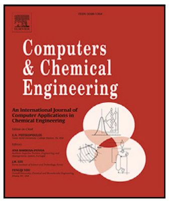
Computers and Chemical Engineering 178 (2023) 108396

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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



Simultaneous multistep transformer architecture for model predictive control

Junho Park [a], Mohammad Reza Babaei [a], Samuel Arce Munoz [a], Ashwin N. Venkat [b], John D. Hedengren [a,*]

[a] *Department of Chemical Engineering, Brigham Young University, Provo, UT, USA*
[b] *Seeq Corporation, Seattle, WA, USA*

ARTICLE INFO

Dataset link: https://github.com/BYU-PRISM/Transformer_MPC

Keywords:
Transformer neural network architecture
Long short-term memory
Model predictive control

ABSTRACT

Transformer neural networks have revolutionized natural language processing by effectively addressing the vanishing gradient problem. This study focuses on applying Transformer models to time-series forecasting and customizing them for a simultaneous multistep-ahead prediction model in surrogate model predictive control (MPC). The proposed method showcases improved control performance and computational efficiency compared to LSTM-based MPC and one-step-ahead prediction models using both LSTM and Transformer networks. The study introduces three key contributions: (1) a new MPC system based on a Transformer time-series architecture, (2) a training method enabling multistep-ahead prediction for time-series machine learning models, and (3) validation of the enhanced time performance of multistep-ahead Transformer MPC compared to one-step-ahead LSTM networks. Case studies demonstrate a significant fifteen-fold improvement in computational speed compared to one-step-ahead LSTM, although this improvement may vary depending on MPC factors like the lookback window and prediction horizon.

1. Introduction

Machine learning technologies have gained attention in many industries and have grown in popularity, especially in areas requiring classification, such as pattern recognition for image and voice. Many researchers in process system engineering have recently renewed interest in machine learning technologies, such as deep learning. An overview of current trends and opportunities for machine learning technologies in process system engineering has been discussed in Lee et al. (2018), Qin and Chiang (2019), Venkatasubramanian (2019) and Gopaluni et al. (2020). Areas using machine learning technologies in process system engineering are control and optimization. Model predictive control (MPC) and real-time optimization (RTO) are two advanced automation systems that require process models. Many MPC application research studies have been conducted using artificial neural network (ANN) models: internal combustion engine (Wang et al., 2006), distillation column and continuous stirred tank reactor (CSTR) example (Lee and Park, 1992; Al Seyab and Cao, 2008), Heating ventilation and air conditioning systems (HVAC) (Wang et al., 2017; Drgoňa et al., 2021), LSTM (Long short term memory network) MPC for continuous pharmaceutical reactors (Wong et al., 2018), and phthalic anhydride synthesis in a fixed-bed catalytic reactor (Wu et al., 2019). A Hybrid model with a Hammerstein model structure, which consists of a static neural net model block and a linear dynamics model block in series, has been investigated in Park et al. (2020a). A gated recurrent unit (GRU)-based encoder–decoder architecture integrated with a physics-based model has been tested for thermal power plants (Machalek et al., 2022). An approximated MPC concept has also been investigated by learning control policy from the closed-loop performance data (Mesbah et al., 2022). Reinforcement learning methods for RTO application have been investigated in Powell et al. (2020), Machalek et al. (2021) and Yoo et al. (2021).

Among the different types of ANNs, RNNs are a natural choice for modeling time-series data because the structure of RNN states has explicit temporal dependence. However, the sequential processing nature of the RNNs makes the model prediction computationally expensive by recursively updating the hidden state. A novel time-series ANN architecture, Transformer neural network, has been introduced to solve the issue by parallelizing the model prediction steps, making it possible to produce multistep predictions simultaneously (Vaswani et al., 2017). State-of-the-art ANN architectures, especially attention mechanisms, improve natural language processing (NLP). Although dynamic process models can be built with time-series data like the NLP problems, the actual issues are vastly different in model usage for prediction in model-based control and optimization applications.

* Corresponding author.
E-mail address: john.hedengren@byu.edu (J.D. Hedengren).

https://doi.org/10.1016/j.compchemeng.2023.108396
Received 3 March 2023; Received in revised form 15 August 2023; Accepted 21 August 2023
Available online 22 August 2023

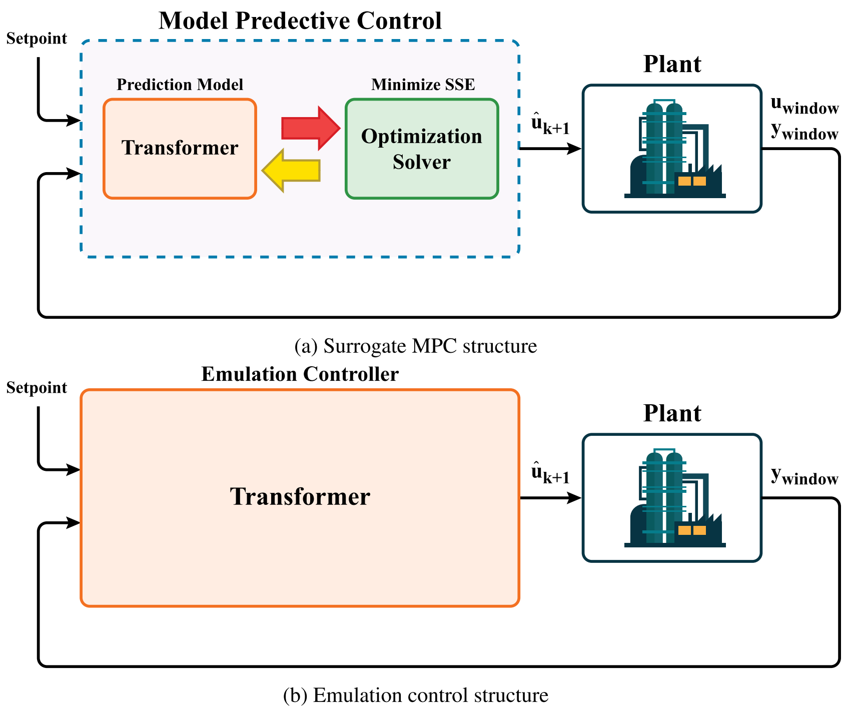


Fig. 1. Two scenarios for Transformers in model based control.

Previous research has recognized the Transformer network for its remarkable computational speed. A distinguishing feature of this network is its attention mechanism, which adeptly handles irregular temporal dependencies. This capability is particularly notable in the context of natural language data. This study further explores the potential for the application of the Transformer network within process control fields by integrating it into the MPC framework as a predictive model. In addition to creating a Transformer-based MPC system, the research introduces a multistep-ahead prediction structure to maximize the benefits derived from the Transformer advancements, particularly within the context of MPC applications. Evidence from three case studies indicates that the multistep ahead Transformer MPC offers superior performance in terms of computational efficiency and setpoint tracking compared to the one-step ahead LSTM MPC.

In an effort to contextualize these findings, the study contrasts the proposed multistep ahead Transformer MPC with the one-step ahead LSTM, which maintains the traditional data-driven time-series format within the surrogate MPC framework. Although all the case studies do not exhaustively examine two alternative model structures – multistep ahead LSTM and one-step ahead Transformer – the first case study (5.1) provides an indirect comparison of these models to the multistep ahead Transformer MPC in terms of prediction accuracy and computational efficiency. Further, sequential neural network models such as LSTM are recognized for a relative lack of prediction precision, particularly when dealing with irregular temporal dependencies prevalent in the multistep-ahead data structure as shown in Fig. 7.

While this paper does not delve into a comprehensive investigation of all model structures, it lays the groundwork for future research. Further studies could provide valuable insights by examining the performance of various models in relation to complexity and ability to handle data irregularities. This could pave the way for a more understanding of model behaviors in different scenarios.

The objective of this work is to develop a simultaneous multistep-ahead prediction method using a Transformer architecture for MPC applications. Fig. 1 gives an overview of two approaches for using Transformer models for model based-control. The Transformer model can emulate the process model (Fig. 1(a)) or the entire model predictive control (Fig. 1(b)). For the scope of this paper, the exploration primarily focuses on the former, investigating the surrogate MPC approach more intensively. A brief review of Transformer Network Architecture is discussed with details of the adaptations for time-series data. Source code is available from https://github.com/BYU-PRISM/Transformer_MPC

2. Transformer network architecture

The original concept of Transformer architecture in “Attention is all you need” (Vaswani et al., 2017) is dedicated to machine translation tasks (Fig. 2). Thus, each element of the original design needs to be customized for the time-series data processing. This section discusses each part of the original Transformer design and the modifications for time-series data, especially for the simultaneous multistep-ahead prediction needed for MPC.

2.1. Encoder–decoder

The Encoder–Decoder concept is initially used in the Sequence to Sequence (Seq2seq) architecture that places two separate RNNs in a row. The Encoder RNN processes the input sequence to produce a context vector. This vector, which encapsulates the information from the input sequence, is then passed to the Decoder RNN to predict the output sequence. The Transformer architecture retains the encoder–decoder framework while using new concepts for specific components.

2.2. Positional encoding

In the RNN models, the hidden state of the previous timesteps returns to the network as an additional input along with the actual input values for the next time step. This recursive data processing is not computationally efficient, but it has an advantage in that it can inherently recognize the order of the input sequence. However, as the Transformer model takes the whole input sequence simultaneously, the inherent temporal dependency in the RNN is no longer available in the Transformer. A positional embedding layer explicitly labels the sequence position to have the Transformer network recognize the sequence order. Trigonometric functions (Vaswani et al., 2017) is a

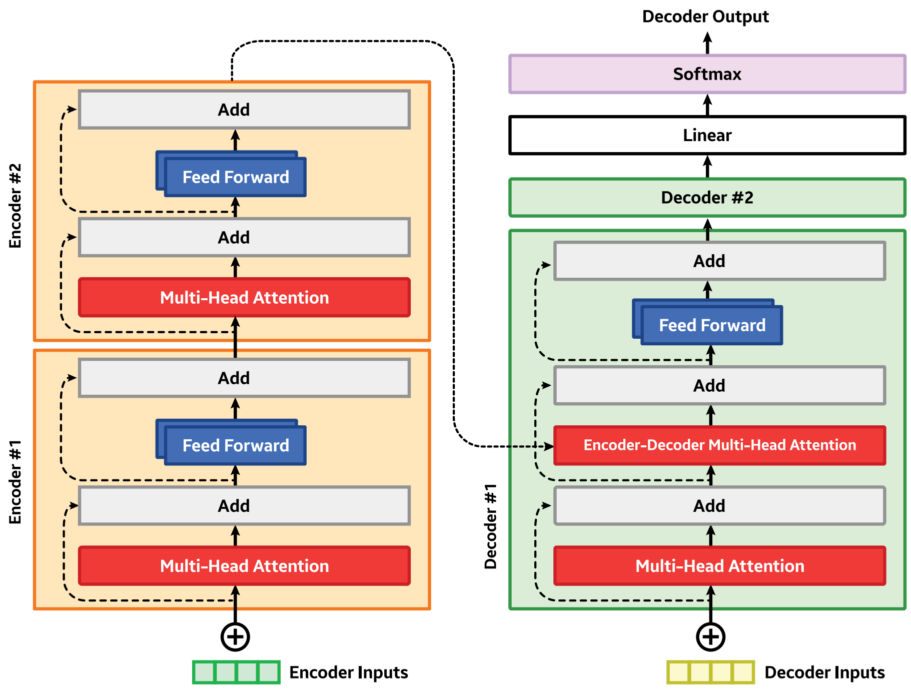


Fig. 2. Transformer structure.

method among many to embed the positional information into the input sequence. This method is particularly effective for sequences that have an irregular length and interval by assigning the unique index in the form of a real number vector. Thus, this method is commonly used for NLP tasks that vary the number of words in every sentence, resulting in different input vector lengths and time intervals. On the other hand, the time-series data usually has a fixed size and interval or can easily be converted to this form by signal processing.

2.3. Attention mechanism

The attention mechanism is a crucial part of the Transformer model. It has been proposed to improve the long-term memory capability in the Seq2seq model (Bahdanau et al., 2015). The encoded input sequence in the Seq2seq model is saved in a fixed-length vector called context vector and passed to the decoder RNN (Cho et al., 2014). Because the context vector in the Seq2seq model is just the last hidden state of the encoder RNN, it cannot accommodate the most critical correlation at earlier time steps caused by vanishing gradients. The memory issue of the RNN can be significant when the time series is long. The attention mechanism improves the memory of the context vector by introducing a probability distribution between a hidden state of the decoder RNN at a specific time step and the hidden states of every time step in the encoder RNN. This probability distribution helps the model focus more on a certain time step in the input sequence regardless of the order. It is converted further into a vector with the same size as the hidden state, called the attention value. This attention value acts as a context vector of the Seq2seq model, holding more useful long-term correlation information between the encoder and decoder. The Transformer architecture extensively employs the attention mechanism, which removes the inherent sequential processing in the RNN models. A conditional probability of a sequence data for the multistep-ahead prediction is shown in Eq. (1).

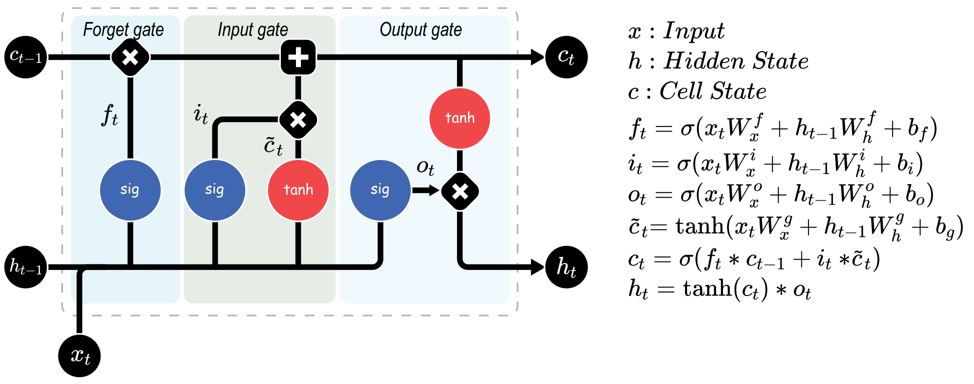
$$\mathbf{p}(\mathbf{y}_{k+1\,:\,k+P}\,|\mathbf{y}_{k-w\,:\,k},\ \mathbf{u}_{k-w\,:\,k+P}) = \prod_{t=k+1}^{k+P} \mathbf{p}(\mathbf{y}_t \mid \mathbf{y}_{t-w-1\,:\,t-1},\ \mathbf{u}_{t-w-1\,:\,t-1}), \tag{1}$$

where, $\mathbf{p}(A \mid B)$ represents the probability of A under the given condition B. $\mathbf{y}$ and $\mathbf{u}$ denote the system output and system input, respectively. The past system output $\mathbf{y}$ is used as one of the NN model inputs, called auto-regressive input, while the system input $\mathbf{u}$ is called exogenous input. The future values $(t > k)$ of $\mathbf{u}$ are also included in the exogenous input as known values for the NN model training, while the $\mathbf{y}$ only include the past values $(t < k)$ as shown in LHS of Eq. (1). k, w, and P denote the current time step, look-back window size, and prediction horizon size, respectively. Using the RNN models to achieve the multistep prediction, it is inevitable to build up the hidden state recursively from time step $k - w$ to k and repeat the same recursive computations for every prediction value in the prediction horizon P. However, the attention mechanism simultaneously computes the probability distribution, also called 'Attention distribution' between the input and output sequence, altering the recurrence in RNN with the matrix multiplications and $softmax$ activation function.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2}$$

where, Q, K, and V represent the weighted sequence data matrices, $Query$, Key, and $Value$. The same input sequence data, X, is assigned for all Q, K, and V matrices for the encoder self-attention layers as $Q = XW_Q$, $K = XW_K$, and $V = XW_V$, where, $X = [\mathbf{y}_{k-w\,:\,k+P};\ \mathbf{u}_{k-w\,:\,k+P}] \in \mathbb{R}^{b\times w+P\times l}$. The b and l denote the batch size and the number of variables in the input matrix, respectively. Note that the future values in sequence $\mathbf{y}$ in the X matrix $(\mathbf{y}_{k+1\,:\,k+P})$ need to be masked with the $\mathbf{y}$ value at the time step k $(\mathbf{y}_k)$ to avoid including the prediction output in the input matrix (X) before training.

The prediction model, to be compatible with the MPC framework, requires both future marked $\mathbf{y}_k$ data points and $\mathbf{u}_{k+1:k+P}$ data points within the X matrix. The $\mathbf{u}_{k+1:k+P}$ data points, in particular, are vital to ensure the degrees of freedom for the MPC optimization solution. Conversely, while the MPC framework does not inherently require future $\mathbf{y}$ points, they are essential for maintaining consistent length between the $\mathbf{y}$ and $\mathbf{u}$ vectors within the X matrix. Including spurious $\mathbf{y}$ values in the prediction horizon portion of the $\mathbf{y}$ vector can lead to anomalous correlations between input and output. Nonetheless, this study demonstrates the adeptness of Transformer networks in managing such data irregularities.



Forget gate
Input gate
Output gate
c_{t-1}
c_t
h_{t-1}
h_t
x_t
f_t
i_t
$\tilde{c}_t$
o_t
sig
tanh
x : Input
h : Hidden State
c : Cell State
$f_t = \sigma(x_t W_x^f + h_{t-1} W_h^f + b_f)$
$i_t = \sigma(x_t W_x^i + h_{t-1} W_h^i + b_i)$
$o_t = \sigma(x_t W_x^o + h_{t-1} W_h^o + b_o)$
$\tilde{c}_t = \tanh(x_t W_x^g + h_{t-1} W_h^g + b_g)$
$c_t = \sigma(f_t * c_{t-1} + i_t * \tilde{c}_t)$
$h_t = \tanh(c_t) * o_t$

Fig. 3. LSTM cell.

3. Long-short term memory networks (LSTM)

The LSTM network is a type of RNN along with a GRU (Gated Recurrent Unit) network. It is composed of a cell, an input gate, an output gate, and a forget gate to reinforce the long-term memory proposed in Hochreiter (1997). LSTMs are well suited for making time-series predictions with longer-term sequences and dependencies with this unique structure. However, the inherent sequential computation in the LSTM leads to an extended processing time that limits the online MPC application as a prediction model where a specific cycle time must be met. From a train-ability point of view, the prediction accuracy with the capability of accommodating the special input-out data structure for multistep-ahead prediction is compared to the proposed Transformer model. The LSTM block and equations for each gate are shown in Fig. 3.

4. Transformer model MPC development

This section discusses the development of a Transformer model-based MPC system. It introduces the training data modification to obtain the multistep prediction model, layer configurations of the Transformer network, and an MPC framework that accommodates the Transformer prediction model.

4.1. Training data preparation

A simple single-input single-output (SISO) first-order plus dead-time (FOPDT) model is chosen for the process model for illustrative purposes. Eq. (3) shows the continuous FOPDT equation between process input (u) and output (y), where K_p, τ_p, and θ_p represent the model parameters for process gain, time constant, and dead time, respectively. The parameter values used for the FOPDT model are also shown in Eq. (3).

$$\tau_p \frac{dy(t)}{dt} = -y(t) + K_p u\left(t - \theta_p\right) \qquad (3)$$
$$\text{where, } K_p = 1, \tau_p = 2, \text{ and } \theta_p = 0$$

The training data is generated by simulating the FOPDT model with randomly created input (u) sequences for 1600 data points. The simulated data is split into two sets, one for training and the other for validation. The validation set is used to monitor and prevent overfitting of the NN model. An additional approach to handle correlated multivariate control inputs is to build a Principal Component Analysis (PCA) based model that reduces the input space or constrains the squared prediction error (SPE) to follow the PCA model during training. With unachievable setpoints, the PCA-based SPE constraint helps to maintain achievable setpoints (Hassanpour et al., 2020, 2022). The data in this study does not have correlated inputs because they are randomly generated. The training and validation data are reshaped to fit the time-series NN model structure. Although the Transformer and LSTM models use entirely different algorithms, the dimension of the input and output vector is the same for the same sequence data. The three-dimensional vector, also called a tensor, can be formed with ($Batch\ size \times Length\ of\ one\ snapshot \times Number\ of\ variables$) and the size of a tensor varies depending on the size of prediction steps such as one-step (os) and multistep (ms) ahead prediction. Fig. 4 visually describes the batch size or Number of snapshots (T), length of one snapshot ($w + P$), and number of variables, respectively. The lengths of the look-back window ($w = 5$) and the prediction horizon ($P = 10$) have been carefully selected to reflect the dynamic characteristics of the process. The mathematical expressions of the data structure are shown in Eqs. (4) and (5) for one-step- and multistep-ahead prediction models, respectively.

$$\begin{aligned} Y_{os} &= \left(y_{k+1}\right) \in \mathbb{R}^{b\times 1\times l} \\ X_{os} &= \left(\left(u_{k-w}, \dots, u_k\right), \left(y_{k-w}, \dots, y_k\right)\right) \in \mathbb{R}^{b\times w\times f} \\ &\quad \text{where, } w \le k \le T - P, \\ &\qquad b = T - w - P \end{aligned} \qquad (4)$$

$$\begin{aligned} Y_{ms} &= \left(y_{k+1}, \dots y_{k+P}\right) \in \mathbb{R}^{b\times P\times l} \\ X_{ms} &= \left(\left(u_{k-w}, \dots u_{k+P}\right), \left(y_{k-w}, \dots y'_{k+1}, \dots y'_{k+P}\right)\right) \\ &\qquad \in \mathbb{R}^{b\times (w+P)\times f} \\ &\quad \text{where, } w \le k \le T - P, \\ &\qquad b = T - w - P, \\ &\qquad y'_{k+i} = y_k \ (1 \le i \le P) \end{aligned} \qquad (5)$$

In Eqs. (4) and (5), Y and X represent the reshaped output and input of the NN training, which are also called $labels$ and $features$, respectively. On the other hand, u and y are the process system input and output of the FOPDT model. As expressed in Eqs. (4) and (5), system input and output (u and y) are in the X as a linear auto-regressive and exogenous part of the NN model input, respectively. The symbols b, T, w, and P denote the batch size, entire time-series length, look-back window size, and prediction horizon length, respectively. Including all snapshots k, l, and f denote the current time step, the number of labels, and the number of features, respectively. A notable difference between one-step prediction and multistep prediction data structures is that the Y_{ms} extends the vector to the $k+P$ time step, including the full P-step-ahead values. In contrast, the Y_{os} includes only the first value in the prediction horizon. A similar difference appears in the input feature. The u and y vector in the X_{ms} extend to the $k+P$, while the ones in the X_{os} stop at the k. Important customization needs to be made for the y vector in the X_{ms}. The y values for future time steps are replaced with the y_k shown as y'_{k+i} in Eq. (5). Prior state values are included in the data row for training to maintain the vector shape consistent with the u vector in the X_{ms}. Fig. 5 visually compares the data structures between one-step- and multistep-ahead models for one snapshot sample in the batch. The modified y vector for the multistep prediction model is illustrated in the dashed box with y_k. Fig. 6 illustrates the training data modification for the multistep prediction models.

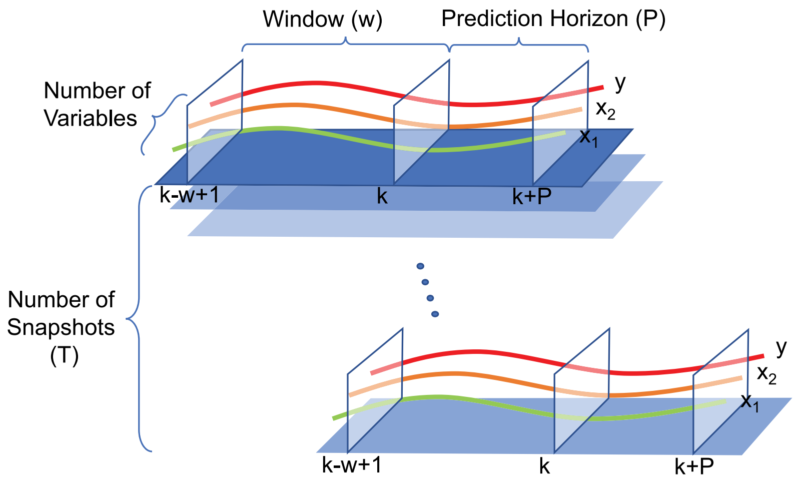


Fig. 4. Training data preparation with receding window snapshots.

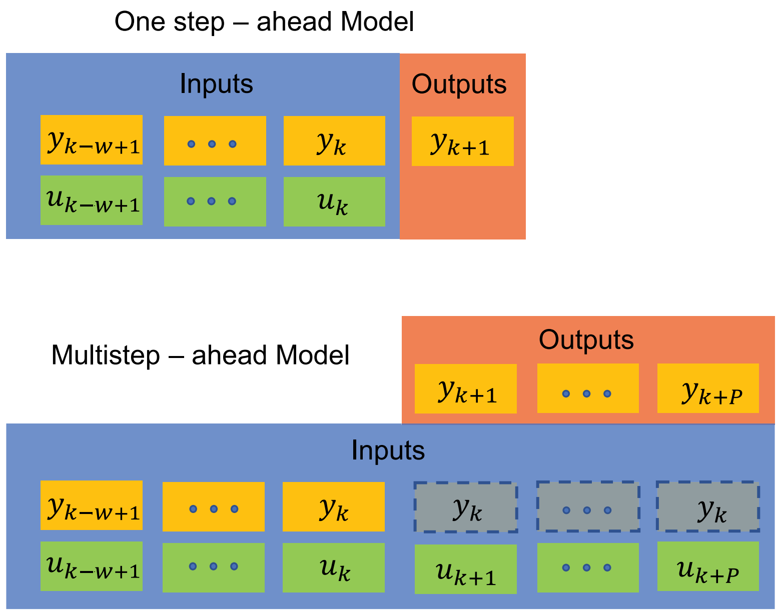


Fig. 5. Training data structure comparison between one-step-ahead model and multistep-ahead model.

4.2. *Transformer model training*

Prepared training data sets for one-step-ahead prediction (Y_{os}, X_{os}) and multistep-ahead prediction (Y_{ms}, X_{ms}) are used for training the Transformer model. Every training result is compared to the result of the counterpart LSTM model. The Transformer model consists of two encoders. The number of encoders and decoders for each system can be determined by hyperparameter optimization. Two encoders were used in this study. Hyperparameter optimization was not used in this study, but could be the subject of future work. Each encoder includes a multi-head attention layer, feedforward layer, dropout layer, and output layer. The multi-head attention layer consists of 10 heads with softmax activation function as described in Eq. (2). The pre-processed inputs X_{os} and X_{ms} are used for Q, K, V vectors for the self-attention mechanism. The attention score vector is concatenated with the input (X) to pose the residual to feed it to the following feedforward layer. One feedforward layer consists of 100 neurons in the hidden layer with a hyperbolic tangent ($tanh$) activation function, and the other output feedforward layer with the same number of neurons with the number of variables in feature X, which is 2 for this study. A dropout layer separates each feedforward layer with a dropout factor of 20%. The attention score vector, which also represents the output of the two encoder blocks, is fed to the final output feedforward layer to map it to the size of the label Y: 1 for the one-step-ahead model and P for the multistep-ahead model.

Meanwhile, the LSTM model consists of three LSTM layers and a dense layer, each separated by a dropout layer with a dropout factor of 20%. Each LSTM layer consists of 100 hidden states, and the current input and the 99 hidden states from the previous time step are fed into the LSTM layers at any time step. The final dense layer takes the 100 outputs generated by the third LSTM layer and maps them to the outputs. Both networks are trained using an Adam optimizer, with loss calculated as a mean-squared error (MSE). The summaries for multistep models are shown in Tables 1 and 2. The number of trainable parameters of weight and bias is significantly less in the Transformer model than in LSTM.

4.3. *Transformer-based surrogate MPC*

Two main parts of MPC are the prediction model and the optimization solver. In the NN-based surrogate MPC, NN models replace the classical transfer function model or physics-based model that generally

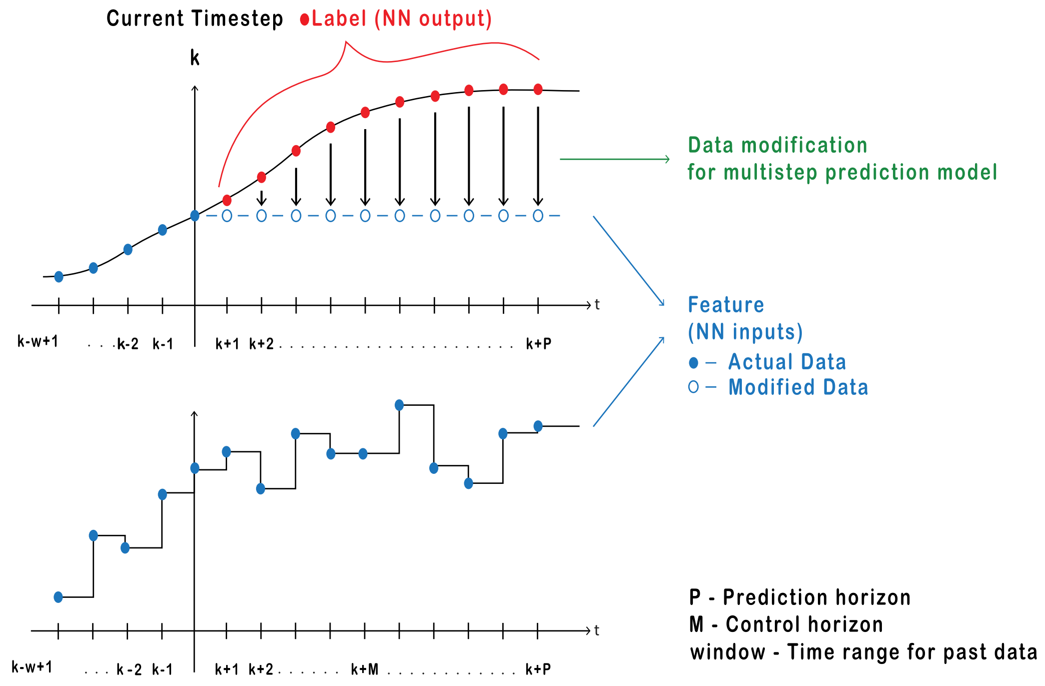

Current Timestep
●Label (NN output)
k
Data modification
for multistep prediction model
Feature
(NN inputs)
● – Actual Data
○ – Modified Data
t
k-w+1 . . . k-2 k-1 k+1 k+2 . . . k+P
k-w+1 . . . k-2 k-1 k+1 k+2 . . . k+M . . . k+P
P - Prediction horizon
M - Control horizon
window - Time range for past data

Fig. 6. Training data modification for multistep-ahead prediction models.

Table 1
Summary of the Transformer model.

Layer	Input dimension	Output dimension	Act. fn
MHA	$Batch \times 15 \times 2$	$Batch \times 15 \times 2$	$SoftMax$
Dense	$Batch \times 15 \times 2$	$Batch \times 15 \times 100$	$tanh$
Dense	$Batch \times 15 \times 100$	$Batch \times 15 \times 2$	$Linear$
MHA	$Batch \times 15 \times 2$	$Batch \times 15 \times 2$	$SoftMax$
Dense	$Batch \times 15 \times 2$	$Batch \times 15 \times 100$	$tanh$
Dense	$Batch \times 15 \times 100$	$Batch \times 15 \times 2$	$Linear$
Dense	$Batch \times 15 \times 2$	$Batch \times 10 \times 1$	$Linear$
Parameters		1758	

Table 2
Summary of the LSTM model.

Layer	Input dimension	Output dimension	Act. fn
LSTM	$Batch \times 15 \times 2$	$Batch \times 15 \times 100$	$tanh$
LSTM	$Batch \times 15 \times 100$	$Batch \times 15 \times 100$	$tanh$
LSTM	$Batch \times 15 \times 100$	$Batch \times 15 \times 100$	$tanh$
Dense	$Batch \times 15 \times 100$	$Batch \times 10 \times 1$	$Linear$
Parameters		203,010	

uses differential equations. The optimization solver minimizes the MPC objective function by searching the best possible u sequences in the control horizon (M). The sum of squared error (SSE) objective function in the general MPC is shown in Eq. (6).

$$\begin{aligned} \min_{\Delta U} \quad & \Phi = (\hat{\boldsymbol{Y}} - \boldsymbol{Y}_{ref})^T \boldsymbol{Q} (\hat{\boldsymbol{Y}} - \boldsymbol{Y}_{ref}) + \Delta \boldsymbol{U}^T \boldsymbol{R} \Delta \boldsymbol{U} \\ s.t. \quad & 0 = f(\dot{x}, x, y, p, d, u) \\ & 0 = g(x, y, p, d, u) \\ & 0 \le h(x, y, p, d, u) \end{aligned} \tag{6}$$

where $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}_{ref}$ are the column vectors for model prediction values and the reference trajectory from the time step $(k+1)$ to $(k+P)$. $\Delta \boldsymbol{U}$ is the column vector for the control moves for the future control horizon (M), from the time step $(k+1)$ to $(k+M)$. $\boldsymbol{Q}$ and $\boldsymbol{R}$ are the diagonal weighting matrices for multiple CVs (Controlled variables) and MVs (Manipulated variables). The detailed descriptions of the MPC equations can be found in Park et al. (2020c).

The trained Transformer models are embedded into the MPC framework serving as prediction models providing the $\hat{\boldsymbol{Y}}$ in Eq. (6). One-step-ahead prediction models for both LSTM and Transformer execute the models P times to get the full prediction for the defined prediction horizon. However, the multistep-ahead prediction models need only one execution for the same task, significantly improving computational efficiency.

5. Case studies for transformer multistep-ahead prediction model MPC

This section discusses three MPC case studies with multistep Transformer models. The case studies include two experiments in the simulation environments and one with an actual temperature control test device, TCLab. The first case study examines one batch of MPC optimization calculation results for a SISO FOPDT process. The second and third case studies discuss continuous multi-input and multi-output (MIMO) MPC tests for a fluidized-bed roaster in gold ore recovery and with the TCLab device. The case studies contrast the improved computational efficiency of the proposed model among the different types of networks (LSTM) and prediction structures (one-step and multistep prediction).

5.1. Case I: First order dynamics model (FOPDT)

Various types and structures of NN models are tested in the surrogate model MPC framework serving as prediction models, and the results are shown in Fig. 7. Trained LSTM and Transformer models with one-step- and multistep-ahead prediction structures provide the model prediction values to the optimization solver. The gradient-based optimization solver iteratively searches the minimum objective function value (SSE). A sequential least-squares programming (SLSQP) solver is used in this study. The surrogate prediction models compute multiple prediction results for every control interval. Fig. 7 shows the optimally predicted CV values based on the given setpoint of 1 and the computation time for different models for one control interval. The plot shows the result of one control interval in Fig. 7 with the time steps in the prediction horizon (P = 10). Compared to the one-step-ahead models, there is a significant improvement in computation

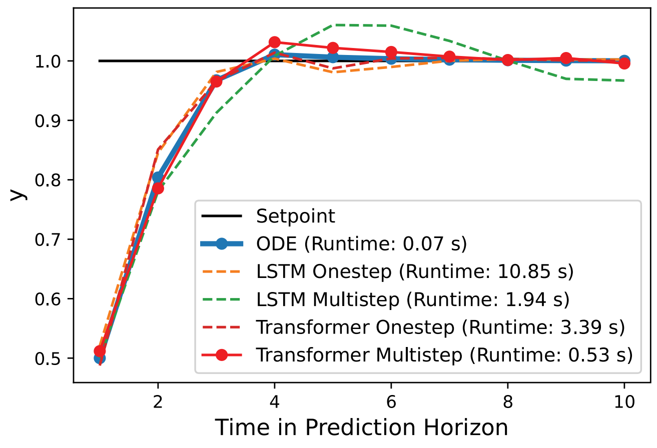
Table 3
Number of sequential computations occurring in one MPC calculation, where, w, P, and n represent a look-back window, prediction horizon, function evaluation of MPC optimization.

Model type	Sequential computation	Multistep prediction	MPC iteration	Number of executions
ODE (Sequential)	1	P	n	$P \times n$
LSTM-OS	w	P	n	$w \times P \times n$
LSTM-MS	$w + P$	1	n	$(w + P) \times n$
Transformer-OS	1	P	n	$P \times n$
Transformer-MS	1	1	n	n

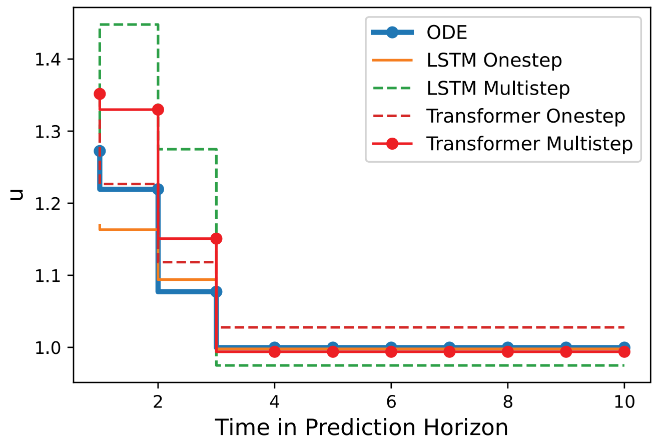
time for both LSTM and transformer with multistep-ahead prediction models. The multistep Transformer model shortens the computation time by about 20 times compared to the one-step LSTM model (from 10.85 s to 0.53 s). Computation time in RNN-based deep learning models is substantially decreased by implementing two key modifications. Firstly, the sequence computation is replaced with a self-attention mechanism, thereby eliminating the recurrence inherent in the one-step ahead LSTM model. Secondly, a multistep ahead prediction structure is applied further enhancing the computational efficiency by calculating the whole length of the prediction simultaneously. The self-attention mechanism in the Transformer model accomplishes the same task with matrix operations reducing the internal sequences by w times less than the LSTM model. The other significant portion that contributes to the computation time is the simultaneous multistep-prediction structure. The multistep prediction structure simultaneously computes the model prediction values for the entire prediction horizon (from $k+1$ to $k+P$), which reduces the number of model calls by P times less than the one-step prediction model. The overall number of sequential computations occurring in one MPC control interval is shown in Table 3.

The number of iterations for MPC calculation (n) in Table 3 can be varied between different models as the numerical solver converges differently with every model. The traditional ODE (Ordinary Differential Equation) model-based MPC is also tested along with the NN models to provide the baseline performance. For a simple FOPDT model, the traditional MPC with a transfer function model outperforms NN models. As system complexity increases, similar NN models are capable of capturing complex, higher-order relationships. In contrast, the effort to build and the computation cost to execute physics-based models increases significantly. The use of multistep Transformers to model complex, nonlinear systems is a natural extension of this work and a topic of ongoing research.

The model accuracy is inferred from Fig. 7. First, the u values are supposed to be at a steady-state '1.0' like the ODE model result because the process gain (K_p) in the FOPDT model is defined as '1.0'. The steady-state u value must be the same as the y value with '1.0' as the output setpoint. However, the u results of the multistep-ahead LSTM model and the one-step-ahead Transformer model converge to slightly different values than 1.0, which means those models have some model mismatch with the process model. In addition, the LSTM multistep model shows a significant model mismatch on both y and u, while the multistep Transformer model follows the ODE model result. The different outcomes can be explained by two distinguished methods to learn the temporal dependency. The suggested training data structure for the multistep-ahead prediction models includes both the past measured data ($k-w : k$) and future prediction horizon data ($k+1 : k+P$), especially for the feature (X) set. As discussed in Section 4.1, this is necessary not only for including the entire prediction horizon data to learn simultaneously but also to pass the control horizon part of the feature ($u_{k+1} : u_{k+M}$) to the SLSQP solver as decision variables to be solved. However, while the prediction horizon is part of the data in the feature set (X), the prediction horizon part of y data ($y_{k+1} : y_{k+P}$) is replaced with the currently measured data (y_k). This modification avoids including the labeled data in the feature set before training while maintaining a consistent data dimension. However, irrelevant data is



(a) Setpoint tracking performance of CV and computation time



(b) Control movement of MV corresponding to given setpoint

Fig. 7. Results of MPC calculation compared to the various types and structures of NN models. Setpoint tracking performance of CV and computation time (Upper), and control movement of MV corresponding to given setpoint (Lower).

included due to this modification, which introduces irregular temporal dependency for time-series data training. The way that the Transformer architecture learns the temporal dependency is completely different than RNNs. The self-attention mechanism employs a conditional probability, selectively including the useful relationship and excluding the irrelevant relationship between input and output data. Therefore, the Transformer is more effective in learning the irregular temporal dependency than LSTM that structurally includes the temporal dependency in a sequential manner.

5.2. Case II: Fluidized-bed roaster model

The roasting process is a part of gold production that recovers the gold from the ore. The substances that negatively affect the carbon-in-leach (CIL) process are removed by the oxidation reaction in the roasting process. The particular type of gold ore, called refractory gold ore, has several characteristics that interfere with satisfactory gold recovery in the traditional CIL. The iron sulfide in the refractory gold ore confines the gold in the minerals that prevent contact with the cyanide solution. The carbonaceous minerals in the ore tend to re-absorb the recovered gold component from the cyanide solution. Thus, the refractory gold ore needs to be pretreated in the roaster, converting the iron sulfide mineral and carbonaceous mineral to oxides. The roaster consists of two stages of fluidized bed reactors (FBRs)

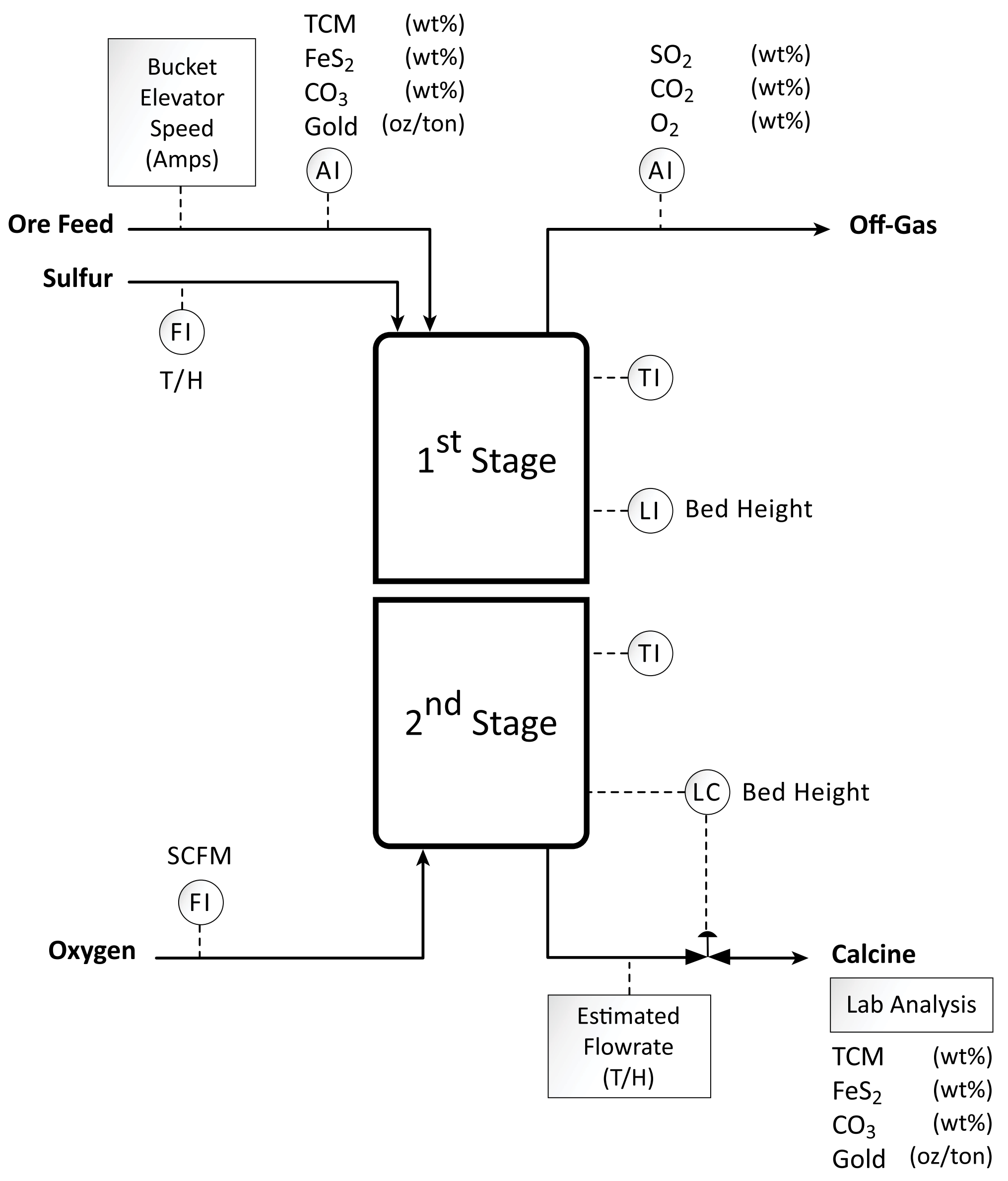


Fig. 8. Schematic drawing of the roaster process.

that oxidize sulfide compounds and organic carbon in the ore. Before the roasting process, the ore is crushed into fine particles with an average size of 75 μm in diameter by passing them through a grinding process. The fine particles are fed into the roaster from the top and are fluidized by the upward force generated by the oxygen flow from the bottom of the roaster. Three critical factors that affect the oxidation reaction are ore particle size, oxygen content in the combustion gas, and bed temperature. Compared with traditional air-roasting, oxygen-roasting yields better gold recovery by improving reaction efficiency. Compared to oxygen roasting, air roasting requires a longer retention time at a higher temperature to maintain a high conversion of sulfidic metals and carbonaceous matter. These conditions easily over-roast and soften the ore particles (glassy flux), blocking the pores of the material from contacting the combustion gas. Over-roasting also affects the subsequent CIL process by encapsulating the gold in the glassy fluxed ore (Wells Jr., 1948; Smith et al., 1990; Thomas and Cole, 2016; Qin et al., 2021). The exothermic reaction generates most of the thermal energy required for the oxidation reaction. Additional heat is provided by feeding sulfur prills as fuel. The major reactions (Thomas and Cole, 2016) are shown below and a diagram of the roasting process is shown in Fig. 8.

Combustion of organic carbon:

$$C(s) + O_2(g) \rightarrow CO_2(g)$$

Combustion of iron sulfides (pyrite and pyrrhotite):

$$7FeS_2(s) + 6O_2(g) \rightarrow Fe_7S_8(s) + 6SO_2(g)$$

$$4Fe_7S_8(s) + 53O_2(g) \rightarrow 14Fe_2O_3(s) + 32SO_2(g)$$

Combustion of sulfur prills (fuel):

$$S(s) + O_2(g) \rightarrow SO_2(g)$$

Retention of Sulfur dioxide:

$$2CaCO_3(s) + 2SO_2(g) + O_2(g) \rightarrow 2CaSO_4(s) + 2CO_2(g)$$

The actual operation data is usually insufficient to train the NN surrogate models because the actual operations are often maintained in a narrow range that does not provide the information for the broader operation range where the optimal operation point may exist. Thus, a physics-based model is developed for process simulation experiments, including data generation for training NN models and emulating the actual process as a digital twin model during the controller performance

Table 4
Lumped parameters from physics-based roaster model.

Quantity	Value
Average surface area of ore particle (A_p)	0.0176 m^2
Average radius of ore particle (R_p)	37.5 μm
Atmospheric pressure (p)	1 atm
Molecular weight of Oxygen (M_{O_2})	32 g/mol
Lennard-Jones potential parameter for Oxygen molecules (σ)	3.433 Å
Lennard-Jones potential parameter for Oxygen molecules (ε/K)	113 K
Average porosity of ore particles (ε_p)	0.5
Effective diffusivity ($\mathscr{D}_e$)	
Diffusivity for gases ($\mathscr{D}_{AM}$)	
Mass transfer coefficient between air and particle (k_g)	
Rate constant for the surface reaction (k'')	
Reactor temperature (T)	
Tortuosity (τ)	
Collision Integral for diffusivity (Ω)	

test. The physics-based model for this research consists of heat and mass balance equations and reaction kinetics. The reaction kinetics uses a shrinking core model (SCM), whose reaction rate is dominated by pore diffusion.

5.2.1. Physics-based model of a roaster process

A physics-based model for the roaster is developed to generate the Transformer model training and validate the Transformer model-based MPC performance. The SCM is used for reaction kinetics of ore particle oxidation reaction shown in Eq. (7). The rate equation of the SCM consists of three parts: Diffusion through gas film controls, diffusion through ash layer control, and chemical reaction controls. The diffusion through gas film control is negligible for engineering applications because it has an effect only for a short period of the initial stage. Once the ash layer starts forming on the particle surface, the ash layer diffusion step controls the overall reaction rate (Kunii and Levenspiel, 1991; Ahn and Choi, 2017). The ash layer diffusion term requires a calculation of the effective diffusivity, $\mathscr{D}_e$. The Chapman–Enskog equation is used for calculating the effective diffusivity ($\mathscr{D}_e$) shown in Eqs. (8)– (10) (see Table 4).

$$-r_i = \frac{A_p C_i}{\frac{1}{k_g} + \frac{R_p}{6\mathscr{D}_e} + \frac{1}{k''}} = \frac{A_p C_i}{\frac{R_p}{6\mathscr{D}_e} + \frac{1}{k''}} \tag{7}$$

$$\mathscr{D}_{AM} = 0.0018583\sqrt{T^3\left(\frac{1}{M_{O_2}} + \frac{1}{M_{O_2}}\right)}\frac{1}{p\sigma_{AB}^2\Omega_{\mathscr{D},AB}} \tag{8}$$

$$\Omega_{\mathscr{D},AB} = \frac{1.06036}{T^{*0.15610}} + \frac{0.19300}{\exp(0.47635T^*)} + \frac{1.03587}{\exp(1.52996T^*)} + \frac{1.76474}{\exp(3.89411T^*)} \quad where,\ T^* = \kappa T/\varepsilon \tag{9}$$

$$\mathscr{D}_e = \frac{\mathscr{D}_{AM}\varepsilon_p}{\tau}, \quad where \quad \tau = \varepsilon_p^{-0.41} \tag{10}$$

$$k'' = A_p T \exp\left(-E_r/RT\right) \tag{11}$$

$$\frac{dN_i}{dt} = F_{i,in} - F_{i,out} + r_i \tag{12}$$

$$N\,C_p\frac{dT_i}{dt} = F_{i,0}H_{in} - F_{i,out}H_{out} + \Delta H_{reaction} \tag{13}$$

Table 5
Operation variables of roaster simulation model.

Variables in roaster model	
Input variables	Output variables
	O_2 in Offgas (wt%)
Ore feed amount (Amp)	CO_2 in Offgas (wt%)
Sulfur prill (T/H)	SO_2 in Offgas (wt%)
Oxygen (SCFM)	TCM in Calcine (wt%)
Carbon content in Ore (wt%)	FeS_2 in Calcine (wt%)
Iron Sulfide content in Ore (wt%)	$CaCO_3$ in Calcine (wt%)
Carbonate content in Ore (wt%)	Stage1 Temperature (F)
	Stage2 Temperature (F)

5.2.2. Train transformer model for roaster MPC

Training data is generated by simulating the physics-based roaster model. The roaster simulation model has fourteen main operation variables, including six input and eight output variables. The first set of input variables is related to the quantity of reactor feed, such as the quantity of ore feed, the mass flow rate of sulfur prill, and the volume flow rate of oxygen. These are the primary operation variables of roaster operation and are the candidates for an advanced process control system manipulated variables (MV). The second set of input variables is the quality of the ore feed, such as the contents of a specific component. Compared to the first set of inputs, the second set is not adjustable for the process operation; thus, it is considered an advanced process control system disturbance variable (DV). The roaster model in this research includes organic carbon, iron sulfide, and carbonate content in the feed ore. The random input signals for the six reactor input variables are generated and simulated in the reactor model, which creates the associated output variable responses. The random input signals and output variable responses are then used for training the Transformer model. The reactor output variables in this case study include the off-gas and calcine specifications and the two reactor temperatures. The input and output variables of the roaster process are shown in Table 5 and the roaster operation range for the random input signals refer to Thomas and Cole (2016). Fifteen thousand sample points are generated for training the Transformer model, and a short segment of the input and output data is shown in Figs. 9 and 10. Fig. 10 displays the Transformer model fitting and Fig. 11 shows the Transformer model validation. The blue dashed line in the figure represents the sampled data from the roaster physics-based simulation model, and the red line shows the prediction results with the Transformer model. The model fitting results validate that the Transformer model can provide accurate model predictions for control.

The network configuration for the Transformer model is shown in Table 6. It consists of Multi-head attention (MHA) layers followed by a set of Feedforward layers, as depicted in Fig. 2 in the previous Section 2. The tensor dimension for each layer represents: $(Batch) \times (Length\ of\ one\ snapshot) \times (Number\ of\ variables)$, where the $Batch$ size for the roaster training is the same as the length of the time series. Each snapshot includes the past data with a length of look-back window size, of six, plus the prediction horizon length, of ten, which forms the second dimension of MHA input tensor, sixteen. The output dimension of the last feedforward layer ($Dense$ layer) is reduced to the size of the multistep prediction result with the length of the prediction horizon, ten, and the number of system output variables, eight: $(Batch \times 10 \times 8)$.

5.2.3. Control results for roaster

An MPC controller is designed as a case study to test the prediction accuracy and solution time of the Transformer based model in the MPC system. This case study selects two controlled variables (CVs) and two manipulated variables (MVs). The two reactor temperatures at each stage of the roaster are chosen for two CVs, and ore feed rate and sulfur prill inlet flows are chosen for the two MVs. The reactor temperature is one of the most critical operation variables directly affected by the reaction status. Thus, it is commonly selected as a major CV in many

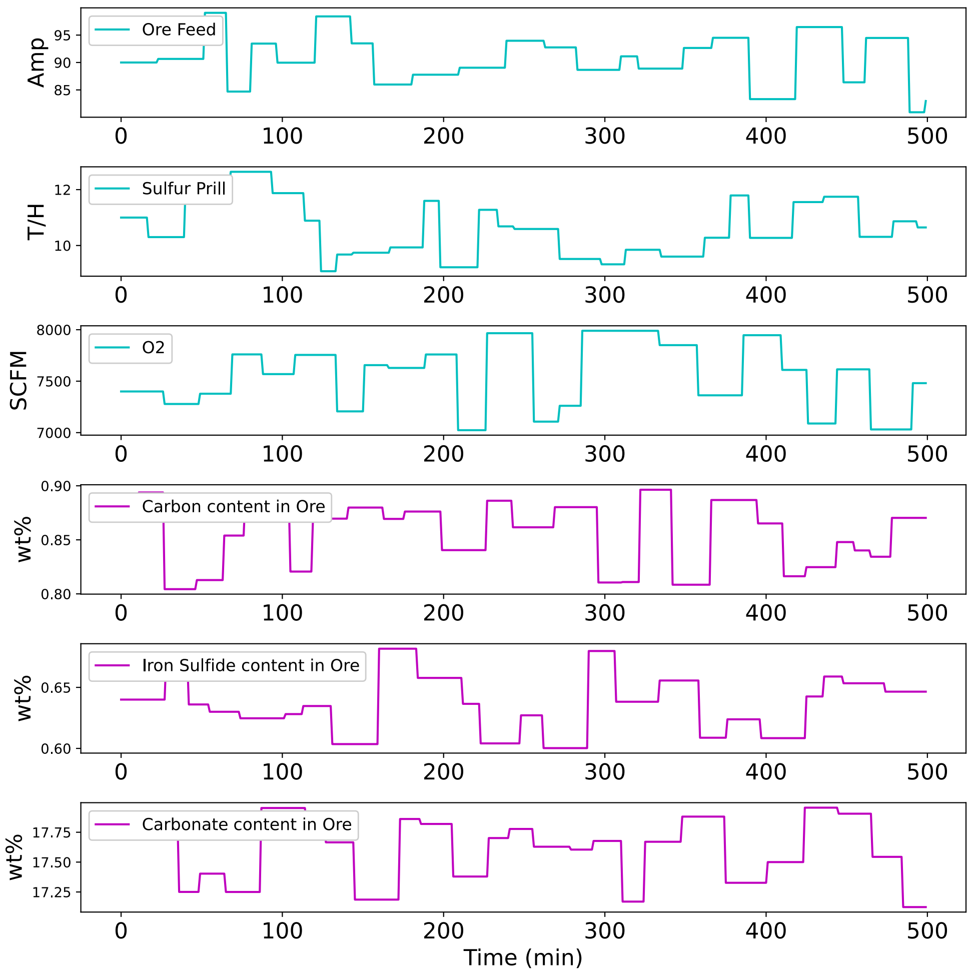


Fig. 9. Roaster input data for Transformer model training.

Table 6
Summary of the Transformer model for roaster data.

Layer	Input dimension	Output dimension	Act. fn
MHA	$Batch \times 16 \times 14$	$Batch \times 16 \times 14$	$SoftMax$
Dense	$Batch \times 16 \times 14$	$Batch \times 16 \times 100$	$tanh$
Dense	$Batch \times 16 \times 100$	$Batch \times 16 \times 14$	$Linear$
MHA	$Batch \times 16 \times 14$	$Batch \times 16 \times 14$	$SoftMax$
Dense	$Batch \times 16 \times 14$	$Batch \times 16 \times 100$	$tanh$
Dense	$Batch \times 16 \times 100$	$Batch \times 16 \times 14$	$Linear$
Dense	$Batch \times 16 \times 14$	$Batch \times 10 \times 8$	$Linear$
Parameter		26,216	

Table 7
Settings of Roaster MPC.

Parameters	Setting
Prediction horizon (P)	10 min
Control horizon (M)	6 min
CV Weight for SSE (q_{T1})	1×10^3
CV Weight for SSE (q_{T2})	1×10^2
MV Weight for rate of change ($r_{Ore\ feed}$)	1×10^3
MV Weight for rate of change ($r_{Sulfur\ prill}$)	1
MV Max move (Ore feed)	1 Amp
MV Max move (Sulfur prill)	0.2 T/H
MV Bounds for SLSQP (Ore feed)	80–100 Amp
MV Bounds for SLSQP (Sulfur prill)	9–13 T/H

advanced process control applications inferring the conversions. The ore particles are transported by the bucket elevator to the 1st stage roaster, and the feed rate is adjusted by changing the speed of the bucket elevator. The sulfur prill feed rate is assigned to the second MV that maintains the reaction temperature requirement to ensure high sulfide sulfur and organic carbon conversion. The other process input variables, including the ore feed composition and oxygen inlet flow, are retained at a constant value during the experiment. The MPC settings are shown in Table 7.

A simulated MPC control result for the roaster process is shown in Fig. 12. In this test, two temperature setpoints are changed to validate the setpoint tracking performance of the Transformer model-based MPC. At 20 min, MPC is activated and starts controlling the temperatures. Ore and sulfur feed rates as MVs move to meet the temperature setpoints based on the MPC control calculation. To increase the temperatures at a given condition, MPC decreases ore feed and increases the sulfur prill feed because the sulfur prill has a higher reaction rate than ore particle as it does not go through the ash layer diffusion. This demonstrates that the Transformer controller provides proper prediction by taking into account the reaction kinetics of gold ore roasting. Although this case study only includes a subset of operation variables, the complete set of variables can be included in future research studies not only for the MPC application but also for the real-time optimization application that deals with an economic objective and steady-state targets.

5.3. Case III: Temperature control lab device

In this section, an experiment is conducted using a temperature control lab (TCLab). The device consists of two heater inputs and

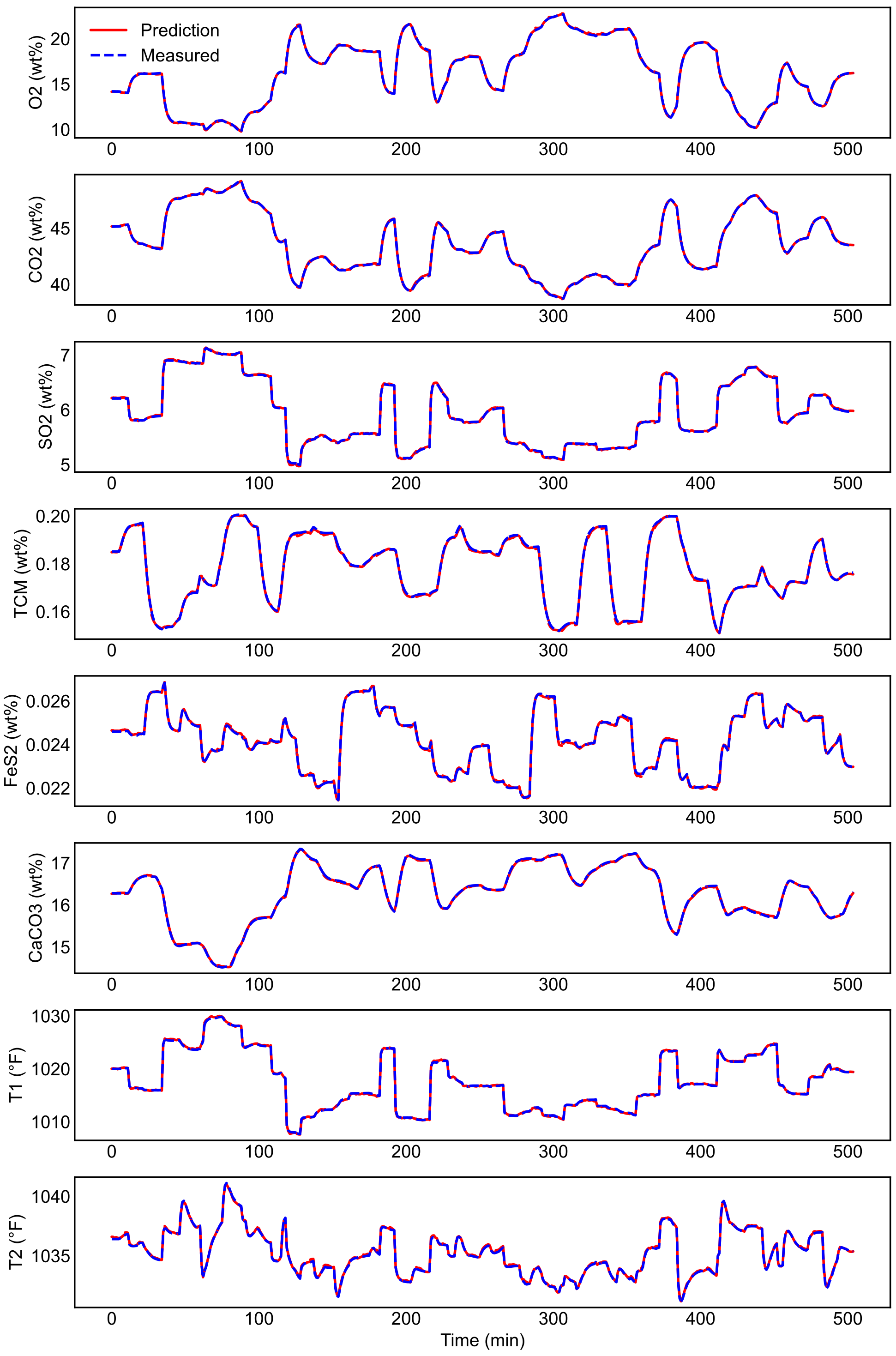


Fig. 10. Transformer model training result for roaster.

two temperature output measurements. This device has been used for control education and research purposes, demonstrating control theory and machine learning-based automation methods (Park et al., 2020b).

In the experiment, the new Transformer multistep-ahead prediction model is developed for the TCLab device and embedded into the MPC algorithm. The newly proposed Transformer multistep-ahead prediction model is compared with three other options, including the Transformer one-step-ahead prediction model and one-step and multistep-ahead prediction models with the LSTM network. In the TCLab case study, all four different model types are tested for offline model development to show model fitness. Then, two of the four models are chosen to validate the online control performance: the multistep Transformer model and the one-step LSTM model. The two models are compared for the computational efficiency and setpoint tracking performance in the online control comparison.

The training data is generated with the TCLab device with random step test input signals for the two heaters. The step-test random signals are generated with a combination of two individual random number generators. The first regulates the amplitude of step changes, while the second generates the number of intervals between the step changes. The

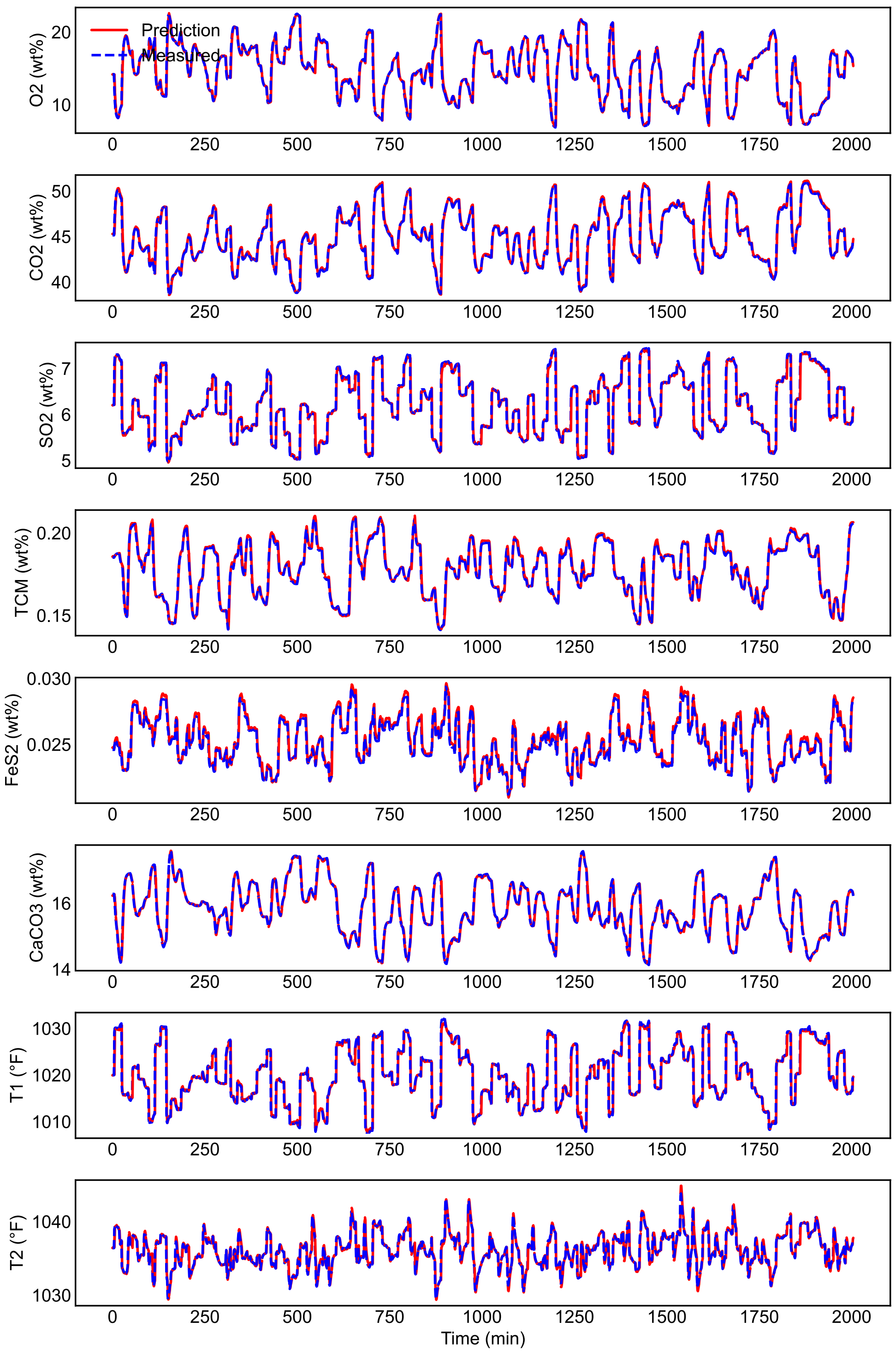


Fig. 11. Transformer model validation result for roaster.

amplitude range for the first random number generator is zero to one hundred percent of the maximum heater value of the TCLab. The step interval range for the second random number generator is in the range of three hundred to six hundred seconds to account for the dynamic behavior of the TCLab. This step-change interval distribution ensures the collection of various input–output relationships with variable frequency responses.

Data is collected every second for twenty-five thousand seconds and down-sampled every 30 s balancing the control calculation speed and control performance in the MPC application. The first training data set includes the first 3000 s of the data to represent a case of insufficient training data. The other training session uses the full 25,000 s of data. The validation results with MAE values for each model are shown in Table 8. In addition, our case studies have certain limitations when confirming the LSTM model performance with irregular temporal dependencies. While this has been repeatedly demonstrated in the field of natural language processing, the model prediction accuracy shown in Table 8 does not fully encapsulate the MS LSTM viability as an MPC prediction model. Thus, a more extensive study with a primary focus on prediction performance could be an interesting topic

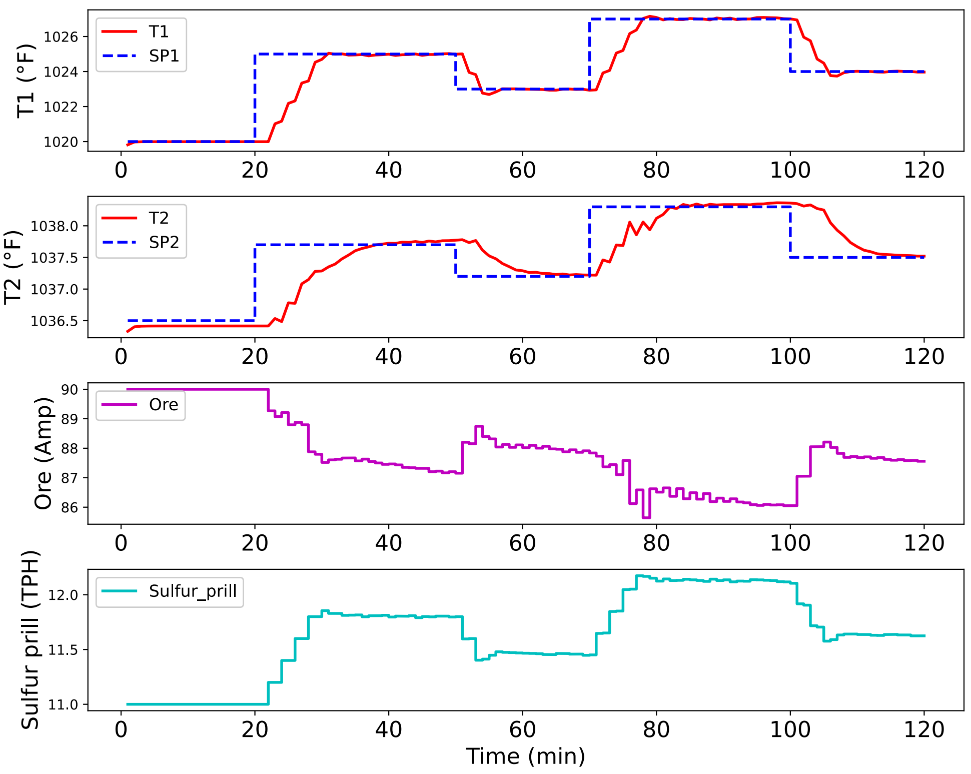


Fig. 12. MPC result for roaster temperature control.

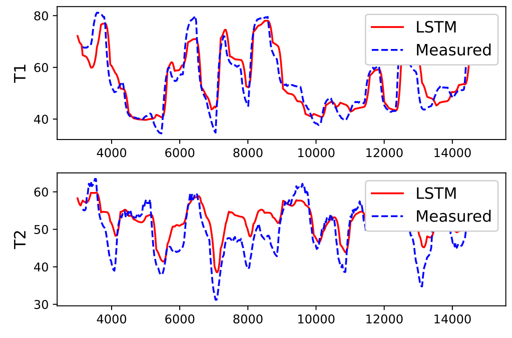
Table 8
Training result comparison.

Prediction type	Network type	Variables	MAE	
			Quantity of training data	
			3000 s	24,000 s
One-step	LSTM	T_1 (°C)	4.03	3.82
		T_2 (°C)	3.88	2.64
	Transformer	T_1 (°C)	2.16	3.32
		T_2 (°C)	5.26	5.63
Multistep	LSTM	T_1 (°C)	3.59	2.59
		T_2 (°C)	2.70	1.57
	Transformer	T_1 (°C)	2.89	2.61
		T_2 (°C)	2.87	1.47

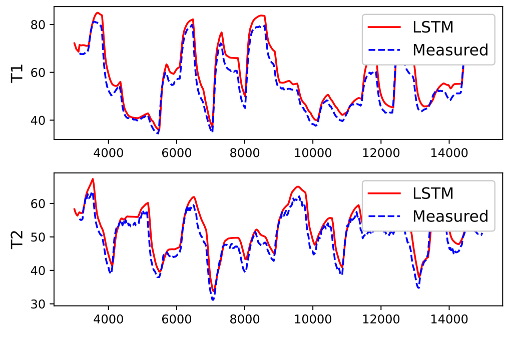
for future research. Fig. 13 depicts the model training results of using either a small data set or large data to train one-step LSTM models and multistep Transformer models on TCLab data. Fig. 13(a), shows the one-step model training result using the small three thousand seconds of data for training. The model predicts the general shape of the measured data but is not able to accurately predict the measured values. Fig. 13(b) shows the training result of the one-step LSTM model trained on the large twenty-five thousand seconds of data. The model accurately predicts both the shape of the measured data and the values of the data over the training range. Fig. 13(c) shows the training result for the Transformer one-step model using three thousand seconds of data. Compared with the LSTM one-step model trained in Fig. 13(a) on the smaller data range, overall accuracy is greatly improved. In addition, when Fig. 13(c) is compared with the LSTM model trained on the large set of data depicted in Fig. 13(b), the Transformer model, despite having less training data, is more accurate in certain intervals than the LSTM. Fig. 13(d) shows the results for using the twenty-five thousand-second data set on the Transformer one-step model. Of the four models in Fig. 13, it is best able to predict the measured data across the entire range.

Fig. 14 depicts the results of using either a small data set or large data to train multi-step LSTM models and Transformer models on TCLab data. Fig. 14(a) shows the results of using the smaller three thousand second data to train a multistep LSTM model. Fig. 14(b) depicts the results of using the larger twenty-five thousand data set to train the multistep LSTM model. The model accurately predicts the measured data. Fig. 14(c) show the results of training a multistep Transformer model on the small three thousand second data. It is more accurate than the multistep LSTM trained on 3000 s of data shown in Fig. 14(a), and in certain ranges, it is more accurate than the multistep LSTM model trained on twenty-five thousands seconds of data shown in Fig. 14(b). Fig. 14(d) depicts the result of using the twenty-five thousand seconds data set to train the multistep Transformer model. This model is the most accurate of the four models, as expected.

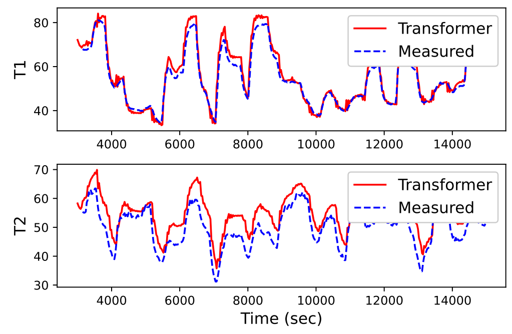
The LSTM one-step prediction model and the Transformer multistep prediction model are chosen to validate the online control performance, as shown in Fig. 15. As observed in the FOPDT model experiment in Section 5.1, a significant difference in MPC calculation time is observed between LSTM one-step model and the Transformer multistep model. The solution time at each control interval is also shown in Figs. 15(a) and 15(b), along with the TCLab system variables. The average solution time of the Transformer multistep model is 1.64 s. In contrast, the LSTM one-step model takes 28.5 s on average, which is more than fifteen times slower than the Transformer multistep model. The MPC solution time affects the control performance in many different ways. The slower solution time must be considered to select the control interval during the MPC design. In this test, for example, the control interval for the Transformer multistep model can be set to 2 s to account for overhead communication time and variable solver iterations. However, the LSTM one-step model control interval is set to 30 s as it needs longer than 28.5 s of average MPC solution time. This slower MPC solution time also affects the overall control performance. The mean absolute error (MAE) value between setpoint and temperatures measures the setpoint tracking performance of each type of controller and is observed visually in Figs. 15(a) and 15(b). Note that the total simulation time to fulfill the identical setpoint sequences for both controllers is different:



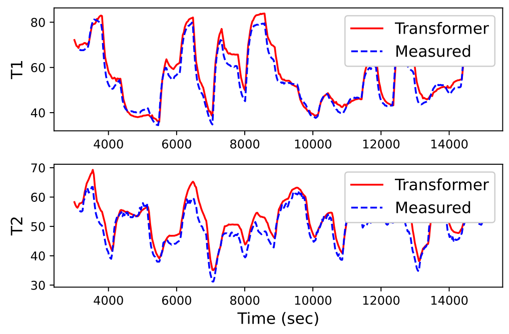
(a) LSTM (one-step model): using 3,000 seconds of data



(b) LSTM (one-step model): using 25,000 seconds of data)

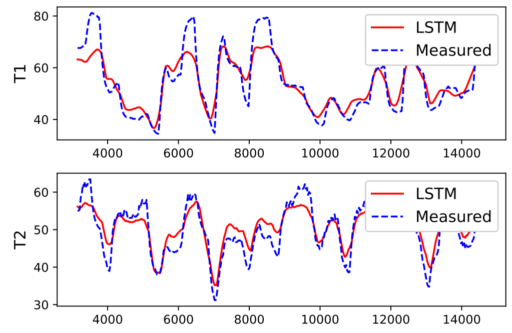


(c) Transformer (one-step model): using 3,000 seconds of data)

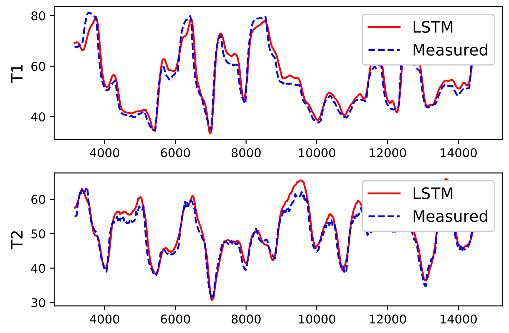


(d) Transformer (one-step model): using 25,000 seconds of data)

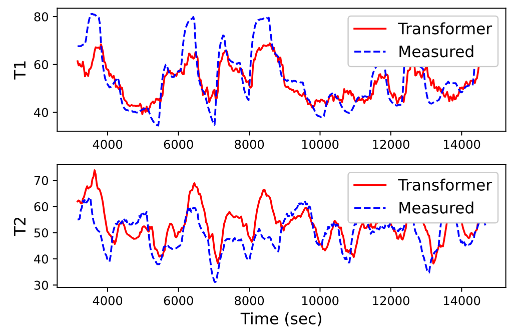
Fig. 13. One-step Models validation for LSTM and Transformer: in comparison with using 3000 (a, c) and 25,000 s data (b, d).



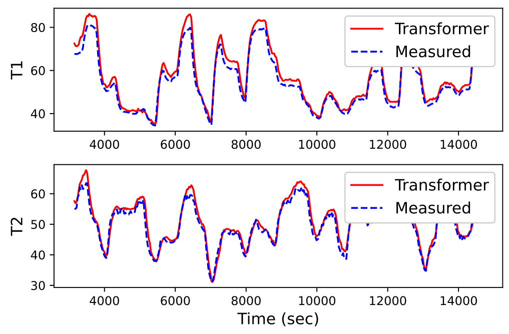
(a) LSTM (Multistep model): using 3,000 seconds of data



(b) LSTM (Multistep model): using 25,000 seconds of data)

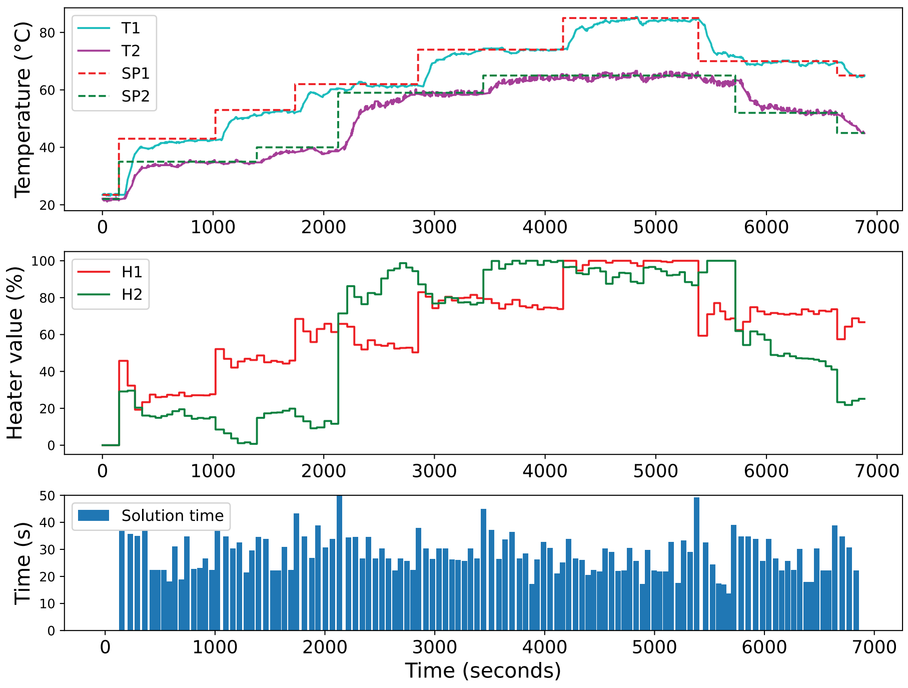


(c) Transformer (Multistep model): using 3,000 seconds of data)

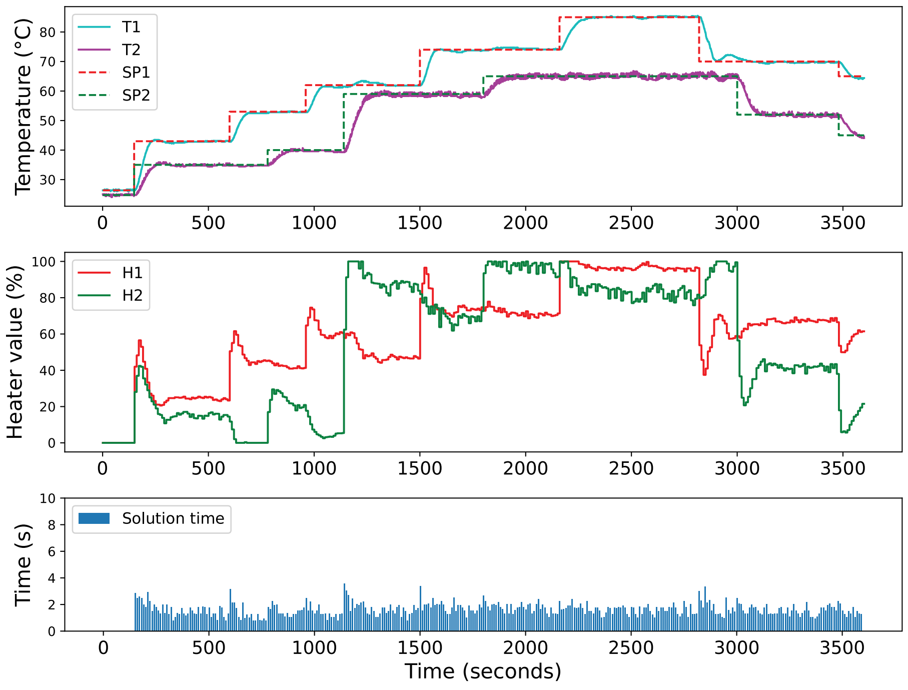


(d) Transformer (Multistep model): using 25,000 seconds of data)

Fig. 14. Multi-step Models validation for LSTM and Transformer: in comparison with using 3000 (a, c) and 25,000 s data (b, d).



(a) TCLab MPC result: LSTM (One-step model)



(b) TCLab MPC result: Transformer (Multistep model)

Fig. 15. TCLab MPC control result comparison between one-step LSTM model and Multistep Transformer model.

sixty-eight hundred seconds for LSTM one-step model and thirty-six hundred seconds for the Transformer multistep model. The surrogate MPC efficiency affects setpoint tracking performance or disturbance compensating performance. For example, the rise time, the amount of time that the controlled variable (CV) reaches the setpoint for the first time since the latest setpoint has been changed, is 100 s, while the rise time of the LSTM takes more than 500 s on average. The weights for the MPC objective function for both controllers are set to the same settings. Table 9 summarizes the control performance metrics.

6. Conclusions

A novel Transformer with a multistep prediction model structure is proposed for MPC. The new multistep prediction model structure is tested with the Transformer NN architecture. The Transformer model demonstrates improved computation time with the multistep prediction structure compared to the one-step prediction structures with the LSTM model. The parallel data processing of the Transformer architecture eliminates the sequential computation nature of RNN-based models, including LSTM. Additionally, the simultaneous multistep prediction

Table 9
MPC performance comparison between LSTM and Transformer models.

Model type		LSTM one-step	Transformer multistep
Simulation time		6885 s	3600 s
Control interval		30 s	10 s
Number of MPC executions		115	345
MPC solution time	Min	13.61 s	0.79 s
	Max	51.07 s	3.60 s
	Avg	28.54 s	1.64 s
MAE (Mean absolute error)	T1 (°C)	10.97	8.26
	T2 (°C)	11.03	8.01

structure removes the recursive procedure of the one-step prediction model that updates the prediction horizon value one step at a time. As a result, the proposed multistep Transformer model calculates the entire length of the model prediction with a single execution of the forward propagation. The increased computational speed of a control system, such as with MPC, substantially improves the control performance by enabling a shorter control interval in real-time control practice.

In addition, the self-attention mechanism in the Transformer model learns the irregular temporal dependencies more effectively than the sequential learning mechanism in the LSTM. The irregular temporal dependencies introduced in the multistep prediction structure are properly handled with the Transformer model compared to the LSTM, which is briefly shown in the first case study result (Fig. 7). The multistep Transformer model shows more dependable accuracy than the multistep LSTM model, even with a smaller model size, containing fewer trainable parameters (Tables 1 and 2).

This study does not carry out a thorough evaluation of the model quality across a variety of problem complexities or examine how different types of models, including alternatives like MS-LSTM and OS-Transformer, would perform under these conditions. Such an investigation could serve as an intriguing focus for future research endeavors.

CRediT authorship contribution statement

Junho Park: Served as the main author of this publication, Contributing significantly to the conceptualization of ideas, Generation of results, Writing of the paper as the lead author. **Mohammad Reza Babaei:** Played a vital role in the project by providing technical assistance, Contributing to the development of ideas, Actively participating in the writing and revision process of the manuscript. **Samuel Arce Munoz:** Provided valuable technical support throughout the project. **Ashwin N. Venkat:** Contributed to the publication by providing direction, Overseeing project management, Acquiring the necessary funding. **John D. Hedengren:** Contributed to the manuscript by providing valuable input on ideas, Participating in the writing process, Assisting in the revision of the manuscript.

Declaration of competing interest

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

Data availability

Link to GitHub repository: https://github.com/BYU-PRISM/Transformer_MPC

Acknowledgment

All authors reviewed and approved the final version of the manuscript. The authors acknowledge the financial and technical support from Seeq Corporation under grant R0602660.

References

Ahn, H., Choi, S., 2017. A comparison of the shrinking core model and the grain model for the iron ore pellet indurator simulation. Comput. Chem. Eng. 97, 13–26. http://dx.doi.org/10.1016/j.compchemeng.2016.11.005.

Al Seyab, R.K., Cao, Y., 2008. Differential recurrent neural network based predictive control. Comput. Chem. Eng. 32 (7), 1533–1545. http://dx.doi.org/10.1016/j.compchemeng.2007.07.007.

Bahdanau, D., Cho, K.H., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. pp. 1–15.

Cho, K., Van, M.B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. J. Clin. Microbiol. 28 (4), 828–829. http://dx.doi.org/10.1128/jcm.28.4.828-829.1990.

Drgoňa, J., Tuor, A.R., Chandan, V., Vrabie, D.L., 2021. Physics-constrained deep learning of multi-zone building thermal dynamics. Energy Build. 243, http://dx.doi.org/10.1016/j.enbuild.2021.110992.

Gopaluni, R.B., Tulsyan, A., Chachuat, B., Huang, B., Lee, J.M., Amjad, F., Damarla, S.K., Woo Kim, J., Lawrence, N.P., 2020. Modern machine learning tools for monitoring and control of industrial processes: A survey. IFAC-PapersOnLine 53 (2), 218–229. http://dx.doi.org/10.1016/j.ifacol.2020.12.126.

Hassanpour, H., Corbett, B., Mhaskar, P., 2020. Integrating dynamic neural network models with principal component analysis for adaptive model predictive control. Chem. Eng. Res. Des. 161, 26–37. http://dx.doi.org/10.1016/j.cherd.2020.03.031.

Hassanpour, H., Corbett, B., Mhaskar, P., 2022. Artificial neural network-based model predictive control using correlated data. Ind. Eng. Chem. Res. 61 (8), 3075–3090.

Hochreiter, S., 1997. Long short-term memory. Neural Comput. 1780, 1735–1780.

Kunii, D., Levenspiel, O., 1991. Fluidization Engineering. Butterworth-Heinemann.

Lee, M., Park, S., 1992. A new scheme combining neural feedforward control with model-predictive control. AIChE J. 38 (2), 193–200. http://dx.doi.org/10.1002/aic.690380204.

Lee, J.H., Shin, J., Realff, M.J., 2018. Machine learning: Overview of the recent progresses and implications for the process systems engineering field. Comput. Chem. Eng. 114, 111–121. http://dx.doi.org/10.1016/j.compchemeng.2017.10.008.

Machalek, D., Quah, T., Powell, K.M., 2021. A novel implicit hybrid machine learning model and its application for reinforcement learning. Comput. Chem. Eng. 155, 107496. http://dx.doi.org/10.1016/j.compchemeng.2021.107496.

Machalek, D., Tuttle, J., Andersson, K., Powell, K.M., 2022. Dynamic energy system modeling using hybrid physics-based and machine learning encoder–decoder models. Energy AI 9, 100172. http://dx.doi.org/10.1016/j.egyai.2022.100172.

Mesbah, A., Wabersich, K.P., Schoellig, A.P., Zeilinger, M.N., Lucia, S., Badgwell, T.A., Paulson, J.A., 2022. Fusion of machine learning and MPC under uncertainty: What advances are on the horizon? In: Proc. of the American Control Conference (ACC). pp. 342–357.

Park, J., Martin, R.A., Kelly, J.D., Hedengren, J.D., 2020a. Benchmark temperature microcontroller for process dynamics and control. Comput. Chem. Eng. 135, 106736. http://dx.doi.org/10.1016/j.compchemeng.2020.106736, URL https://linkinghub.elsevier.com/retrieve/pii/S0098135419310129.

Park, J., Martin, R.A., Kelly, J.D., Hedengren, J.D., 2020b. Benchmark temperature microcontroller for process dynamics and control. Comput. Chem. Eng. 135, http://dx.doi.org/10.1016/j.compchemeng.2020.106736.

Park, J., Price, C., Pixton, D., Aghito, M., Nybø, R., Bjørkevoll, K., Hedengren, J.D., 2020c. Model predictive control and estimation of managed pressure drilling using a real-time high fidelity flow model. ISA Trans. 105, 256–268. http://dx.doi.org/10.1016/j.isatra.2020.05.035.

Powell, B.K.M., Machalek, D., Quah, T., 2020. Real-time optimization using reinforcement learning. Comput. Chem. Eng. 143, 107077. http://dx.doi.org/10.1016/j.compchemeng.2020.107077.

Qin, S.J., Chiang, L.H., 2019. Advances and opportunities in machine learning for process data analytics. Comput. Chem. Eng. 126, 465–473. http://dx.doi.org/10.1016/j.compchemeng.2019.04.003.

Qin, H., Guo, X., Tian, Q., Yu, D., Zhang, L., 2021. Recovery of gold from sulfide refractory gold ore: Oxidation roasting pretreatment and gold extraction. Miner. Eng. 164 (January), http://dx.doi.org/10.1016/j.mineng.2021.106822.

Smith, J.C., McCord, T.H., O'Neil, G.R., 1990. Treating refractory gold ores via oxygen-enriched roasting.

Thomas, K., Cole, A., 2016. Roasting Developments – Especially Oxygenated Roasting. Ken Thomas & Murray Pearson, pp. 373–392. http://dx.doi.org/10.1016/b978-0-444-63658-4.00023-2.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Process. Syst. 2017-Dec (Nips), 5999–6009.

Venkatasubramanian, V., 2019. The promise of artificial intelligence in chemical engineering: Is it here, finally? AIChE J. 65 (2), 466–478. http://dx.doi.org/10.1002/aic.16489.

Wang, Y., Velswamy, K., Huang, B., 2017. A long-short term memory recurrent neural network based reinforcement learning controller for office heating ventilation and air conditioning systems. Processes 5 (3), http://dx.doi.org/10.3390/pr5030046.

Wang, S.W., Yu, D.L., Gomm, J.B., Page, G.F., Douglas, S.S., 2006. Adaptive neural network model based predictive control of an internal combustion engine with a new optimization algorithm. Proc. Inst. Mech. Eng. D http://dx.doi.org/10.1243/095440706X72754.

Wells Jr., C., 1948. Oxygen Engiched Atmosphere Roasting (Ph.D. thesis). (1948), Montana School of Mines.

Wong, W.C., Li, J., Wang, X., Chee, E., Li, J., Wang, X., 2018. Recurrent neural network-based model predictive control for continuous pharmaceutical manufacturing. Mathematics 6 (11), http://dx.doi.org/10.3390/math6110242, URL http://arxiv.org/abs/1807.09556.

Wu, Z., Tran, A., Ren, Y.M., Barnes, C.S., Chen, S., Christofides, P.D., 2019. Model predictive control of phthalic anhydride synthesis in a fixed-bed catalytic reactor via machine learning modeling. Chem. Eng. Res. Des. 145, 173–183. http://dx.doi.org/10.1016/j.cherd.2019.02.016.

Yoo, H., Byun, H.E., Han, D., Lee, J.H., 2021. Annual reviews in control reinforcement learning for batch process control : Review and perspectives. Annu. Rev. Control 52 (July), http://dx.doi.org/10.1016/j.arcontrol.2021.10.006.