 10^4$ kJ/kmol
$k_0 = 72 \times 10^9$ min^{-1}	$\Delta H = -4.78 \times 10^4$ kJ/kmol
$C_p = 0.239$ kJ/(kg K)	$R = 8.314$ kJ/(kmol K)
$\rho = 1000$ kg/m^3	$C_{A0_s} = 1.0$ kmol/m^3
$Q_s = 0.0$ kJ/min	$C_{A_s} = 0.57$ kmol/m^3
$T_s = 395.3$ K	

To place Equation (15) in the form of the class of nonlinear systems of Equation (1), deviation variables are used in this example, such that the equilibrium point of the system is at the origin of the state-space. $x^T = [C_A - C_{As}\ T - T_s]$ represents the state vector in deviation variable form, and $u^T = [\Delta C_{A0}\ \Delta Q]$ represents the manipulated input vector in deviation variable form.

The explicit Euler method with an integration time step of $h_c = 10^{-5}$ min is applied to numerically simulate the dynamic model of Equation (15). The nonlinear optimization problem of the LMPC of Equation (14) is solved using the IPOPT software package [26] with the sampling period $\Delta = 10^{-3}$ min.

We construct a Control Lyapunov Function using the standard quadratic form $V(x) = x^T P x$, with the following positive definite P matrix:

$$P = \begin{bmatrix} 9.35 & 0.41 \\ 0.41 & 0.02 \end{bmatrix} \tag{16}$$

Under the LMPC of Equation (14) without cyber-attacks, closed-loop stability is achieved for the nominal system of Equation (15) in the sense that the closed-loop state is always bounded in the stability region Ω_ρ with $\rho = 0.2$ and ultimately converges to $\Omega_{\rho_{min}}$ with $\rho_{min} = 0.002$ around the origin. However, if a min-max cyber-attack is added to tamper the sensor measurement of temperature of the system of Equation (15), closed-loop stability is no longer guaranteed. Specifically, the min-max cyber-attack is designed to be of the following form:

$$\bar{x}_1 = x_1 \tag{17a}$$

$$\bar{x}_2 = \min\{\arg\max_{x_2 \in \mathbf{R}}\{x^T P x \leq \rho\}\} \tag{17b}$$

where $x_1 = C_A - C_{As}$, $x_2 = T - T_s$, and $\bar{x}_1$, $\bar{x}_2$ are the corresponding state measurements under min-max cyber-attacks. In this example, the min-max cyber-attack of Equation (17) is designed such that the measurement of concentration remains unchanged, and the measurement of temperature is tampered to be the minimum value that keeps the state at the boundary of the stability region Ω_ρ.

In Figures 5 and 6, the temperature sensor measurement is intruded by a min-max cyber-attack at time $t = 0.067$ min. Without any cyber-attack detection system, it is shown in Figure 5 that the LMPC of Equation (14) keeps operating the system of Equation (15) using false sensor measurements blindly and finally drives the closed-loop state out of the stability region Ω_ρ.

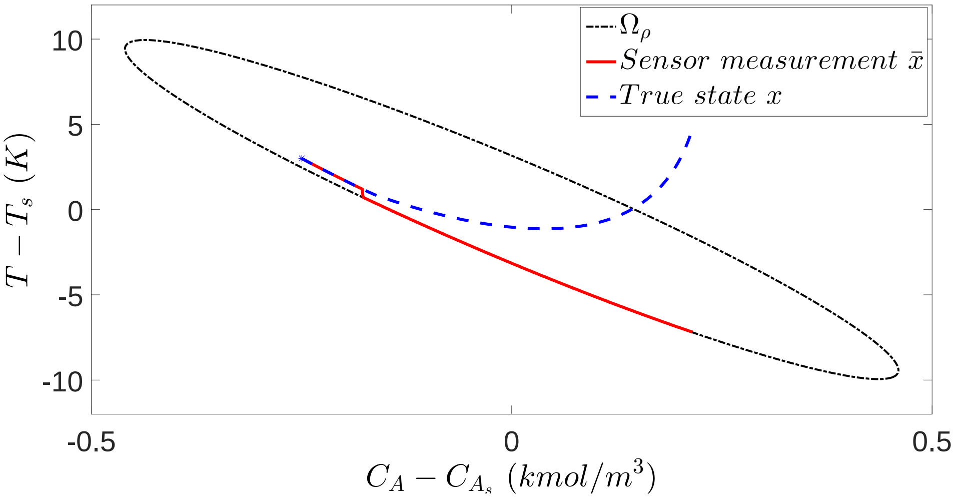

Figure 5. The state-space profile for the CSTR of Equation (15) under the LMPC of Equation (14) and under a min-max cyber-attack for the initial condition (−0.25, 3).

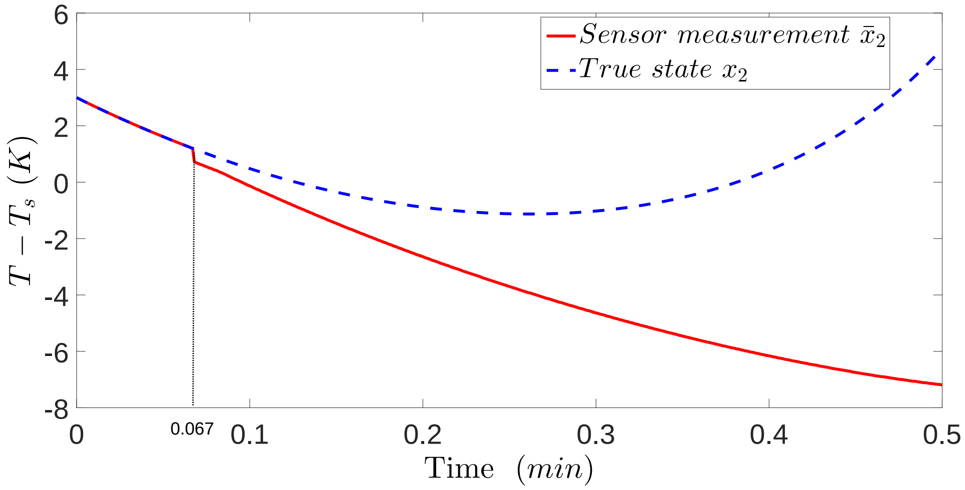

Figure 6. The true state profile ($x_2 = T - T_s$) and the sensor measurements ($\bar{x}_2 = \bar{T} - T_s$) of the closed-loop system under the LMPC of Equation (14) and under a min-max cyber-attack for the initial condition (−0.25, 3), where the vertical dotted line shows the time the cyber-attack is added.

To handle the min-max cyber-attack, the model-based detection system of Equation (5) and the NN-based detection method are applied to the system of Equation (15). The simulation results are shown in Figures 7–13. Subsequently, the application of the NN-based detection method to the system under other cyber-attacks and the presence of disturbances is demonstrated in Figures 14–16. Specifically, we first demonstrate the application of the model-based detection system of Equation (5) and of the LMPC of Equation (14), where $S_{TH} = 1$ and $b = -0.5$ are chosen through closed-loop simulations. In Figure 7, the min-max cyber-attack of Equation (17) is added at 0.06 min and is detected at 0.1 min before the closed-loop state comes out of Ω_ρ. The variation of the CUSUM statistic $S(k)$ is shown in Figure 8, in which $S(k)$ remains at b when there is no cyber-attack and exceeds S_{TH} at 0.1 min. After the min-max cyber-attack is detected, the true states are obtained from redundant, secure sensors and the LMPC of Equation (14) drives the closed-loop state into $\Omega_{\rho_{min}}$.

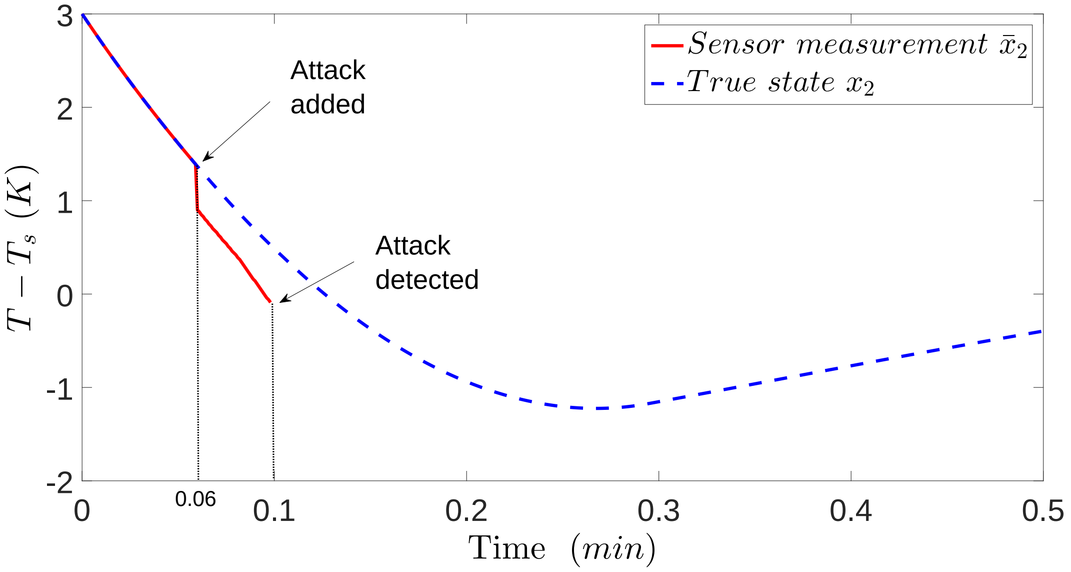

Figure 7. Closed-loop state profiles ($x_2 = T - T_s$, $\bar{x}_2 = \bar{T} - T_s$) for the initial condition $(-0.25, 3)$ under the LMPC of Equation (14) and the model-based detection system.

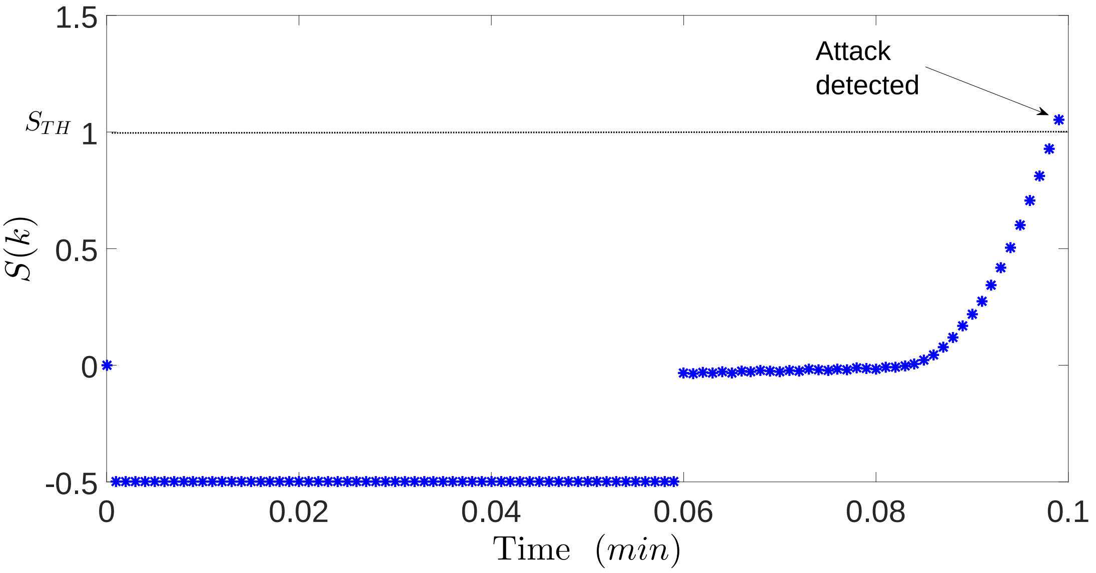

Figure 8. The variation of $S(k)$ for the initial condition $(-0.25, 3)$ under the LMPC of Equation (14) and the model-based detection system.

Next, the NN-based detection system and the LMPC of Equation (14) are implemented to mitigate the impact of cyber-attacks. The feed-forward NN model with two hidden layers is built in Python using the Keras library. Specifically, 3000 time-series balanced data samples of the closed-loop states of the nominal system, the system with disturbances, and the system under min-max cyber-attacks from

$t = 0$ to $t = 1$ min are used to train the neural network to generate the classification of three classes, where class 0, 1, and 2 stand for the system under min-max cyber-attacks, the nominal system and the system with disturbances, respectively. It is demonstrated that 3000 time-series data is sufficient to build the NN for the CSTR example because dataset size smaller than 3000 leads to lower classification accuracy while the increase of dataset size over 3000 does not significantly improve the classification accuracy but brings more computation time as found in our calculations. 3000 data samples are split into 2000 training data, 500 validation data and 500 test data, respectively. $V(x) = x^T P x$ is utilized as the input vector to the NN model. The structure of the NN model is listed in Table 2. Additionally, to improve the performance of the NN model, batch normalization is utilized after each hidden layer to improve the performance of the NN algorithm.

Table 2. Feed-forward NN model.

	Neurons	**Activation Functions**
First Hidden Layer	120	ReLu
Second Hidden Layer	100	ReLu
Output Layer	1	Softmax

To apply the NN-based detection method, we first investigate the relationship of the classification accuracy of the NN with respect to the size of the dataset. Specifically, assuming that the min-max cyber-attack occurs at a random sampling step before 0.1 min, the first NN model $M_{0.1}$ is trained at $t = 0.1$ min using the data of states from $t = 0$ to 0.1 min. As shown in Figure 9, early-stopping is activated at the 8th iteration (epoch) of training when validation accuracy ceases to increase. The averaged classification accuracy at $t = 0.1$ min is obtained by training the same model $M_{t=0.1}$ for 10 times independently. The above process is repeated by increasing the size of the dataset by 0.02 min every time to derive the models for different time instances (i.e., $M_{t=0.12}$, $M_{t=0.14}$, ...). The minimum, the maximum and the averaged classification accuracy at each detection time instance are shown in Figure 10.

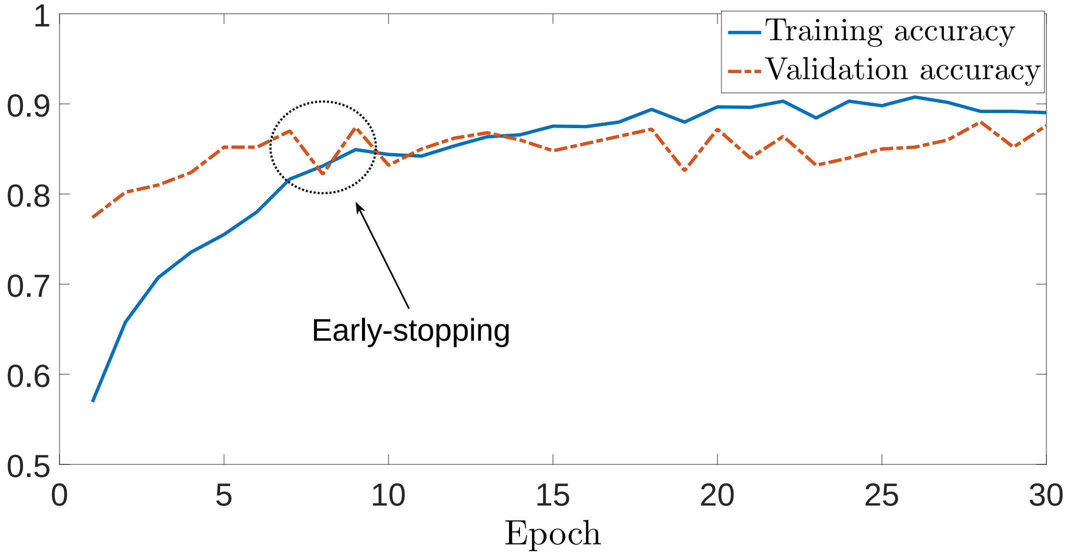

Figure 9. The variation of training accuracy and validation accuracy for the NN model $M_{0.1}$, where early-stopping is activated at the 8th epoch of training.

Figure 10 shows that the averaged test accuracy increases as more state measurements are collected after the cyber-attack occurs, and is up to 95% with state measurements for a long period of time. This suggests that the detection based on recent models is more reliable and deserves higher weights in the sliding window. The confusion matrix of the above NN for three classes: the system under min-max cyber-attack, the nominal system, and the system with disturbances is given in Table 3. Additionally, besides the NN method, other supervised learning-based classification methods including k-NN, SVM and random forests are also applied to the same dataset and obtained the averaged test accuracies, sensitivities and specificities within 0.28 min as listed in Table 4.

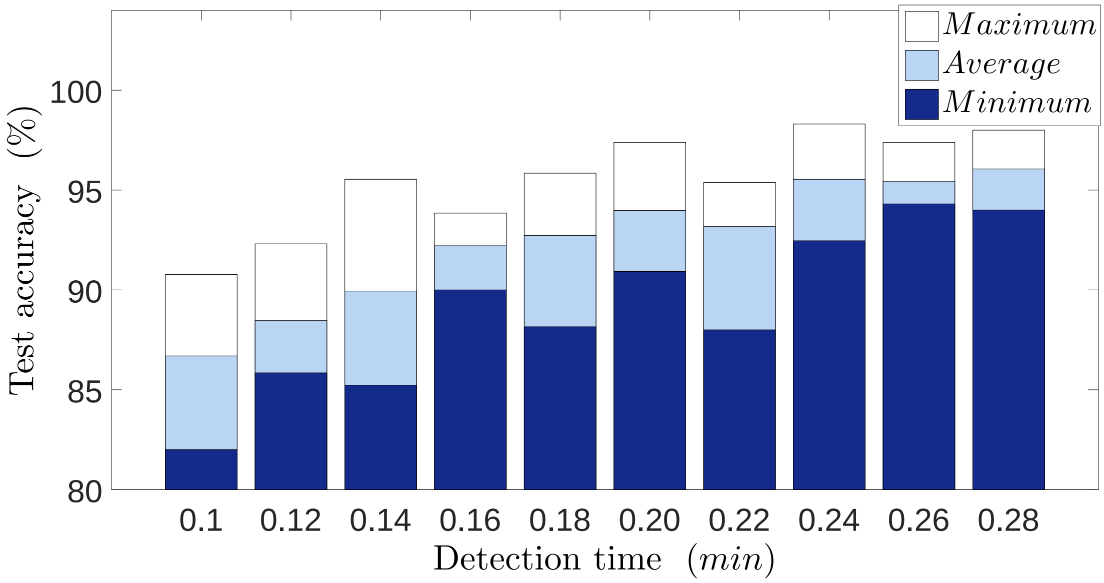

Figure 10. The test accuracy of neural network with respect to the size of training and test data.

Table 3. Confusion matrix of the neural network.

	Actual Class 0: **Min-Max Cyber-Attack**	**Actual Class 1:** **Nominal System**	**Actual Class 2:** **The System with Disturbances**
Predicted Class 0:	198	1	3
Predicted Class 1:	0	140	10
Predicted Class 2:	0	0	148

Table 4. Comparison of the performance of different detection models.

Models	**Test Accuracy**	**Sensitivity**	**Specificity**
k-NN	71.1%	90.9%	99.5%
SVM	83.0%	93.0%	87.8%
Random Forest	96.2%	100.0%	96.2%
Neural Network	95.8%	98.0%	98.6%

When the detection of cyber-attacks is incorporated into the closed-loop system of Equation (15) under the LMPC of Equation (14), the detection system is called every $N_a = 5$ sampling periods. The sliding window length is $N_s = 15$ sampling periods and the threshold for the detection indicator is $D_{TH} = 1.6$. The detection system is activated from $t = 0.1$ min such that a desired test accuracy is achieved with enough data. The closed-loop state-space profiles under the NN-based detection system with the stability region Ω_ρ check and the detection system without the Ω_ρ check are shown in Figures 11 and 12.

Specifically, in Figure 11, it is demonstrated that without the stability region check, the closed-loop state leaves Ω_ρ before the cyber-attack is confirmed. However, under the detection system with the boundedness check of Ω_ρ, the closed-loop state is always bounded in Ω_ρ by switching to redundant sensors at the first detection of min-max cyber-attacks. In Figure 12, it is shown that after the min-max cyber-attack is confirmed at $t = 0.115$ min, the misbehaving sensor is isolated and the LMPC of Equation (14) starts using the measurement of temperature from redundant sensors and re-stabilizes the system at the origin. The simulations demonstrate that it takes around 0.8 min for the closed-loop state trajectory to enter and remain in $\Omega_{\rho_{min}}$ under the LMPC of Equation (14) once the min-max cyber-attack is detected. The corresponding input profiles for the closed-loop system of Equation (1) under the NN-based detection system with the Ω_ρ check are shown in Figure 13, where it is observed that a sharp change of ΔC_{A0} occurs from $t = 0.095$ min to $t = 0.115$ min due to the min-max cyber-attack.

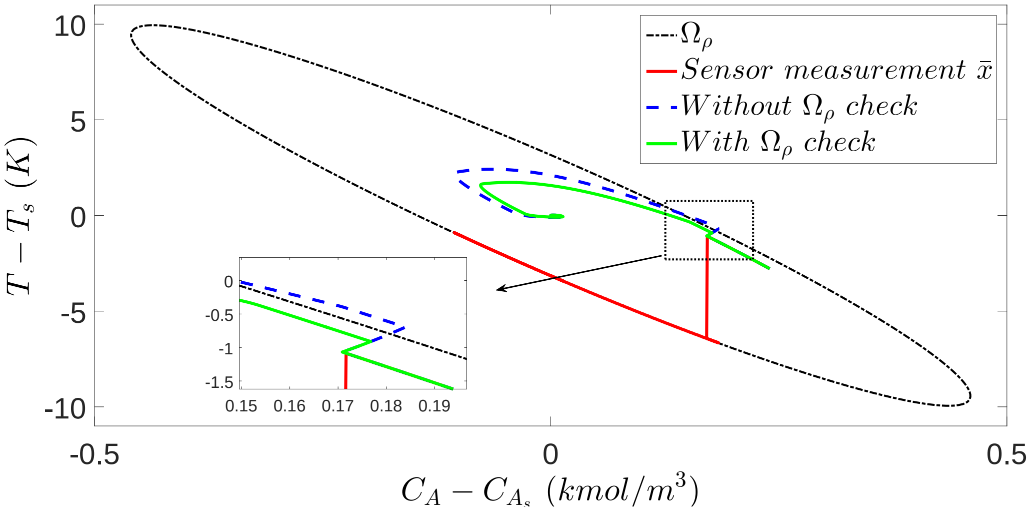

Figure 11. The state-space profile for the closed-loop CSTR with the initial condition (0.24, −2.78), where a min-max cyber-attack is detected by the NN-based detection system and mitigated by the LMPC of Equation (14).

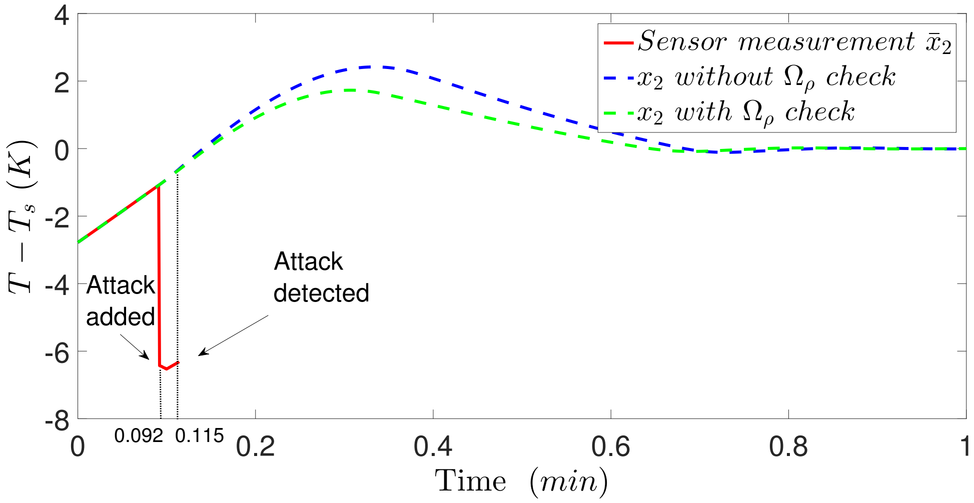

Figure 12. Closed-loop state profiles ($x_2 = T - T_s$, $\bar{x}_2 = \bar{T} - T_s$) for the initial condition (0.24, −2.78) under the LMPC of Equation (14) and the NN-based detection system.

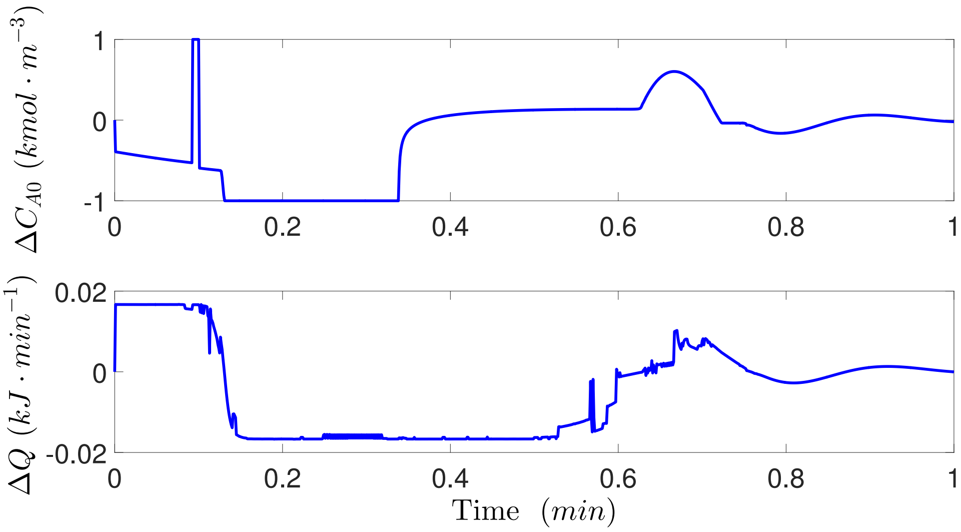

Figure 13. Manipulated input profiles ($u_1 = \Delta C_{A0}$, $u_2 = \Delta Q$) for the initial condition (0.24, -2.78) under the LMPC of Equation (14) and the NN-based detection system.

Additionally, when both disturbances and min-max cyber-attacks are present, it is demonstrated that the NN-based detection system is still able to detect the min-max cyber-attack and re-stabilize the closed-loop system of Equation (15) in the presence of disturbances by following the same steps as in the pure-cyber-attack case. The bounded disturbances w_1 and w_2 are added in Equation (15a,15b) as standard Gaussian white noise with zero mean and variances $\sigma_1 = 0.1$ kmol/(m^3 min) and $\sigma_2 = 2$ K/min, respectively. Also, the disturbance terms are bounded as follows: $|w_1| \leq 0.1$ kmol/(m^3 min), and $|w_2| \leq 2$ K/min, respectively. The closed-loop state and input profiles are shown in Figures 14–16. Specifically, in Figure 15, it is demonstrated that the min-max cyber-attack occurs at 0.08 min and is confirmed at 0.115 min before the closed-loop state leaves Ω_ρ. In the presence of disturbances, the misbehaving sensor is isolated and the closed-loop states are driven to a neighborhood around the origin under the LMPC of Equation (14). In Figure 16, it is demonstrated that the manipulated inputs show variation around the steady-state values (0, 0) when the closed-loop system reaches a neighborhood of the steady-state due to the bounded disturbances.

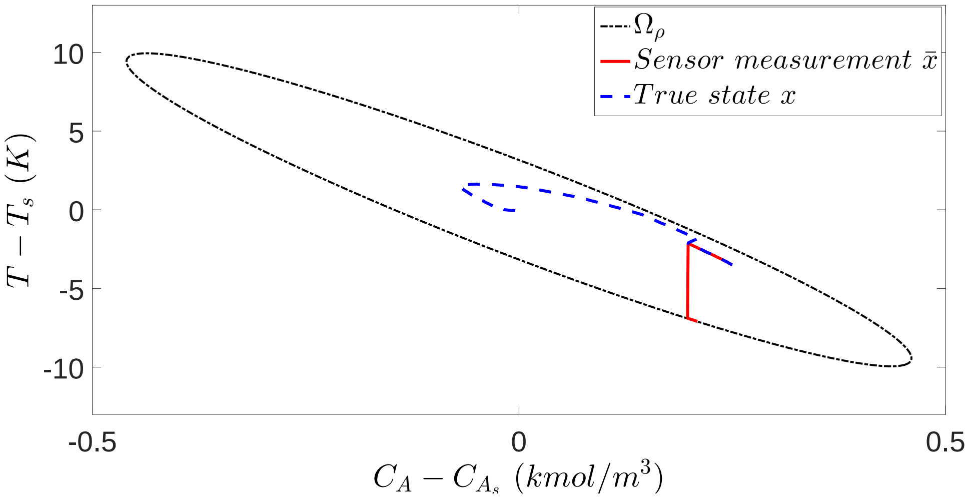

Figure 14. The state-space profiles for the closed-loop CSTR with bounded disturbances and the initial condition (0.25, -3), where a min-max attack is detected by the NN-based detection system and mitigated by the LMPC of Equation (14).

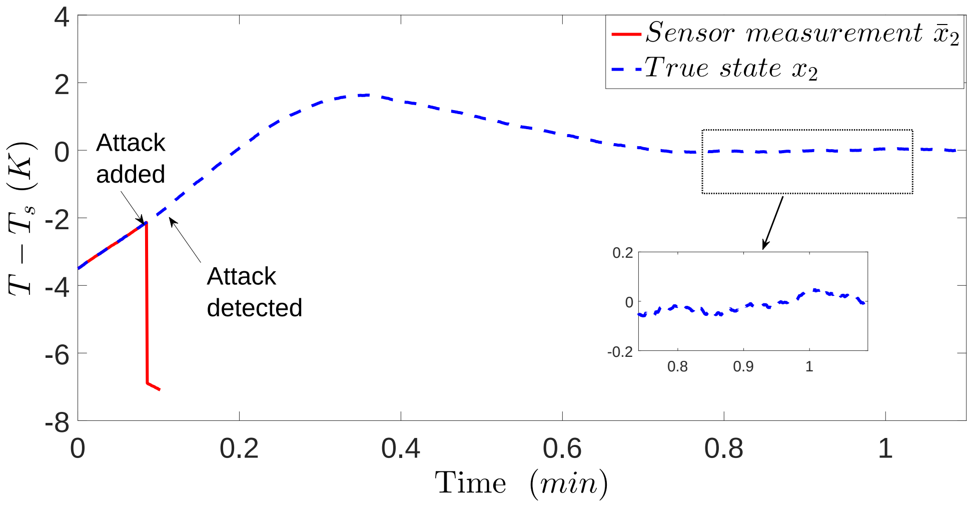

Figure 15. State profiles ($x_2 = T - T_s$, $\bar{x}_2 = \bar{T} - T_s$) for the closed-loop CSTR with bounded disturbances and the initial condition (0.25, -3) under the LMPC of Equation (14) and the NN-based detection system.

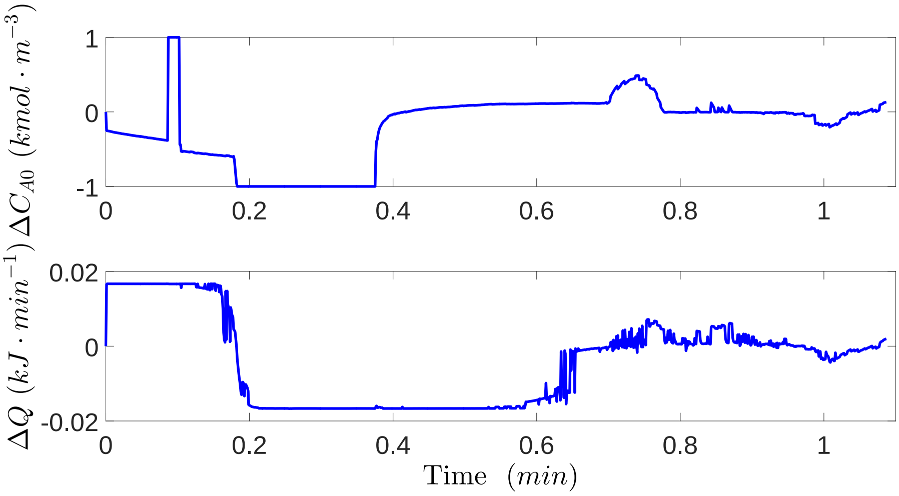

Figure 16. Manipulated input profiles ($u_1 = \Delta C_{A0}$, $u_2 = \Delta Q$) for the closed-loop CSTR with bounded disturbances and the initial condition (0.25, -3) under the LMPC of Equation (14).

Lastly, since the surge cyber-attack of Equation (6) is undetectable by the model-based detection method, we also test the performance of the NN-based detection on the surge cyber-attack due to the similarity between surge cyber-attacks and min-max cyber-attacks (i.e., the surge cyber-attack works as a min-max attack for the first few sampling steps). It is demonstrated in simulations that 89% of surge cyber-attacks can be detected by the NN-based detection system that is trained for min-max cyber-attacks only, which implies that the NN-based detection method can be applied to many other cyber-attacks with similar properties.

Moreover, when cyber-attacks with different properties are taken into account, for example, the replay attack (i.e., $\bar{x} = X$, where X is the set of past measurements of states), the NN-based detection system can still efficiently distinguish the type of cyber-attacks and disturbances by re-training the NN model. The new NN model is built with labeled training data for the case of min-max, replay, nominal and with disturbances, for which the classification accuracy within 0.28 min is up to 85%. As a result, the NN-based detection model can be readily updated with the data of new cyber-attacks without changing the entire structure of detection or control systems.

6. Conclusions

In this work, we proposed an integrated NN-based detection and model predictive control method for nonlinear process systems to account for potential cyber-attacks. The NN-based detection system was first developed with the sliding detection window to detect cyber-attacks. Based on that, the Lyapunov-based MPC was developed with the stability region check triggered by the detection indicator to achieve closed-loop stability in the sense that the closed-loop state remained within a well-characterized stability region and was ultimately driven to a small neighborhood around the origin. Finally, the proposed integrated NN-based detection and LMPC method was applied to a nonlinear chemical process example. The simulation results demonstrated that the min-max cyber-attack was successfully detected before the state exited the stability region, and the closed-loop system was stabilized under the LMPC by using the measurements from redundant secure sensors. The good performance of the proposed approach with respect to surge and replay cyber-attacks was also demonstrated. The value and importance of the NN-based detection method is twofold. First, the NN-based detection method is able to detect cyber-attacks without having to know the process model if a large amount of past data is available. This is very important as nowadays most SCADA systems are large-scale networks with complicated process models, while the big data processing becoming available in both storage and computation. Second, compared to other detection methods, the NN-based detection is easy to implement. The proposed detection and control method can improve the safeness of processes by effectively detecting known (or similar to known) cyber-attacks and also can be readily updated to handle new, unknown cyber-attacks. However, NN-based detection method also has its limitations. Although it achieves desired performance for a trained, known cyber-attack, it is not guaranteed to work for an unknown, new cyber-attack unless it shares similar properties with known cyber-attacks.

Author Contributions: Investigation, Z.W., F.A., J.Z., Z.Z. and H.D.; Methodology Z.W., F.A., J.Z., Z.Z. and H.D.; Writing, Z.W. and H.D.; Supervision, P.D.C.

Funding: Financial support from the National Science Foundation and the Department of Energy is gratefully acknowledged.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding the publication of the research article.

References

1. Ye, N.; Zhang, Y.; Borror, C.M. Robustness of the Markov-chain model for cyber-attack detection. *IEEE Trans. Reliab.* **2004**, *53*, 116–123. [CrossRef]
2. Fawzi, H.; Tabuada, P.; Diggavi, S. Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Trans. Autom. Control* **2014**, *59*, 1454–1467. [CrossRef]
3. Ding, D.; Han, Q.L.; Xiang, Y.; Ge, X.; Zhang, X.M. A survey on security control and attack detection for industrial cyber-physical systems. *Neurocomputing* **2018**, *275*, 1674–1683. [CrossRef]
4. Cárdenas, A.A.; Amin, S.; Lin, Z.S.; Huang, Y.L.; Huang, C.Y.; Sastry, S. Attacks against process control systems: Risk assessment, detection, and response. In Proceedings of the 6th ACM Symposium on Information, Computer And Communications Security, Hong Kong, China, 22–24 March 2011; pp. 355–366.
5. Singh, J.; Nene, M.J. A survey on machine learning techniques for intrusion detection systems. *Int. J. Adv. Res. Comput. Commun. Eng.* **2013**, *2*, 4349–4355.
6. Ozay, M.; Esnaola, I.; Vural, F.T.Y.; Kulkarni, S.R.; Poor, H.V. Machine learning methods for attack detection in the smart grid. *IEEE Trans. Neural Netw. Learn. Syst.* **2016**, *27*, 1773–1786. [CrossRef] [PubMed]
7. Satchidanandan, B.; Kumar, P.R. Dynamic watermarking: Active defense of networked cyber–physical systems. *Proc. IEEE* **2017**, *105*, 219–240. [CrossRef]
8. Pajic, M.; Weimer, J.; Bezzo, N.; Sokolsky, O.; Pappas, G.J.; Lee, I. Design and implementation of attack-resilient cyberphysical systems: With a focus on attack-resilient state estimators. *IEEE Control Syst.* **2017**, *37*, 66–81.
9. Dolk, V.S.; Tesi, P.; De Persis, C.; Heemels, W.P.M.H. Event-triggered control systems under denial-of-service attacks. *IEEE Trans. Control Netw. Syst.* **2017**, *4*, 93–105. [CrossRef]

10. Rawlings, J.B.; Mayne, D.Q. *Model Predictive Control: Theory and Design*; Nob Hill Pub.: San Francisco, CA, USA, 2009.
11. Mhaskar, P.; El-Farra, N.H.; Christofides, P.D. Stabilization of nonlinear systems with state and control constraints using Lyapunov-based predictive control. *Syst. Control Lett.* **2006**, *55*, 650–659. [CrossRef]
12. Muñoz de la Peña, D.; Christofides, P.D. Lyapunov-based model predictive control of nonlinear systems subject to data losses. *IEEE Trans. Autom. Control* **2008**, *53*, 2076–2089. [CrossRef]
13. Wu, Z.; Albalawi, F.; Zhang, Z.; Zhang, J.; Durand, H.; Christofides, P.D. Control Lyapunov-barrier function-based model, predictive control of nonlinear systems. In Proceedings of the American Control Conference, Milwaukee, WI, USA, 27–29 June 2018; pp. 5920–5926.
14. Durand, H. A Nonlinear Systems Framework for Cyberattack Prevention for Chemical Process Control Systems. *Mathematics* **2018**, *6*, 169. [CrossRef]
15. Narasingam, A.; Kwon, J.S.I. Data-driven identification of interpretable reduced-order models using sparse regression. *Comput. Chem. Eng.* **2018**. [CrossRef]
16. Narasingam, A.; Siddhamshetty, P.; Kwon, J.S.I. Temporal clustering for order reduction of nonlinear parabolic PDE systems with time-dependent spatial domains: Application to a hydraulic fracturing process. *AIChE J.* **2017**, *63*, 3818–3831. [CrossRef]
17. Sidhu, H.S.; Narasingam, A.; Siddhamshetty, P.; Kwon, J.S.I. Model order reduction of nonlinear parabolic PDE systems with moving boundaries using sparse proper orthogonal decomposition: Application to hydraulic fracturing. *Comput. Chem. Eng.* **2018**, *112*, 92–100. [CrossRef]
18. Lin, Y.; Sontag, E.D. A universal formula for stabilization with bounded controls. *Syst. Control Lett.* **1991**, *16*, 393–397. [CrossRef]
19. Li, Y.; Shi, L.; Cheng, P.; Chen, J.; Quevedo, D.E. Jamming attacks on remote state estimation in cyber-physical systems: A game-theoretic approach. *IEEE Trans. Autom. Control* **2015**, *60*, 2831–2836. [CrossRef]
20. Tsai, C.F.; Hsu, Y.F.; Lin, C.Y.; Lin, W.Y. Intrusion detection by machine learning: A review. *Expert Syst. Appl.* **2009**, *36*, 11994–12000. [CrossRef]
21. Bishop, C.M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*; Springer: New York, NY, USA, 2006; p. 049901.
22. Alexandridis, K.; Maru, Y. Collapse and reorganization patterns of social knowledge representation in evolving semantic networks. *Inf. Sci.* **2012**, *200*, 1–21. [CrossRef]
23. Daqi, G.; Yan, J. Classification methodologies of multilayer perceptrons with sigmoid activation functions. *Pattern Recognit.* **2005**, *38*, 1469–1482. [CrossRef]
24. Xu, B.; Liu, X.; Liao, X. Global exponential stability of high order Hopfield type neural networks. *Appl. Math. Comput.* **2006**, *174*, 98–116. [CrossRef]
25. Zhang, Z.; Wu, Z.; Durand, H.; Albalawi, F.; Christofides, P.D. On integration of feedback control and safety systems: Analyzing two chemical process applications. *Chem. Eng. Res. Des.* **2018**, *132*, 616–626. [CrossRef]
26. Wächter, A.; Biegler, L.T. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math. Programm.* **2006**, *106*, 25–57. [CrossRef]