

Machine learning modeling and predictive control of nonlinear processes using noisy data

Zhe Wu¹ | David Rincon¹ | Junwei Luo¹ | Panagiotis D. Christofides^{1,2} 

¹Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, California

²Department of Electrical and Computer Engineering, University of California, Los Angeles, California

Correspondence

Panagiotis D. Christofides, Department of Chemical and Biomolecular Engineering, University of California, Los Angeles, CA, USA.
Email: pdc@seas.ucla.edu

Abstract

This work focuses on machine learning modeling and predictive control of nonlinear processes using noisy data. We use long short-term memory (LSTM) networks with training data from sensor measurements corrupted by two types of noise: Gaussian and non-Gaussian noise, to train the process model that will be used in a model predictive controller (MPC). We first discuss the LSTM training with noisy data following a Gaussian distribution, and demonstrate that the standard LSTM network is capable of capturing the underlying process dynamic behavior by reducing the impact of noise. Subsequently, given that the standard LSTM performs poorly on a noisy data set from industrial operation (i.e., non-Gaussian noisy data), we propose an LSTM network using Monte Carlo dropout method to reduce the overfitting to noisy data. Furthermore, an LSTM network using co-teaching training method is proposed to further improve its approximation performance when noise-free data from a process model capturing the nominal process state evolution is available. A chemical process example is used throughout the manuscript to illustrate the application of the proposed modeling approaches and demonstrate their open- and closed-loop performance under a Lyapunov-based model predictive controller with state measurements corrupted by industrial noise.

KEYWORDS

long short-term memory, machine learning, model predictive control, noisy data, nonlinear systems

1 | INTRODUCTION

Data-driven modeling has historically received significant attention in the context of model predictive control (MPC), for example References 1–7. Over the past few decades, system identification methods such as singular value decomposition,⁸ N4SID,⁹ autoregressive models with exogenous inputs,^{10–12} artificial neural networks¹³ have been widely used to develop data-driven models for MPC (see, References 14–18 for reviews of system identification). Among many data-driven modeling approaches, machine learning modeling has demonstrated its superiority in modeling complex nonlinear systems due to the high degree of freedom (e.g., number of neurons, various types of activation functions, and of training algorithms)^{19–22} in capturing

nonlinearities. In order to develop a machine learning model with a desired prediction accuracy for MPC, a high-quality data set that can be generated from industrial process sensors, lab experiments, or extensive computer simulations is required, from which supervised machine learning models can learn the nonlinear relationship between network inputs and outputs. However, real industrial measurements often involve noise stemming from different sources, such as sensors variability and common plant variance, which is a critical point in machine learning modeling that has not been addressed yet in the context of machine-learning-based MPC.

Many methodologies have been proposed for dealing with noisy measurements in the literature. For example, the ARX and ARARX processes have been extended for the case of additive white noise on

the input and output observation.²³ Similarly, subspace identification based on principal component analysis has been proposed by estimating the noise term.²⁴ Under different noise models using closed-loop operation data, several subspace identification methods (i.e., canonical variate analysis, CVA, N4SID, PLS, ARX) were able to identify correctly the process models.²⁵ Additionally, the Kalman filter is one of the most well-known methodologies for state estimation of linear systems under Gaussian white noise.^{26,27} Similarly, state estimation techniques have been proposed for denoising industrial measurements for nonlinear systems, such as the extended Kalman filter, unscented Kalman filter, and moving horizon estimation.²⁶ In general, these methodologies are built under assumptions on the type of noise and system structure.²⁶ As a result, a priori process knowledge such as first-principles models are often needed, which restricts the use of these methods in real industrial systems.^{27,28} On the other hand, the Savitzky–Golay filter is one of the most popular methods for smoothing noisy data²⁹ without any knowledge of process model. This method fits a moving low-order polynomial on adjacent data points by tuning the polynomial order and the size of the data window. However, tuning the parameters could be time demanding and the computational cost is proportional to the window width, which limits its application in real-time control problems. Therefore, how to develop a data-driven process model that can directly predict process dynamic behavior from noisy measurement data remains a challenge.

In recent years, machine learning has attracted an increased level of attention in model identification.^{30–32} Recurrent neural networks (RNN) have been one of the widely used modeling approaches due to their ability of representing temporal dynamic behavior through feedback loops in neurons, and have been successfully incorporated in model predictive control (MPC).^{33,34} When the data training set is not very informative, a model based on principal component analysis (PCA) and RNN has been proposed and tested with MPC in which an alternative closed-loop model re-identification step is proposed.³⁵ Long short-term memory network (LSTM) is a type of RNN model that has been used to maintain information in memory for long periods of time due to its special units in addition to standard units. LSTM has also been used as prediction model in MPC in Reference 36, and has been implemented with economic MPC in its encoder-decoder LSTM structure.³⁷ Additionally, machine learning techniques have also been improved by incorporating feature engineering ideas to efficiently develop models for large-scale systems. For example, in Reference 38, it was demonstrated that both machine learning and projection to latent structure (PLS) methods show poor performance on raw vibration signals from a laboratory-scale water flow system, and therefore, further treatment of raw data sets, such as feature-based monitoring that can significantly improve model prediction^{38–40} is needed. An implementation of similar machine learning structures can be found for nonlinear systems in References 41–43, and their potential role in Industry 4.0 was also highlighted in (bio)chemical processes.⁴⁴ While machine learning techniques have shown great potential in modeling big data sets, machine learning modeling using noisy data has been one of the key issues hampering their implementation to chemical plant data since

most machine learning applications are still limited to deterministic cases (i.e., noise-free data sets) in the literature.²⁷

Many in-silico studies adopt Gaussian noise for testing the robustness of the proposed machine learning methods; however, more realistic conditions such as non-Gaussian noise have not been studied. Unlike Gaussian noise that can be handled by many standard machine learning modeling approaches, non-Gaussian noise may lead to incorrect mapping from input to its (noise-free) ground-truth output due to its overfitting to the noisy pattern of the training data set corrupted by nonstationary noise. A simple, effective technique to reduce overfitting to non-Gaussian noisy data without having any a priori process knowledge is to employ a dropout method in the neural network training process. Specifically, a dropout method randomly drops the connections between units in adjacent layers during training, and provides an efficient way that approximately combines many different neural network architectures together to improve the prediction performance.^{45,46} Recent works^{47,48} have extended dropout to approximate Bayesian inference with theoretical results that provide insights into the use of dropout in RNN models. Moreover, when noise-free process data generated using a priori process knowledge are available (for example, computer simulations based on first-principle models that approximate real processes), the co-teaching method may provide a potential solution to further improve model performance by using the noise-free data set. The co-teaching method was originally proposed to solve classification problems with noisy data (i.e., mislabeled data) in the machine learning community.⁴⁹ Specifically, the co-teaching method with a symmetric or asymmetric mode depending on the noise level of the data set, is able to learn the noise-free pattern through noisy data by using two models that share clean information between each other during each training step.⁴⁹ However, to our knowledge, little attention has been paid to the extension of the co-teaching method in regression problems.

Motivated by the above, this work studies machine learning modeling of nonlinear processes using noisy data via novel neural networks techniques with dropout and co-teaching methods. Specifically, we first investigate the standard LSTM models' capability of modeling noise-free process dynamics using process data with Gaussian noise. Then, we present the Monte Carlo dropout technique and demonstrate the implementation of dropout LSTM in handling noisy industrial-data sets from ASPEN that follows a non-Gaussian distribution. Lastly, we discuss the co-teaching method in the context of regression problems and demonstrate its improved modeling performance by further accounting for noise-free data in the training process. The rest of this article is organized as follows: in Section 2, the notations, the class of nonlinear systems considered, the long short-term memory network and the formulation of LSTM-based model predictive controller are given. In Section 3, we use a chemical process example to study the de-noising capability of LSTM networks using a noisy training data set of Gaussian distribution, and the dropout LSTM modeling approach for handling non-Gaussian noise. In Section 4, we propose the co-teaching scheme using both industrial noisy data and first-principles solutions to improve LSTM training performance. Open-loop and closed-loop simulations under a Lyapunov-based MPC using the aforementioned LSTM models are carried out to compare their performance.

2 | PRELIMINARIES

2.1 | Notation

The notation $|\cdot|$ is used to denote the Euclidean norm of a vector. x^T denotes the transpose of x . The notation $L_f V(x)$ denotes the standard Lie derivative $L_f V(x) := \frac{\partial V(x)}{\partial x} f(x)$. Set subtraction is denoted by “ \setminus ”, that is, $A \setminus B = \{x \in \mathbf{R}^n \mid x \in A, x \notin B\}$. The function $f(\cdot)$ is of class C^1 if it is continuously differentiable in its domain.

2.2 | Class of systems

The class of continuous-time nonlinear systems considered is described by the following system of first-order nonlinear ordinary differential equations:

$$\begin{aligned} \dot{x} &= F(x, u) := f(x) + g(x)u, x(t_0) = x_0 \\ y &= x + w \end{aligned} \quad (1)$$

where $x \in \mathbf{R}^n$ is the state vector, $u \in \mathbf{R}^m$ is the manipulated input vector, $y \in \mathbf{R}^n$ is the vector of state measurements that are sampled continuously, and $w \in \mathbf{R}^n$ is the noise vector. The control actions are constrained by $u \in U := \{u_i^{\min} \leq u_i \leq u_i^{\max}, i = 1, \dots, m\} \subset \mathbf{R}^m$. $f(\cdot)$ and $g(\cdot)$ are sufficiently smooth vector and matrix functions of dimensions $n \times 1$ and $n \times m$, respectively. Throughout the manuscript, we assume that the initial time t_0 is zero ($t_0 = 0$), and $f(0) = 0$ such that the origin is a steady-state of the nominal (i.e., $w(t) \equiv 0$) system of Equation (1) (i.e., $(x_s^*, u_s^*) = (0, 0)$, where x_s^* and u_s^* represent the steady-state state and input vectors, respectively). Note that the machine learning modeling methods that will be developed in this manuscript are not restricted to the system with all states assumed to be measurable. If there exist unmeasured states in an actual process, the machine learning modeling approaches can still be applied using system inputs and outputs only provided that the system dynamics that are not observable from the outputs are asymptotically stable. Alternatively, state estimation techniques can be applied with system inputs and outputs to handle unmeasured states when the system is observable.

2.3 | Stabilization via control Lyapunov function

Consider the nominal system of Equation (1) with noise-free state measurement available (i.e., $y(t) = x(t)$ with $w(t) \equiv 0$). To guarantee that the closed-loop system is stabilizable, a stabilizing control law $u = \Phi(x) \in U$ that renders the origin of the nominal system of Equation (1) (i.e., $w(t) \equiv 0$) exponentially stable is assumed to exist. Following converse Lyapunov theorems,⁵⁰ there exists a C^1 Control Lyapunov function $V(x)$ such that the following inequalities hold for all x in an open neighborhood D around the origin:

$$c_1 |x|^2 \leq V(x) \leq c_2 |x|^2, \quad (2a)$$

$$\frac{\partial V(x)}{\partial x} F(x, \Phi(x)) \leq -c_3 |x|^2, \quad (2b)$$

$$\left| \frac{\partial V(x)}{\partial x} \right| \leq c_4 |x| \quad (2c)$$

where c_1, c_2, c_3 , and c_4 are positive constants. $F(x, u)$ represents the nonlinear system of Equation (1). The universal Sontag control law⁵¹ is a candidate controller for $u = \Phi(x)$. A set of states $\phi_u \in \mathbf{R}^n$ is first characterized where Equation (2) is satisfied under $u = \Phi(x)$. Then a level set of the Lyapunov function inside ϕ_u is used as the closed-loop stability region Ω_ρ for the nonlinear system of Equation (1) as follows: $\Omega_\rho := \{x \in \phi_u \mid V(x) \leq \rho\}$, where $\rho > 0$ and $\Omega_\rho \subset \phi_u$.

2.4 | Long short-term memory network

Recurrent neural networks (RNN) have been utilized in numerous applications to model nonlinear dynamical systems with time-series process data due to their ability to represent temporal behavior using the feedback loops in the hidden layer. Among many types of RNN models, long short-term memory (LSTM) networks have received increasing attention due to their ability to model long-term sequential dependencies with the use of three gates (the input gate, the forget gate, and the output gate) that help avoid vanishing gradients during training.

In this work, we develop an LSTM model with the following general form to approximate the nonlinear system of Equation (1):

$$\dot{\hat{x}} = F_{nn}(\hat{x}, u) := A\hat{x} + \Theta^T z \quad (3)$$

where $\hat{x} \in \mathbf{R}^n$ is the LSTM state vector, and $u \in \mathbf{R}^m$ is the manipulated input vector. $z = [z_1 \cdots z_{n+m+1}]^T = [H(\hat{x}_1) \cdots H(\hat{x}_n) \ u_1 \cdots u_m \ 1]^T \in \mathbf{R}^{n+m+1}$ is a vector of both the network states \hat{x} and the inputs u , where $H(\cdot)$ represents nonlinear activation functions in each LSTM unit, and “1” represents the bias term. $A = \text{diag}\{-\alpha_1 \cdots -\alpha_n\} \in \mathbf{R}^n \times \mathbf{R}^n$ is a diagonal coefficient matrix, and $\Theta = [\theta_1 \cdots \theta_{n+m+1}] \in \mathbf{R}^{(n+m+1) \times n}$ with $\theta_i = \beta_i [\omega_{i1} \cdots \omega_{i(n+m)} \ b_i]$, $i = 1, \dots, n$. α_i and β_i are constants, and ω_{ik} is the weight connecting the k th input to the i th neuron where $i = 1, \dots, n$ and $k = 1, \dots, (n+m)$, and b_i is the bias term for $i = 1, \dots, n$. α_i are assumed to be positive constants such that the state vector \hat{x} is bounded-input bounded-state stable. From the general form of the continuous-time LSTM networks of Equation (3), it is clear that the evolution of LSTM states over time can be obtained based on the current state and manipulated inputs. However, considering that the neural network training data sets typically consist of sampled-data that are discrete in time, the following equations are practically utilized by the LSTM network to calculate the predicted output sequence $\hat{x}(k)$ from the input sequence $m(k)$, $k = 1, \dots, T$:

$$\hat{x}(k) = \sigma(\omega_i^m m(k) + \omega_i^h h(k-1) + b_i) \quad (4a)$$

$$f(k) = \sigma(\omega_f^m m(k) + \omega_f^h h(k-1) + b_f) \quad (4b)$$

$$c(k) = f(k)c(k-1) + i(k)\tanh(\omega_c^m m(k) + \omega_c^h h(k-1) + b_c) \quad (4c)$$

$$o(k) = \sigma(\omega_o^m m(k) + \omega_o^h h(k-1) + b_o) \quad (4d)$$

$$h(k) = o(k)\tanh(c(k)) \quad (4e)$$

$$\hat{x}(k) = \omega_y h(k) + b_y \quad (4f)$$

where $m(k)$ denotes the k th element in the input sequence $m \in \mathbf{R}^{(n+m) \times T}$ that contains the measured states $x \in \mathbf{R}^n$ and the manipulated inputs $u \in \mathbf{R}^m$ with a sequence length of T , and $\hat{x} \in \mathbf{R}^{n \times T}$ denotes the LSTM network output sequence. T is the number of measured states of the sampled-data system of Equation (1). $h(k)$, $c(k)$, $i(k)$, $f(k)$, and $o(k)$ are the internal state, the cell state, the outputs from the input gate, the forget gate, and the output gate, respectively. $\tanh(\cdot)$ is the hyperbolic tangent activation function and $\sigma(\cdot)$ is the sigmoid activation function. ω_f^m and ω_f^h represent the weight matrices for the LSTM input vector m , and the hidden state vector in the input gate, respectively. Similarly, ω_c^m , ω_c^h , ω_i^m , ω_i^h , ω_o^m , ω_o^h represent the weight matrices for the input vector m and the hidden state vector h in calculating the cell state c , the forget gate f , and the output gate o , respectively, with b_f , b_c , b_o , b_y representing the corresponding bias terms. The predicted state vector \hat{x} given by the LSTM network is the linear combination of the internal state h , where ω_y and b_y denote the weight matrix and bias vector for the output, respectively. Readers can refer to³⁶ for both the structures of an unfolded LSTM network (i.e., a general form of RNN model) and of the internal setup of an LSTM network (e.g., forget, input and output gates).

The training and validation data sets for developing the LSTM model are generated using extensive open-loop simulations of the nonlinear system of Equation (1) with various initial conditions $x_0 \in \Omega_\rho$ and control actions $u \in U$. Specifically, the control actions are fed into the nonlinear system of Equation (1) in a sample-and-hold fashion, that is, $u(t) = u(t_k)$, $\forall t \in [t_k, t_{k+1})$, where $t_{k+1} = t_k + \Delta$ and Δ is the sampling period. The continuous-time nonlinear system of Equation (1) is integrated using explicit Euler method with a sufficiently small integration time step $h_c < \Delta$; in practice, time-series process data may be used. In each simulation run, we choose an initial condition and a set of control actions, and carry out the open-loop simulation for the system of Equation (1) over a fixed length of time $(P+1) \cdot \Delta$, where P is an integer and $P \geq 1$. The states x are measured continuously at each integration time step, and therefore, a total of $(P+1) \cdot \frac{\Delta}{h_c}$ measured states are saved in each simulation run. As a result, the LSTM network will use the measured states within the first P sampling periods (i.e., $y(t)$, $\forall t \in [0, P \cdot \Delta]$) and the manipulated inputs $u(t)$, $\forall t \in (\Delta, [P+1] \cdot \Delta)$ to predict the states in the second P sampling period (i.e., $y(t)$, $\forall t \in [\Delta, (P+1) \cdot \Delta]$), from which the predicted state in the last sampling period, that is, $y(t)$, $t \in (P \cdot \Delta, [P+1] \cdot \Delta)$, is obtained based on the previous state measurements. In this case, the number of measured states T in the LSTM network of

Equation (4) is set to $P \cdot \frac{\Delta}{h_c}$, and the input sequence m to the LSTM network is formulated as follows:

$$m = [x(0), x(h_c), \dots, x(P \cdot \Delta - h_c); u(\Delta), u(\Delta + h_c), \dots, u((P+1) \cdot \Delta - h_c)] \quad (5)$$

The LSTM data set is then partitioned into training, validation and testing data sets, and the training process is carried out using the neural-network library, Keras, with an optimal LSTM structure (e.g., number of layers, units, and weight initialization approaches) to achieve a desired model accuracy that satisfies the closed-loop stability requirements of MPC that will be introduced in the next subsection. The LSTM model is developed to predict one sampling period forward; however, by applying the LSTM model in a rolling horizon manner, it can predict future states over a much longer period of time, and therefore, can be used as the prediction model in MPC.

Remark 1. To simplify the discussion of modeling nonlinear systems using LSTM models, Equation (3) presents a general form of one-hidden-layer LSTM (or any RNN) network with n states \hat{x} to approximate the true process state $x \in \mathbf{R}^n$ in Equation (1). However, the LSTM modeling approach is not restricted to a one-hidden layer structure with n states only. In practical implementation, the LSTM output layer corresponds to the states of the nonlinear system of Equation (1) (see Equation (4)), and multiple hidden layers and a sufficient number of LSTM units can be used to improve model accuracy according to the approximation theorem saying that an RNN model with a sufficient number of neurons can approximate any nonlinear dynamic system on compact subsets of the state-space for finite time.⁵² Note that the linear unit is only used between the last hidden layer and the output layer, while nonlinear activation functions such as tanh and sigmoid functions are used in the previous hidden layers to introduce nonlinearities into the LSTM model. Additionally, LSTM gates typically use sigmoid for the input/output/forget gates and tanh for the cell state and the internal state. The saturation limit is to regulate whether there is no flow or complete flow of information through the gates. In addition, tanh function is used because its second derivative can sustain for a long range before going to 0, therefore can overcome the vanishing gradient problem in conventional RNN's.

Remark 2. Note that extensive open-loop simulations under different initial conditions and control actions is not the only way to generate the machine learning data set. The data generation can also be done with a single continuous trajectory (e.g., under a pseudorandom binary input sequence) that is commonly adopted in classical subspace algorithms.

Remark 3. The LSTM inputs and outputs in this work are different from those in Reference 36, which only uses the state measurement $x(t)$ at the current sampling time, for example, $t = t_k$, to predict the states in the following sampling period, that is, x

(t), $\forall t \in (t_k, t_{k+1})$. Although the proposed LSTM structure in Reference 36 works well when noise-free state measurement is available, that is, $w \equiv 0$ in Equation (1), the state prediction based on the current state measurement may show significant deviation from the true state in the presence of sensor noise. As a result, we introduce a window of length P in the LSTM input and output sequences in this work to reduce the dependence of LSTM network on the current state measurement by accounting for past (noisy) state measurements over P sampling periods to provide better predictions.

2.5 | Model predictive control using LSTM models

The Lyapunov-based model predictive control (LMPC) scheme using the LSTM model is given by the following optimization problem:

$$\mathcal{J} = \min_{u \in S(\Delta)} \int_{t_k}^{t_{k+N}} L(\bar{x}(t), u(t)) dt \quad (6a)$$

$$\text{s.t. } \dot{\bar{x}}(t) = F_{nn}(\bar{x}(t), u(t)) \quad (6b)$$

$$u(t) \in U, \forall t \in [t_k, t_{k+N}) \quad (6c)$$

$$\dot{V}(x(t_k), u) \leq \dot{V}(x(t_k), \Phi_{nn}(x(t_k))), \text{ if } x(t_k) \in \Omega_p \setminus \Omega_{\rho_{nn}} \quad (6d)$$

$$V(\bar{x}(t)) \leq \rho_{nn}, \forall t \in [t_k, t_{k+N}), \text{ if } x(t_k) \in \Omega_{\rho_{nn}} \quad (6e)$$

where \bar{x} is the predicted state trajectory, N is the number of sampling periods in the prediction horizon, and $S(\Delta)$ is the set of piecewise constant functions with period Δ . $\Phi_{nn}(x)$ is the stabilizing control law that renders the origin of the LSTM system of Equation (3) exponentially stable. Ω_p is the stability region for the closed-loop LSTM system, and $\Omega_{\rho_{nn}}$ is the target region that the predicted state is ultimately driven into. $\dot{V}(x, u)$ represents the time-derivative of the Lyapunov function V , that is, $\frac{\partial V(x)}{\partial x}(F_{nn}(x, u))$. The LMPC computes the optimal input sequence $u^*(t)$ over the prediction horizon $t \in [t_k, t_{k+N})$, and send the first control action, $u^*(t_k)$, to the system to be applied for the next sampling period. Then, the LMPC horizon is rolled one sampling step forward, and is resolved at the next sampling time.

The optimization problem of Equation (6) is to minimize the time-integral of the function $L(\bar{x}(t), u(t))$ that has its minimum value at the steady-state $(x_s^*, u_s^*) = (0, 0)$ over the prediction horizon subject to the constraints of Equations (6b)–(6e). The constraint of Equation (6b) is the LSTM model of Equation (3) that is used to predict the states of the closed-loop system. Equation (6c) defines the input constraints applied over the entire prediction horizon. The constraint of Equation (6d) drives the closed-loop state toward the origin if $x(t_k) \in \Omega_p \setminus \Omega_{\rho_{nn}}$, in which the time-derivative of V in the constraint of Equation (6d) is approximated using a forward finite difference method. The constraint of Equation (6e) maintains the state within

$\Omega_{\rho_{nn}}$ for the remaining time after the state enters $\Omega_{\rho_{nn}}$. Additionally, since the LSTM model is developed using sampled-data collected every integration time step h_c in open-loop simulation, the state is measured with the same integration time step in MPC and fed into the LSTM model of Equation (6b) at each sampling time. When the noise-free state measurement is available, closed-loop stability is guaranteed for the nonlinear system of Equation (1) under LMPC in the sense that the closed-loop state is bounded in Ω_p for all times and can be ultimately driven into a small neighborhood around the origin for any initial condition $x_0 \in \Omega_p$. The reader is referred to Reference 33 for detailed closed-loop stability and feasibility analysis.

3 | NEURAL NETWORK TRAINING USING NOISY DATA

While LSTM networks developed using noise-free computational simulation data have been demonstrated to be able to predict process dynamics accurately, LSTM training with noisy data has been a challenging research problem when dealing with industrial process data that contain noise. Traditional approaches for reducing noise in time-series data include the use of a filter, for example, Savitzky–Golay filter, to smoothen the data without distorting the signal tendency. However, the filtering performance relies on the sliding time window length, which also affects the computation time in real-time implementation. Therefore, in this section, we discuss LSTM training methods that directly use noisy time-series data to predict true states. Specifically, we will discuss two types of sensor noise, Gaussian noise and non-Gaussian noise in the following subsections.

3.1 | LSTM training with Gaussian noise

We first consider a Gaussian white noise $w \sim \mathcal{N}(0, \sigma^2)$ in sensor measurement y of Equation (1), and demonstrate that the LSTM network is able to predict the true states well using the noisy state measurements following Gaussian distribution. We use a chemical process example to illustrate the de-noising capability of LSTM network with Gaussian noise data. Specifically, a well-mixed, nonisothermal continuous stirred tank reactor (CSTR) where an irreversible second-order exothermic reaction takes place is considered. The reaction transforms a reactant A to a product B ($A \rightarrow B$). The inlet concentration of A , the inlet temperature and feed volumetric flow rate of the reactor are C_{A0} , T_0 , and F , respectively. The CSTR is equipped with a heating jacket that supplies/removes heat at a rate Q . The CSTR dynamic model is described by the following material and energy balance equations:

$$\frac{dC_A}{dt} = \frac{F}{V}(C_{A0} - C_A) - k_0 e^{\frac{-E}{RT}} C_A^2 \quad (7a)$$

$$\frac{dT}{dt} = \frac{F}{V}(T_0 - T) + \frac{-\Delta H}{\rho_L C_p} k_0 e^{\frac{-E}{RT}} C_A^2 + \frac{Q}{\rho_L C_p V} \quad (7b)$$

where C_A is the concentration of reactant A in the reactor, V is the volume of the reacting liquid in the reactor, T is the temperature of the reactor and Q denotes the heat input rate. The concentration of reactant A in the feed is C_{A0} . The feed temperature and volumetric flow rate are T_0 and F , respectively. The reacting liquid has a constant density of ρ_L and a heat capacity of C_p . ΔH , k_0 , E , and R represent the enthalpy of reaction, pre-exponential constant, activation energy, and the ideal gas constant, respectively. Process parameter values can be found in Reference 34, and are omitted here. Additionally, for this particular example, we have shown in Reference 34 that a machine learning model was able to better capture the nonlinear dynamics of the CSTR process than a data-driven linear state-space model using a noise-free data set. Beyond its poor performance under noise-free data set, a linear state-space model is not applicable in this example because the sensor measurements are now corrupted by industrial noise, and thus, the evolution of state variables can no longer rely on the current state measurement.

We study the operation of CSTR under LMPC around the unstable steady-state $(C_{As}, T_s) = (1.95 \text{ kmol/m}^3, 402 \text{ K})$, and $(C_{A0}, Q_s) = (4 \text{ kmol/m}^3, 0 \text{ kJ/hr})$. The manipulated inputs are the inlet concentration of species A and the heat input rate, which are represented by the deviation variables $\Delta C_{A0} = C_{A0} - C_{A0s}$, $\Delta Q = Q - Q_s$, respectively. The manipulated inputs are bounded as follows: $|\Delta C_{A0}| \leq 3.5 \text{ kmol/m}^3$ and $|\Delta Q| \leq 5 \times 10^5 \text{ kJ/hr}$. Therefore, the states and the inputs of the closed-loop system are $x^T = [C_A - C_{As} \quad T - T_s]$ and $u^T = [\Delta C_{A0} \quad \Delta Q]$, respectively, such that the equilibrium point of the system is at the origin of the state-space, (i.e., $(x_s^*, u_s^*) = (0, 0)$). The Lyapunov function $V(x) = x^T P x$ is designed with $P = \begin{bmatrix} 1060 & 22 \\ 22 & 0.52 \end{bmatrix}$. Then, the closed-loop stability region Ω_ρ for the CSTR is characterized as a level set of the Lyapunov function with $\rho = 368$, from which the origin can be rendered exponentially stable under the controller $u = \Phi(x) \in U$.

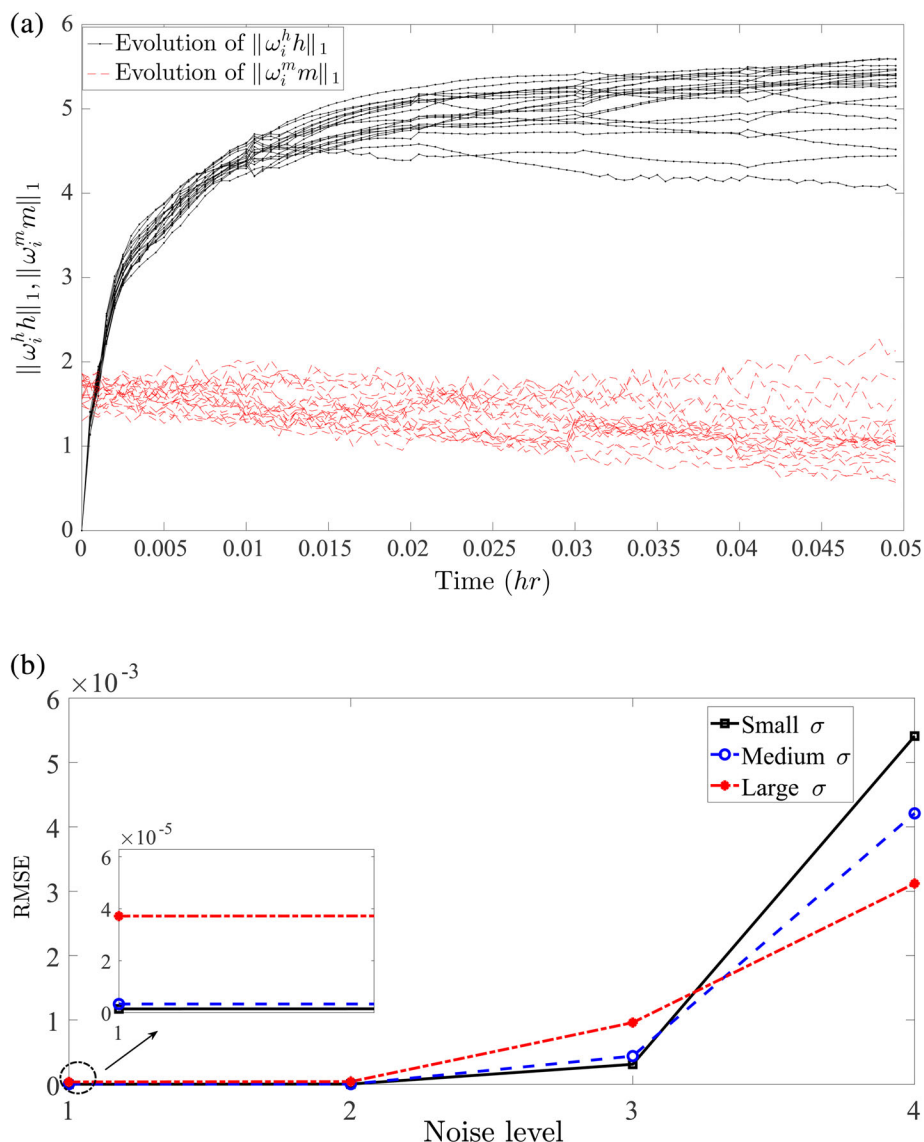
The explicit Euler method with an integration time step of $h_c = 10^{-4} \text{ hr}$ is used to numerically simulate the dynamic model of Equation (7). The LMPC optimization problem is solved using the python module of the IPOPT software package,⁵³ named Pylpopt with the sampling period $\Delta = 10^{-2} \text{ hr}$. The forward finite difference method is used to approximate the first-order derivatives in IPOPT by adding a small perturbation Δu on the optimized variable (i.e., the control actions u). The LSTM models are generated following the standard data generation and learning algorithm in References 33, 54.

We first carry out extensive open-loop simulations of Equation (7) with various initial conditions and control actions to generate the clean and noisy data sets. However, it should be noted that in reality, the noisy data comes from chemical plants, and computational simulations based on first-principles models are only used to generate noise-free data. In open-loop simulation, the state x is measured every integration time step h_c and saved in the clean data set without introducing any measurement noise, while a Gaussian noise $w_1 \sim \mathcal{N}(0 \text{ kmol/m}^3, 2.5 \times 10^{-3} (\text{kmol/m}^3)^2)$ and $w_2 \sim \mathcal{N}(0 \text{ K}, 25 \text{ K}^2)$

are added on concentration and temperature measurements, respectively in the noisy data set. Then, we train the LSTM network using the noisy data set following the standard LSTM training process without accounting for the fact that the data set is corrupted by Gaussian noise. Additionally, an LSTM network using the noise-free data set is also generated as a baseline for comparison. Common techniques that optimize a neural network and perform hyperparameter tuning include preprocessing of data, optimization of network topology, and adjusting learning rate, optimizer, loss function, and number of epochs. In this study, the two LSTM models are developed with the same structure, and the mean squared errors (MSE) between the LSTM predicted values and the actual noise-free value are calculated as follows to indicate the capability of the two LSTMs in approximating true state trajectories: (a) the MSE of x_1 and x_2 for the standard LSTM using clean data set are $3.816 \times 10^{-5} (\text{kmol/m}^3)$ and $0.0837 (\text{K})$, respectively, and (b) the MSE of x_1 and x_2 for the standard LSTM using noisy data set are $1.275 \times 10^{-4} (\text{kmol/m}^3)$ and $1.053 (\text{K})$, respectively. It can be seen that the LSTM trained with clean data set achieves better MSE results than that using noisy data set; however, both networks achieve sufficiently small MSEs, which implies that the standard LSTM formulation is able to suppress the contributions from noisy state measurements, and learn the noise-free pattern that captures the true process dynamics. Following the study in Reference 55, we provide an insight on the de-noising capability of LSTM by computing the relative contributions of the LSTM internal dynamics and of the noisy input (i.e., noisy state measurements). Figure 1a shows the time evolution of the LSTM internal states (i.e., $\|w_i^h h\|_1$ in a black dotted line) and LSTM input network (i.e., $\|w_i^m m\|_1$ in a red dashed line) states for 20 data samples using the same noise level. From any of the data sample, it is observed that the internal state evolution is smoother than the input network and that the norm of the internal states is larger than that of the input network. Therefore, it is concluded that the internal states play a dominant role in the prediction of LSTM outputs when using a noisy data set. This explains why the LSTM network is capable of predicting true states well even using noisy state measurements.

Additionally, we study the de-noising capability of LSTM networks for different noise levels. We calculate the root-mean-square errors (RMSE) of the three LSTM networks trained against different noise levels (i.e., small, medium and large σ in Figure 1b) in predicting true states using the noisy data sets with various noise levels (i.e., noise level 1, 2, 3, and 4 in Figure 1b). Noise level 1, 2, 3, 4 correspond to the white Gaussian noise with $\sigma_1^A = 0.05$, $\sigma_1^T = 5$, $\sigma_2^A = 0.1$, $\sigma_2^T = 10$, $\sigma_3^A = 0.15$, $\sigma_3^T = 15$, and $\sigma_4^A = 0.2$, $\sigma_4^T = 20$ on $x_1 = C_A - C_{As}$ and $x_2 = T - T_s$, respectively, where the subscript number denotes the noise level, and the units are omitted here. The three LSTM networks trained against the small, medium and large noise levels represent the LSTM models using the training data sets with level 1, 2, and 3 Gaussian noise, respectively. It is shown in Figure 1b that the LSTM model trained against the noise with a small σ perform better in small noise level, but is sensitive to the variation of noise level. On the other

FIGURE 1 (a) Time evolution of the long short-term memory (LSTM) internal states $\|\omega_i^h h\|_1$ and the input states $\|\omega_i^m m\|_1$ under Gaussian noise for multiple data sequences, and (b) RMSE in terms of the noise level for LSTM models trained against different noise levels [Color figure can be viewed at wileyonlinelibrary.com]



hand, the LSTM model trained against the noise with a large σ is less sensitive to the noise level since the network relies more on its internal dynamics as shown in Figure 1a. Therefore, Figure 1b provides an insight on the optimal range of noise levels where different LSTM networks can be applied. Additionally, it also explains why the neural networks trained on a noise-free data set become extremely sensitive to small perturbations on testing data, and as a result, we sometimes intentionally introduce noise into neural network training process in order to improve its robustness in practical implementation.

3.2 | LSTM training with non-Gaussian noise

Through the investigation of the case of Gaussian noise in training data sets, we have demonstrated that the standard LSTM is able to reduce the impact of Gaussian noise by utilizing its internal dynamics. In this section, we further study the LSTM capability of handling non-

Gaussian noise and introduce a dropout method that is used to reduce overfitting in LSTM when trained with a noisy data set. While there are many types of non-Gaussian noise that are difficult to formalize the distribution using an equation, we study a specific industrial noise in chemical processes to illustrate the application of LSTM networks.

The top figure in Figure 2 shows the original dimensionless data from ASPEN (public domain data), from which it can be seen that the data noise follows a non-Gaussian distribution due to the irregular drifts. To extract the noise information, a Savitzky-Golay filter is used to approximate the underlying noise-free profile, which is represented by the red line in the top figure of Figure 2. By subtracting the raw data from its noise-free trajectory and then performing normalization, the main feature of this industrial data noise is extracted, which is shown as the normalized noise in the bottom of Figure 2. In the following simulations, the normalized industrial data noise will be added into LSTM data sets generated from open-loop simulations of the CSTR of Equation (7) to mimic the industrial data for the CSTR process we considered.

3.3 | LSTM networks using dropout layers

The dropout technique is widely used in neural networks to prevent overfitting by randomly masking network units.⁴⁶ Although the standard LSTM model is demonstrated to predict process dynamics well using the sensor measurements with Gaussian noise, it performs poorly on training data corrupted by non-Gaussian noise as shown in Figure 3. In this section, we take advantage of Monte Carlo dropout method (MC dropout) that was recently proposed in References 47, 48, in which RNN models using MC dropout method were interpreted as probabilistic models. Specifically, the RNN weights are treated as random variables, and the posterior distribution of the RNN model weights is obtained by sampling the network with randomly dropped out weights at test time (termed Monte Carlo samples). It was demonstrated in Reference 48 that by using dropout techniques, a large RNN model with a sufficient number of neurons may lead to improved denoised results compared to small RNN models that were commonly used in the past to avoid overfitting.

To simplify the discussion, we present the Monte Carlo method using the general form of LSTM models in Equation (3). However, the

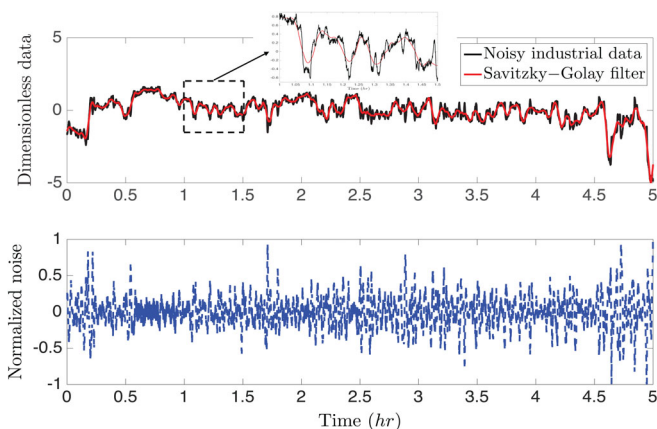


FIGURE 2 Noisy industrial-data sets from ASPEN (black line in the top figure), the denoised result using Savitzky-Golay filter (red line in the top figure), and the extracted (normalized) noise from ASPEN industrial data (bottom figure) [Color figure can be viewed at wileyonlinelibrary.com]

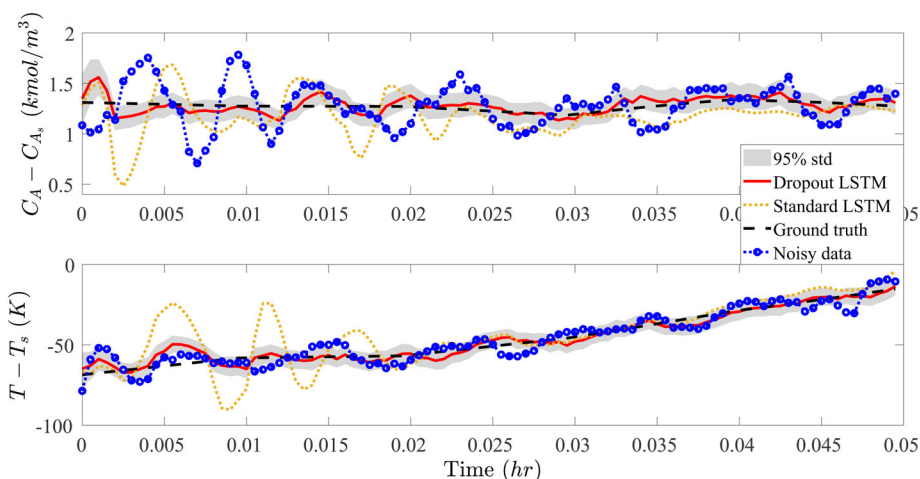


FIGURE 3 State profiles predicted by the dropout long short-term memory (LSTM) and the standard LSTM, where the red line is dropout LSTM, the black, dashed line is the ground truth, the yellow line is the standard LSTM, and the blue, dotted line is the noisy state measurement [Color figure can be viewed at wileyonlinelibrary.com]

following discussion can be readily generalized to the detailed LSTM network of Equation (4), and other RNN models, for example, gated recurrent unit (GRU) as shown in Reference 48. Let $\mathbf{W} = \{W_i\}_{i=1}^L$ denote the weight matrix of the LSTM model of Equation (3) including all the weights and bias terms to be optimized, where W_i is the weight matrix of dimension $K_i \times K_{i-1}$ for each LSTM layer i , and L is the number of layers. Given the LSTM training data that include the data pair of (\mathbf{M}, \mathbf{X}) , where \mathbf{M} and \mathbf{X} represent the LSTM input and output matrices, respectively, the goal of the LSTM model using MC dropout method (termed dropout LSTM) is to find the posterior distribution over the weights $p(\mathbf{W} | \mathbf{M}, \mathbf{X})$. We first define a binary variable z_{ij} of Bernoulli distribution following⁴⁷:

$$z_{ij} \sim \text{Bernoulli}(p_i) \quad (8)$$

where $z_{ij} = 0$, $i = 1, \dots, L$, $j = 1, \dots, K_{i-1}$ represents the j th weight between layer $i-1$ and layer i being dropped out with probability $1 - p_i$, and $z_{ij} = 1$ represents the weight remaining unchanged with probability p_i . Therefore, the weight matrix W_i can be represented as follows for $i = 1, \dots, L$.

$$W_i = B_i \cdot \text{diag}(z_i) \quad (9)$$

where B_i are the variational variables to be optimized. Since the posterior distribution $p(\mathbf{W} | \mathbf{M}, \mathbf{X})$ is intractable in practice, Reference 47 proposed to use the approximating distribution $q(\mathbf{W})$ of Equations (8) and (9) and minimize the Kullback-Leibler (KL) divergence between the full posterior and $q(\mathbf{W})$. Therefore, an approximate predictive distribution of LSTM output can be given as follows:

$$p(\mathbf{x}^* | \mathbf{m}^*, \mathbf{X}, \mathbf{M}) = \int p(\mathbf{x}^* | \mathbf{m}^*, \mathbf{W}) q(\mathbf{W}) d\mathbf{W} \quad (10)$$

where \mathbf{m}^* is the LSTM input in testing data sets, \mathbf{x}^* is the corresponding LSTM predicted output, and \mathbf{X}, \mathbf{M} are the LSTM inputs and outputs in training data set. Additionally, Equation (10) can be approximated by performing Monte Carlo dropout at test time and calculating the averaged results as follows:

$$p(\mathbf{x}^* | \mathbf{m}^*, \mathbf{X}, \mathbf{M}) \approx \frac{1}{N_t} \sum_{k=1}^{N_t} p(\mathbf{x}^* | \mathbf{m}^*, \mathbf{W}_k) \quad (11)$$

where $\mathbf{W}_k \sim q(\mathbf{W})$, and N_t is the number of realizations in dropout LSTM. It is noted that unlike normal dropout approach that does not apply dropout at test time, the main difference of MC dropout is that it applies dropout at both training and testing phases, thereby the LSTM prediction is no longer deterministic. Given a same testing input, by performing dropout LSTM prediction multiple times, we will be able to generate random predictions, from which an approximate probabilistic distribution of LSTM output is obtained. In this way, dropout LSTM provides a potential solution to learn the ground truth (i.e., nominal state trajectories) from sensor measurement data corrupted by a complex, non-Gaussian noise. This is conceptually equivalent to the stochastic optimization problem that uses Monte Carlo method to simulate an uncertain process to predict state propagation. In fact, the dropout LSTM (i.e., the LSTM model using MC dropout) itself can be considered as a complex nonlinear process model with the dropped LSTM weights being uncertain variables. The application of dropout LSTM is illustrated using the same CSTR example that has been introduced in the previous section, and the simulation results are presented below.

The open-loop prediction results for standard LSTM and dropout LSTM are summarized in Table 3, and will be discussed in detail after we introduce the co-teaching method in the next section. Overall, it can be concluded from Table 3 that the dropout LSTM achieves better prediction results in terms of smaller validation MSE. Figure 3 shows one of the open-loop prediction results from the dropout LSTM and the standard LSTM using the same noisy data set for training. Specifically, we ran dropout LSTM prediction 300 times to obtain the distribution of predicted state trajectories, from which we show the mean state trajectory (red line) and the 95% SD interval (gray region) in Figure 3. It is demonstrated that the mean state trajectory predicted by the dropout LSTM is much closer to the ground truth (i.e., the nominal state trajectory in black) compared to the state trajectory (yellow line) predicted by the standard LSTM.

Since dropout LSTM uses Monte Carlo simulation to perform forward prediction, which is one of the most time-consuming parts in solving MPC,⁶ whether nonlinear first-principles models or machine learning models are used, parallel computing is utilized to reduce the computation time by running multiple realizations in different cores/

TABLE 1 Comparison of computation time under the standard long short-term memory (LSTM) and dropout LSTM networks

Model	Computation time (seconds)
(1) Standard LSTM	0.0165
(2) Dropout LSTM with serial computing	0.0536
(3) Dropout LSTM with parallel computing	0.0183

nodes. Note that the time complexity of the forward pass algorithm of machine learning models (more precisely, the number of multiplications involved in the linear combinations) not only depends on the dimension of inputs, but also depends on the number of layers and the size of each layer, which could be large for a complex nonlinear system. The parallelization of dropout LSTM is particularly useful when it is incorporated in MPC for real-time control implementation. Table 1 summarizes the averaged computation time for predicting one state trajectory using standard LSTM, and dropout LSTM under serial and parallel computing, respectively with a testing data set of 100 data samples. It is noted that in the parallel implementation of dropout LSTM, 60 nodes were reserved in UCLA Hoffman 2 computing cluster to carry out 60 simulation runs such that each simulation run was assigned to a single worker node. The mean state trajectory and standard deviation are computed in the host node after synchronization operation. It is clearly seen from Table 1 that the computation time increases dramatically as the number of simulation runs increase under serial computation of dropout LSTM. On the other hand, the parallel computation of dropout LSTM can instead significantly reduce the computation time to the level that is comparable to the standard LSTM.

Remark 4. The standard and dropout LSTMs in serial mode perform the function evaluations in the host node only, while the dropout LSTM predictions in parallel mode are performed in the worker nodes, where each worker node calculates one dropout realization. Note that the computation time reported in Table 1 accounts for the time of data normalization and denormalization before and after LSTM predictions as well as the time of the synchronization step (in parallel computing). The computation time for all of these pre-processing and post-processing calculations is around 0.0158 s. Therefore, the

TABLE 2 Tuning the threshold for the symmetry co-teaching methodology

Threshold	Clean seqs./total seqs.
0.01	(267/306)
0.003	(96/102)
0.002	(37/38)
0.0014	(22/23)

TABLE 3 Statistical analysis of the open-loop predictions under non-Gaussian noise

Methods	MSE x_1	MSE x_2
(1a) LSTM: noise-free data only	0.0011	8.2056
(1b) LSTM: mixed data	0.0258	22.1795
(1c) LSTM: noisy data only	0.0328	29.5571
(2) Co-teaching LSTM	0.0053	7.2123
(3) Dropout LSTM	0.0052	19.2123

Abbreviation: LSTM, long short-term memory.

computation time for prediction step only is 0.0378 s for serial dropout LSTM, which is around 15 times longer than 0.0025 s for parallel dropout LSTM. It is noticed that the computation time for serial dropout LSTM is not exactly 60 times longer than that for parallel dropout LSTM because in parallel computing, a synchronization step is performed to ensure that each task in worker node blocks until all tasks in the computing group reach the host node. As a result, the computation time for parallel dropout LSTM depends heavily on the slowest worker node. Despite the slight performance degradation due to synchronization step, the computational efficiency is still much improved under parallel computing of dropout models.

4 | CO-TEACHING METHOD

Unlike the standard and dropout LSTM networks that use noisy data only for training, the co-teaching method that we will discuss in this section takes advantage of noise-free data sets that can be obtained from first-principles modeling and simulation of chemical processes to further improve LSTM prediction accuracy. Co-teaching has been originally proposed for the classification problem with noisy labels,^{49,56} for example, in image classification, one of the most popular applications of neural networks that classifies the images into one of a number of predefined classes. However, in many cases, the training data set for image classification task is not totally clean in the sense that some images are mislabeled, where the term “noisy label” is often used to represent the image data with incorrect labels. Without any treatment on data set, training neural networks using such a

noisy data set may lead to undesired model accuracy. However, cleaning noisy labels manually also appears impractical when dealing with a high-dimensional data set.

Co-teaching is an algorithm that trains models with noisy labels by training two networks simultaneously using different data sets of the targeted system.⁴⁹ Figure 4 shows a schematic diagram of the co-teaching method with two networks: A and B. The intuition of co-teaching is straightforward and it is based on the observations that neural networks tend to fit simple pattern at the early stage of training process.⁴⁹ As a result, noise-free data will achieve a low loss function value, while noisy data typically has a high loss function value. Specifically, the co-teaching training method (Algorithm 1) works as follows: in each mini-batch during training epoch k , each model checks its data sequences (i.e., each pair of data labeled as input and output), and generates a small data set with all the data that has a low loss function value. This new data set can be approximately regarded as noise-free data sets, and will be shared between two networks. Subsequently, after receiving the new noise-free data sets from the peer network, the weights are updated and the training is resumed for one more epoch. The above process is repeated until the all the training epochs are completed.

One of the main benefits of co-teaching method is that each network can filter noisy labels in different ways due to distinct learning abilities.⁴⁹ In order for the algorithm to perform accurately, it is important to have a considerable number of clean sequences during the training step such that clean sequences can be detected and shared within the two networks. However, to our knowledge, little attention has been paid to the implementation of co-teaching method in solving regression problems. In this study, we take advantage of the idea of co-teaching method, and adapt it to LSTM modeling of nonlinear processes using noisy data. Specifically, we first develop a noise-free data

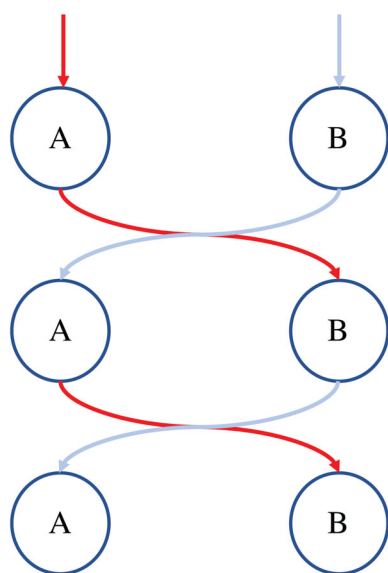


FIGURE 4 The symmetric co-teaching framework that trains two networks (A and B) simultaneously [Color figure can be viewed at wileyonlinelibrary.com]

Algorithm 1

Co-teaching Algorithm

D is the original mixed data set, l_{max} is the maximum number of iterations, x is the data sequence, $loss(A, x)$ calculates the loss function value for data x under model A, $loss_T$ is the threshold for identifying small-loss data sequences, and η is the learning rate.

for $i = 0$ **to** l_{max} **do**.

Select a mini-batch D_m from D .

Obtain the small-loss data sequences from model A:

$$D_A = \{x \in D_m \mid loss(A, x) \leq loss_T\}.$$

Obtain the small-loss data sequences from model B:

$$D_B = \{x \in D_m \mid loss(B, x) \leq loss_T\}.$$

Update the weight matrix of model A:

$$\mathbf{W}_A = \mathbf{W}_A - \eta \nabla loss(A, D_B).$$

Update the weight matrix of model B:

$$\mathbf{W}_B = \mathbf{W}_B - \eta \nabla loss(B, D_A).$$

end

set from extensive first-principles model simulations. Although an accurate process model is generally unavailable for industrial chemical processes, a first-principles model based on well-known mass, energy, and momentum balances can be developed to approximate process dynamics. This noise-free data set can then be used in co-teaching training to guide LSTM networks to learn the underlying (noise-free) dynamics. When using co-teaching method for solving regression problems, it is important to tune the LSTM structure, develop a high-quality data set consisting of both noise-free and noisy data, and choose the threshold that can identify the clean data in the mixed data set. Specifically, the number of units in each network should be carefully chosen in order to achieve a balanced performance between the noisy and clean pattern. For example, a large number of hidden units are not preferred since the network tends to fit both clean and noisy data at the same time (i.e., overfitting). Second, the ratio between clean data obtained from the approximated model and noisy data should also be carefully chosen. If clean data are insufficient, the networks are not able to share the noise-free process dynamics information during training phase; however, if too much clean data are included in the mixed data set, then the network would simply learn the process dynamics from approximated first-principles model instead of the actual process dynamics from noisy data. Additionally, given a mixed data set, it is also important to identify the clean data by evaluating loss function value as introduced at the beginning of the section. To better understand the role of threshold in choosing clean data, we perform the following case study using the symmetry co-teaching method with model *A* and *B* that share clean data during the training step. Model *B* uses a data set of 500 sequences in which the first 100 data are noisy and the following 400 data are clean sequences. Subsequently, at the end of each epoch, based on the threshold we pre-determined, we select the clean sequences by evaluating their contribution to the loss function. Table 2 shows the selected clean sequences for four thresholds $loss_T$ during the same epoch. It is demonstrated in Table 2 that under the threshold of 0.01, 306 sequences are selected to be shared with model *A* in which 267 are truly clean sequences (i.e., the accuracy is approximately 87%). By decreasing the threshold value, it is shown in Table 2 that fewer data sequences are selected as clean data; however, the accuracy is increasing (e.g., the ratio of true clean data are increasing to 95.6% for the threshold 0.0014). Therefore, the threshold for identifying clean data also plays an important role in the co-teaching training performance.

Another type of co-teaching framework uses an asymmetric structure (termed asymmetric co-teaching). Unlike the symmetric co-teaching that trains two models using the same data set, asymmetric co-teaching methodology trains two models (i.e., *A* and *B*) using a noise-free and a noisy data set, respectively. In this work, the noise-free data set for model *A* comes from extensive open-loop simulation for the CSTR model of Equation (7) under 2000 initial conditions that cover the entire stability region and 100 pairs of manipulated inputs. Based on the open-loop simulation data set, the noisy data set is generated by adding the industrial noise of Figure 2 on state measurements. As a result, the two data sets have the same number of data

sequences and the data are also organized in the same order. In other words, the k_{th} data sequence in the two data sets are using the same initial condition and the same control actions with the only difference that the industrial noise is added on state measurements. In asymmetric co-teaching framework, clean data set is extracted and shared only from model *A* to model *B*, and not in the opposite direction. Similar to Algorithm 1, in each training epoch, model *A* sends a subset of clean data sequences to model *B*. To update the noisy labels during the co-teaching implementation in the asymmetric co-teaching method, the noisy labels are randomly updated using its noise-free counterpart from the clean data set. It will be shown in the next section that LSTM training using asymmetric co-teaching method can achieve better model accuracy than the LSTM using standard training algorithm with the same mixed data set of clean and noisy data. To implement the co-teaching method in Keras, the training process of the neural network is discretized by using a loop structure in order to update the noisy LSTM outputs between two consecutive epochs. After updating the outputs, the training is resumed and the noisy outputs will be updated again at the end of next training epoch.

Remark 5. Bias is intentionally introduced into the LSTM models in order to improve their performance in predicting underlying (nominal) process dynamics. Note that as we assume no information on industrial noise is available, the proposed LSTM models do not account for any error term in its formulation—the LSTM inputs are process states and control actions only. Also, since the training data are noisy for both LSTM inputs and outputs, the objective of dropout and co-teaching LSTM modeling approaches is to avoid overfitting (overfitting typically exhibits low bias and high variance, where low bias means that the model fits well on the training set, and high variance means that the model is unable to make accurate predictions on the validation set). In fact, no bias is undesired in this case because it means that the LSTM model fits the noisy pattern (i.e., noisy outputs) very well, and cannot predict the true (nominal) state.

4.1 | Open-loop simulation results of three LSTM models

We carry out the open-loop simulations using standard LSTM, dropout LSTM, and co-teaching LSTM models trained with the same noisy data set, and show the MSE results in Table 3. It should be noted that all the LSTM models in Table 3 are developed using the same structure in terms of neurons, layers, epochs, and activation functions. Additionally, the MSE is calculated as the difference between the predicted state trajectories and the underlying (noise-free) state trajectories since the goal is to capture the nominal process dynamics using noisy data. Therefore, an LSTM model with a low MSE value implies that the model is able to predict smooth state trajectories that are close to the ground-truth trajectories.

As shown in Table 3, we first train three standard LSTM models using the noise-free data set (baseline case), the mixed data set, and the noisy data set, respectively, to demonstrate that the standard LSTM modeling approach cannot achieve a desired model accuracy without a high-quality data set. However, compared with the standard LSTM using noisy data only (i.e., 1c in Table 3), it is observed that the LSTM using mixed data (i.e., 1b in Table 3) shows a slight improvement of model prediction with the use of noise-free data in training. The co-teaching LSTM and dropout LSTM models are then developed using the same noisy data set. Specifically, the co-teaching LSTM starts with a noise data set, and receives clean data sequences from the peer model as the training process evolves. As a result, the weight matrices of the co-teaching model are able to capture a balanced pattern that accounts for both noisy and clean data. The dropout LSTM is trained using the noisy data set only and the final prediction results are obtained by averaging all Monte Carlo realizations that apply dropout at test time. It is shown in Table 3 that the dropout LSTM outperforms the standard LSTM model in terms of a lower MSE, and the co-teaching LSTM achieves the best performance among all the models.

4.2 | Closed-loop simulation results under LMPC

We run closed-loop simulations on the CSTR system of Equation (7) under the LMPC of Equation (6) using different LSTM models (i.e., standard LSTM, co-teaching LSTM, and dropout LSTM). The LMPC receives noisy measurements at each integration time step, and solves for the optimal control actions at every sampling time. The control objective is to operate the CSTR at the unstable equilibrium point (C_{As}, T_s) by manipulating the heat input rate ΔQ and the inlet concentration ΔC_{A0} under LMPC. The LMPC objective function is designed with the following form:

$$L(x, u) = |x|_{Q_1}^2 + |u|_{Q_2}^2 \quad (12)$$

where Q_1 and Q_2 are coefficient matrices that balance the contributions of state convergence and of control actions (i.e., energy and reactant consumption). Table 4 summarizes the closed-loop performance of different LSTM models by using the objective function of Equation (12) as an indicator. Specifically, industrial noise of four different levels (i.e., tiny, small, medium, and large levels) are added on closed-loop state measurements for each LSTM model. The four noise levels correspond to the multiplication of the normalized industrial

noise (Figure 2) with the coefficients $\sigma_i^{C_A}$ and σ_i^T , $i = 1, 2, 3, 4$ that have been reported in Section 3.1. Each level of noise is tested in closed-loop simulation using five different initial conditions in the stability region. Then, we use the LMPC objective function value to analyze its closed-loop performance in the way that a large objective function value implies a slow convergence and a high-energy consumption. We calculate the objective function value over time for each closed-loop state trajectory and compute the averaged result, that is, $\frac{1}{N_s} \sum_{i=1}^{N_s} \int_{t=0}^{t=t_s} L(x, u) dt$ for each LSTM model, where N_s is the number of simulation runs ($N_s = 5$ corresponding to five different initial conditions in this study), and t_s is the closed-loop simulation time. The mean objective function value is finally normalized with respect to its biggest value in all simulation runs to eliminate the impact of initial conditions on closed-loop performance analysis.

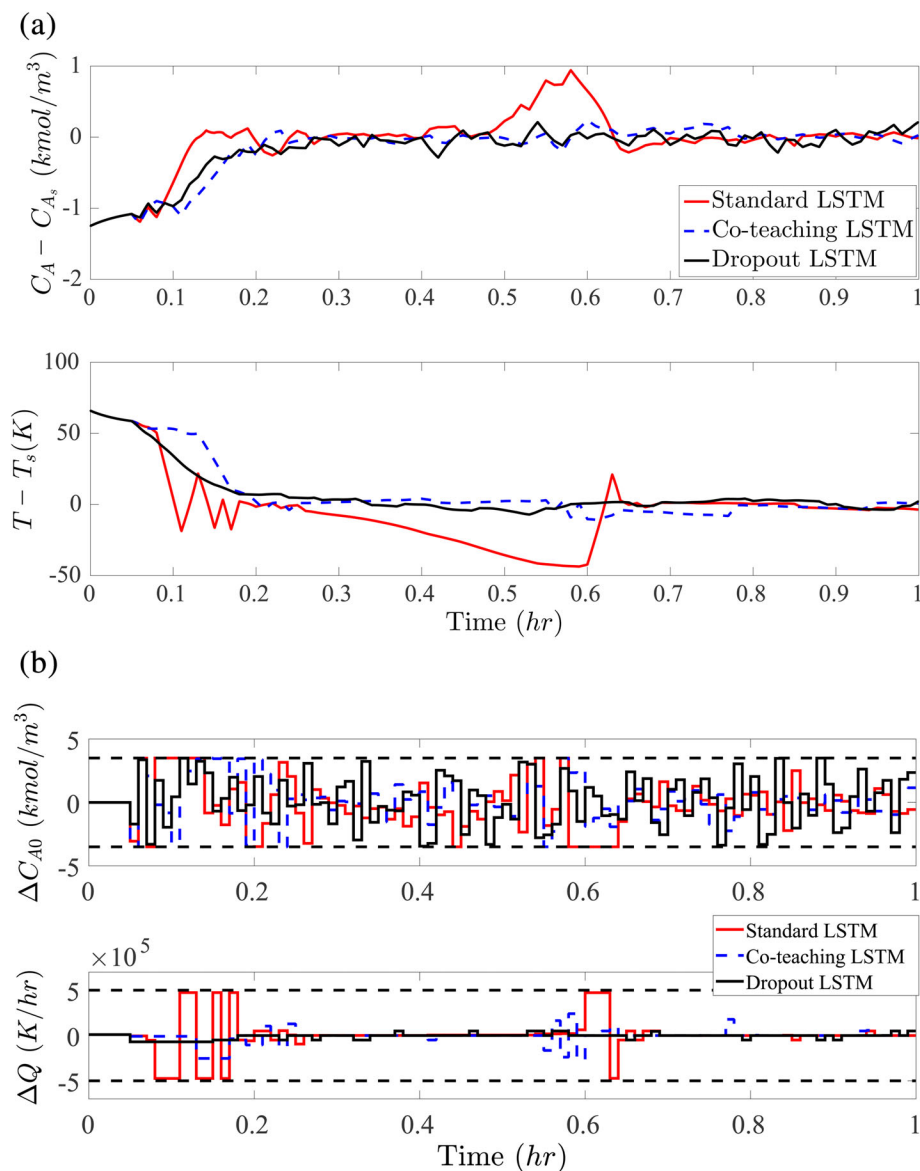
As shown in Table 4, the standard LSTM has the worst closed-loop performance in terms of the largest objective function value among all three models. Additionally, co-teaching and dropout LSTMs achieve a significant improvement (around 20%) for the small and medium noise levels, while the improvement is less than 10% for tiny and large noise levels. Specifically, it is observed from closed-loop simulations that under a tiny noise level, the LMPCs using the three LSTM models all drive the closed-loop state to the steady-state quickly. This is consistent with the fact that the machine-learning-based LMPC is robust to a sufficiently small disturbance as shown in Reference 33. Additionally, all the LSTM models perform poorly under a large noise level since the noisy state measurements deviate far from its true state in the presence of a large noise, which makes it challenging for LSTM to predict the true states. As a result, LMPC is unable to solve for the optimal control actions that can drive the closed-loop state toward the steady-state. In fact, all the closed-loop state trajectories show significant oscillations under a large noise level. Practically, such a large noise implies that the problem is beyond the scope of sensor noise, and equipment maintenance may be needed before bringing controller system on-line. In the case of a small and a medium noise level, it is shown in Table 4 that closed-loop performances under the co-teaching and dropout LSTMs are much improved compared with the standard LSTM. This implies that the proposed LSTM modeling approaches are preferred in handling noise in such a range of noise levels. From Table 4, we can conclude that the improvements of model accuracy under dropout/co-teaching methods vary depending on the noise levels. Therefore, just like any other machine learning models, whether the dropout and co-teaching LSTM models achieve desired performance and how much improvement they have can only be assessed after the training process (i.e., open-loop performance) and also in the closed-loop operation (i.e., closed-loop performance).

In Figure 5a,b, we show one of the closed-loop simulation results under a medium noise level. It is observed in state profiles of Figure 5a that the standard LSTM model shows significant variation when the closed-loop state approaches the steady-state. This is due to the fact that states predicted by the standard LSTM are far from the true states, which mislead the LMPC to give a solution that drives the state in a wrong direction. However, since the co-teaching LSTM and dropout LSTM models have been demonstrated

TABLE 4 Statistical analysis of the closed-loop simulation results under non-Gaussian noise using four noise levels

Methods	Tiny	Small	Medium	Large
(1) Standard LSTM	0.8786	0.9796	0.9430	0.8768
(2) Co-teaching LSTM	0.8110	0.7587	0.7749	0.7850
(3) Dropout LSTM	0.7162	0.6889	0.8078	0.7928

FIGURE 5 (a) Closed-loop state profiles and (b) Manipulated input profiles ($u_1 = \Delta C_{A0}$, $u_2 = \Delta Q$) for the initial condition $(-1.25, 66)$ under LMPC using standard LSTM (red), co-teaching LSTM (blue), and dropout LSTM (black) [Color figure can be viewed at wileyonlinelibrary.com]



to achieve a desired model accuracy in Section 4.1, it is expected that the LMPC using these two models can maintain the state in a small neighborhood around the steady-state more smoothly. Note that the MPC will not be able to stabilize the system exactly at the steady-state due to the sample-and-hold implementation of control actions, and the model mismatch between the LSTM model and the actual nonlinear process. Therefore, if the state trajectory of the closed-loop system starting from the stability region Ω_p remains bounded in Ω_p and converges to a small compact set around the origin where it will be maintained thereafter, then the system is considered practically stable under the sample-and-hold implementation of MPC. The control action ΔC_{A0} in Figure 5b shows variation over the simulation period because the closed-loop states in Figure 5a move dynamically within $\Omega_{p_{nn}}$ under MPC. Since both the reactant concentration C_A and reactor temperature T remain close to the steady-state value after $t = 0.25$ hr in Figure 5a, the system is practically stable for the initial condition $(-1.25, 66)$ under LSTM-based MPC.

Remark 6. The closed-loop control performance is influenced by the model accuracy after the training process. Since the dropout/co-teaching LSTM modeling approaches are purely data-driven, no stability guarantees can be derived priori to the training. In Reference 34, we have demonstrated that closed-loop stability is ensured for the machine-learning-based MPC provided that the machine learning model achieves a desired model accuracy in a given operating region. Therefore, if the dropout and co-teaching LSTM models are developed with a desired model accuracy, the stability analysis also holds for the MPC using these LSTM models.

5 | CONCLUSION

In this work, we presented machine learning approaches that can predict underlying nonlinear process dynamics from noisy data. We

initially investigated the case of Gaussian noise and demonstrated that the standard LSTM using noisy data set was able to achieve a desired model accuracy since the internal states of the neural network played a dominant role in prediction. Then, we investigated the case of a non-Gaussian noise from ASPEN industrial-data sets, for which the standard LSTM networks showed poor prediction performance. To handle industrial noisy data, dropout LSTM and co-teaching LSTM schemes were proposed to learn process dynamics using noisy data only, and using both noisy data and noise-free first-principles model simulation data, respectively. The chemical process example was utilized to demonstrate the improved model accuracy achieved by the dropout and co-teaching LSTM models through both open-loop and closed-loop simulations.

ACKNOWLEDGMENT

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

AUTHOR CONTRIBUTIONS

Panagiotis Christofides: Methodology; supervision; writing-review and editing. **Zhe Wu:** Conceptualization; data curation; formal analysis; writing-original draft. **Junwei Luo:** Data curation; formal analysis; investigation. **David Rincon:** Conceptualization; data curation; formal analysis; writing-original draft.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Panagiotis D. Christofides  <https://orcid.org/0000-0002-8772-4348>

REFERENCES

1. Aggelogiannaki E, Sarimveis H. Nonlinear model predictive control for distributed parameter systems using data driven artificial neural network models. *Comput Chem Eng*. 2008;32:1225-1237.
2. Al Seyab R, Cao Y. Nonlinear system identification for predictive control using continuous time recurrent neural networks and automatic differentiation. *J Process Control*. 2008;18:568-581.
3. Aumi S, Corbett B, Clarke-Pringle T, Mhaskar P. Data-driven model predictive quality control of batch processes. *AIChE J*. 2013;59:2852-2861.
4. Chaffart D, Ricardez-Sandoval LA. Optimization and control of a thin film growth process: a hybrid first principles/artificial neural network based multiscale modelling approach. *Comput Chem Eng*. 2018;119:465-479.
5. Garg A, Mhaskar P. Utilizing big data for batch process modeling and control. *Comput Chem Eng*. 2018;119:228-236.
6. Xie W, Bonis I, Theodoropoulos C. Data-driven model reduction-based nonlinear mpc for large-scale distributed parameter systems. *J Process Control*. 2015;35:50-58.
7. Zeng J, Gao C, Su H. Data-driven predictive control for blast furnace ironmaking process. *Comput Chem Eng*. 2010;34:1854-1862.
8. C. Moore. Application of singular value decomposition to the design, analysis, and control of industrial processes. Paper presented at the 1986 American Control Conference, 643-650. IEEE, 1986.
9. Van Overschee P, De Moor B. N4SID: subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica*. 1994;30:75-93.
10. Huusom JK, Poulsen NK, Jørgensen SB, Jørgensen JB. Tuning siso offset-free model predictive control based on arx models. *J Process Control*. 2012;22:1997-2007.
11. Menezes JMP Jr, Barreto GA. Long-term time series prediction with the narx network: an empirical evaluation. *Neurocomputing*. 2008;71:3335-3343.
12. Siegelmann HT, Horne BG, Giles CL. Computational capabilities of recurrent narx neural networks. *IEEE Trans Syst Man Cybern B Cybern*. 1997;27:208-215.
13. Prasad V, Bequette BW. Nonlinear system identification and model reduction using artificial neural networks. *Comput Chem Eng*. 2003;27:1741-1754.
14. Delgado A, Kambhampati C, Warwick K. Dynamic recurrent neural network for system identification and control. *IEE Proceed-Cont Theory Appl*. 1995;142:307-314.
15. Hong X, Mitchell RJ, Chen S, Harris CJ, Li K, Irwin GW. Model selection approaches for non-linear system identification: a review. *Int J Syst Sci*. 2008;39:925-946.
16. J. Sjöberg, Q. Zhang, L. Ljung, A. Benveniste, B. Deylon, P. Glorennec, H. Hjalmarsson, and A. Juditsky. Nonlinear black-box modeling in system identification: a unified overview. Linköping University, 1995.
17. Zhu Y, Butoyi F. Case studies on closed-loop identification for mpc. *Control Eng Pract*. 2002;10:403-417.
18. Zhu Y, Patwardhan R, Wagner SB, Zhao J. Toward a low cost and high performance mpc: the role of system identification. *Comput Chem Eng*. 2013;51:124-135.
19. Ali JM, Hussain MA, Tade MO, Zhang J. Artificial intelligence techniques applied as estimator in chemical process systems—a literature survey. *Expert Syst Appl*. 2015;42:5915-5931.
20. Kosmatopoulos EB, Polycarpou MM, Christodoulou MA, Ioannou PA. High-order neural network structures for identification of dynamical systems. *IEEE Trans Neural Netw*. 1995;6:422-431.
21. Trischler AP, D'Eleuterio GM. Synthesis of recurrent neural networks for dynamical system simulation. *Neural Netw*. 2016;80:67-78.
22. Wong W, Chee E, Li J, Wang X. Recurrent neural network-based model predictive control for continuous pharmaceutical manufacturing. *Mathematics*. 2018;6(11):242.
23. Diversi R, Guidorzi R, Soverini U. Identification of ARX and ARARX models in the presence of input and output noises. *Eur J Control*. 2010;16:242-255.
24. Wu P, Pan H, Ren J, Yang C. A new subspace identification approach based on principal component analysis and noise estimation. *Ind Eng Chem Res*. 2015;54:5106-5114.
25. Juricek BC, Larimore WE, Seborg DE. Reduced-rank ARX and subspace system identification for process control. *IFAC Proceed*. 1998;31:247-252.
26. Patwardhan SC, Narasimhan S, Jagadeesan P, Gopaluni B, Shah SL. Nonlinear bayesian state estimation: a review of recent developments. *Control Eng Pract*. 2012;20:933-953.
27. Yeo K, Melnyk I. Deep learning algorithm for data-driven simulation of noisy dynamical system. *J Comput Phys*. 2019;376:1212-1231.
28. Hamilton F, Berry T, Sauer T. Ensemble Kalman filtering without a model. *Phys Rev X*. 2016;6:011021.
29. Savitzky A, Golay MJE. Smoothing and differentiation of data by simplified least squares procedures. *Anal Chem*. 1964;36:1627-1639.
30. Bishop CM. *Pattern recognition and machine learning*. New York: Springer-Verlag; 2006.
31. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Stat Comput*. 2004;14:199-222.
32. Wang H, Chaffart D, Ricardez-Sandoval LA. Modelling and optimization of a pilot-scale entrained-flow gasifier using artificial neural networks. *Energy*. 2019;188:116076.

33. Wu Z, Tran A, Rincon D, Christofides PD. Machine learning-based predictive control of nonlinear processes. Part I: theory. *AIChE J.* 2019;65:e16729.
34. Wu Z, Tran A, Rincon D, Christofides PD. Machine-learning-based predictive control of nonlinear processes. Part II: computational implementation. *AIChE J.* 2019;65:e16734.
35. Hassanpour H, Corbett B, Mhaskar P. Integrating dynamic neural network models with principal component analysis for adaptive model predictive control. *Chem Eng Res Des.* 2020;161:26-37.
36. Chen S, Wu Z, Rincon D, Christofides PD. Machine learning-based distributed model predictive control of nonlinear processes. *AIChE J.* 2020;66:e17013.
37. Ellis MJ, Chinde V. An encoder-decoder LSTM-based EMPC framework applied to a building HVAC system. *Chem Eng Res Des.* 2020; 160:508-527.
38. Shah D, Wang J, He QP. Feature engineering in big data analytics for IoT-enabled smart manufacturing—comparison between deep learning and statistical learning. *Comput Chem Eng.* 2020;141:106970.
39. He QP, Wang J. Statistical process monitoring as a big data analytics tool for smart manufacturing. *J Process Control.* 2018;67:35-43.
40. He QP, Wang J, Shah D. Feature space monitoring for smart manufacturing via statistics pattern analysis. *Comput Chem Eng.* 2019; 126:321-331.
41. Lee JH, Shin J, Realf MJ. Machine learning: overview of the recent progresses and implications for the process systems engineering field. *Comput Chem Eng.* 2018;114:111-121.
42. Qin SJ, Chiang LH. Advances and opportunities in machine learning for process data analytics. *Comput Chem Eng.* 2019;126:465-473.
43. Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: Lstm cells and network architectures. *Neural Comput.* 2019;31:1235-1270.
44. Udugama IA, Gargalo CL, Yamashita Y, et al. The role of big data in industrial (bio) chemical process operations. *Ind Eng Chem Res.* 2020; 59:15283-15297.
45. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.
46. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res.* 2014;15:1929-1958.
47. Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: representing model uncertainty in deep learning. Paper presented at the *Proceedings of the International Conference on Machine Learning* 1050-1059, 2016.
48. Y. Gal and Z. Ghahramani. A theoretically grounded application of dropout in recurrent neural networks. Paper presented at the *Proceedings of the Advances in Neural Information Processing Systems*, 1019-1027, 2016.
49. B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama. Co-teaching: robust training of deep neural networks with extremely noisy labels. Paper presented at the *Proceedings of the Advances in Neural Information Processing Systems*, pages 8527-8537, 2018.
50. Khalil HK. *Nonlinear systems*. Vol 3. Upper Saddle River, NJ: Prentice Hall; 2002.
51. Lin Y, Sontag ED. A universal formula for stabilization with bounded controls. *Syst Control Lett.* 1991;16:393-397.
52. Funahashi K, Nakamura Y. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Netw.* 1993;6: 801-806.
53. Wächter A, Biegler LT. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math Program.* 2006;106:25-57.
54. Wu Z, Christofides PD. Economic machine-learning-based predictive control of nonlinear systems. *Mathematics.* 2019;7:494.
55. K. Yeo. Short note on the behavior of recurrent neural network for noisy dynamical system. *arXiv preprint arXiv:1904.05158*, 2019.
56. F. Yang, K. Li, Z. Zhong, Z. Luo, X. Sun, H. Cheng, X. Guo, F. Huang, R. Ji, and S. Li. Asymmetric co-teaching for unsupervised cross-domain person re-identification., 2020.

How to cite this article: Wu Z, Rincon D, Luo J, Christofides PD. Machine learning modeling and predictive control of nonlinear processes using noisy data. *AIChE J.* 2021; 67:e17164. <https://doi.org/10.1002/aic.17164>