A Multiple Local Model Learning for Nonlinear and Time-Varying Microwave Heating Process

Tong Liu and Shan Liang
School of Automation
Chongqing University
Chongqing 400044, China
E-mail: tl3n18@soton.ac.uk

Sheng Chen and Chris J. Harris
School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, U.K.
E-mail: sqc@ecs.soton.ac.uk

***Abstract*—This paper proposes a multiple local model learning approach for nonlinear and nonstationary microwave heating process (MHP). The proposed local learning framework performs model adaption at two levels: (1) adaptation of the local linear model set, which adaptively partitions the process's data into multiple process states, each fitted with a local linear model; (2) online adaptation of model prediction, which selects a subset of candidate local linear models and linearly combines them to produce the model prediction. Adaptive process state partition and fitting a new local linear model to the newly emerging process state is based on statistical hypothesis testing, and the optimal combining coefficients of the selected subset linear models are obtained by minimizing the mean square error with the constraint that the sum of these coefficients is unity. A case study involving a real-world industrial MHP is used to demonstrate the superior performance of the proposed multiple local model learning approach, in terms of online modeling accuracy and computational efficiency.**

I. Introduction

Microwave heating technology has found wide-ranging applications in industry due to its many advantages over conventional heating methods, which include selective and volumetric heating, rapid heat transfer and pollution-free environment [1]. However, a major drawback associated with microwave heating is the temperature runaway, caused by properties of material and the inner electromagnetic field distribution [2], which may lead to unwanted combustion and destruction in industrial processes. To improve the safety of microwave heating technology in industrial applications, an accurate model is required for the purpose of temperature prediction and control. From the underlying physics, microwave heating process (MHP) can be modeled by several partial differential equations (PDEs) [3]. These PDEs describe the characteristics of thermodynamics and model the conversion of microwave energy, which are highly complex. Although numerical techniques can be adopted to solve these PDEs, they impose heavy computational burden. Furthermore, the model so obtained is very difficult to be adopted in online control of the MHP. Modeling MHP from data offers a practical alternative.

For industrial processes exhibiting both nonlinear and time-varying characteristics, batch global nonlinear modeling approaches, such as [4]–[6], cannot be applied. Adaptive global nonlinear modeling of nonstationary processes is a challenging task, since both the model parameter values and the model structure must be adapted sufficiently fast in order to timely capture the changing characteristics of the underlying process. However, most of the existing adaptive nonlinear modeling approaches do not perform online model structure updating and they only use the recursive least squares (RLS) algorithm to adapt the model parameter values [7]–[10]. In particular, the extreme learning machine (ELM) for single-hidden-layer neural networks places sufficiently dense number of fixed nodes in the input space and only sequentially updates the model weights using the RLS algorithm. Because the size of the nonlinear model has to be very large for ELM, online adaptation of the model weights is computationally costly and, moreover, it takes time to sufficiently change the model weights to match the changing nonlinear characteristics of the underlying process. Therefore, the online sequential ELM (OS-ELM) approach only works well for relatively slow time varying nonlinear industrial processes.

An alternative to global nonlinear modeling is to adopt the multiple local models, which are capable of capture severe nonlinearity too [11]–[13]. Based on this principle, a multiple local modeling framework is proposed in [14], [15] for nonlinear and nonstationary processes, which comprises a set of radial basis function (RBF) sub-models. Each local RBF model tracks the incoming data independently by updating its weights online, and a subset of these local RBF models are selected to produce the output by a linear combiner of the selected sub-models. However, the model structures of the candidate local RBF models are fixed during the initial training, and they do not change during online operation. In a sense, this multiple local modeling framework is similar to the OS-ELM, and suffers from the same drawback. Specifically, the performance of this approach depends on the coverage of the initially fixed candidate sub-models. The difference with the OS-ELM is basically that the ELM approach employs a large number of hidden nodes to cover the entire model input space, while this multiple local modeling approach 'partitions' the model input space into multiple 'regions', each covered

T. Liu would like to thank the sponsorship of Chinese Scholarship Council for funding his research at University of Southampton. This work was supported in part by the National Natural Science Foundation of China under Grant 61771077.

by a local model. However, during the online operation of a time-varying industrial process, the process dynamics can vary significantly and the process may enter an operating region which is completely outside the initial modeling space, which will degrade the performance of both this adaptive local modeling approach and the OS-ELM.

In order to accurately model nonlinear and nonstationary processes, a multiple local model approach must be able to adaptively generate a new local model timely and efficiently for the newly emerging operating environment. In the online soft sensor design, this capability has been demonstrated to be vital to achieve excellent performance [16], [17]. This motivates our work. In this paper, we propose a multiple local model learning approach for nonlinear and time-varying industrial processes, in which the set of local linear models are self adapted to capture the newly emerging process state, and the prediction of the process output is also adapted based on an optimally selected ensemble of subset linear local models. Similar to [16], [17], which consider a different application of soft sensor design, our proposed multiple local model learning approach performs the model adaptation at two levels. At the level of local model development, a newly emerging process state in the incoming data is automatically identified and a new local linear model is fitted to this newly emerged process state. At the level of modeling update or online prediction, a subset of candidate local linear models are optimally selected and the prediction of the process output is computed as an optimal linear combiner of the selected subset local linear models. A case study involving MHP demonstrates the effectiveness of our multiple local model learning approach, in terms of online prediction accuracy and computational efficiency.

II. Proposed multiple local model learning

A. Adaptation of local linear models

Given the data sample set $\{\boldsymbol{x}(t), y(t)\}_{t=1}^{N}$, where $\boldsymbol{x}(t) \in \mathbb{R}^{m}$ and $y(t) \in \mathbb{R}$ are the system's input vector and output, respectively, the task is to construct the local linear models $\{f_l\}_{l=1}^{L}$ that are valid in their corresponding process states represented by their respective sub-datasets $\{\boldsymbol{X}_l, \boldsymbol{y}_l\}_{l=1}^{L}$.

Without loss of generality, let a data window $\mathcal{W}_{ini} = \big\{\boldsymbol{X}_{ini} \in \mathbb{R}^{W\times m}, \boldsymbol{y}_{ini} \in \mathbb{R}^{W}\big\}$ with W consecutive time samples $\{\boldsymbol{x}(t), y(t)\}_{t=t_{ini}}^{t_{ini}+W}$ be initially set, and a local linear model f_{ini} is built on it as

$$\widehat{\boldsymbol{y}}_{ini} = f_{ini}\big(\boldsymbol{X}_{ini}\big) = \boldsymbol{\Phi}\boldsymbol{\beta} \tag{1}$$

where $\boldsymbol{\Phi} = \big[\boldsymbol{1}_W \ \boldsymbol{X}_{ini}\big] \in \mathbb{R}^{W\times(1+m)}$ and $\boldsymbol{1}_W$ denotes the W-dimensional vector whose elements are all one, while the model parameter vector $\boldsymbol{\beta} \in \mathbb{R}^{(1+m)}$ is solved by the least square (LS) algorithm as

$$\boldsymbol{\beta} = \big(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\big)^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y}_{ini}. \tag{2}$$

The predicted error or residual vector of this local model is

$$\boldsymbol{e}_{ini} = \boldsymbol{y}_{ini} - f_{ini}\big(\boldsymbol{X}_{ini}\big) \in \mathbb{R}^{W}. \tag{3}$$

After an initial local model f_{ini} is built, a shifted window $\mathcal{W}_{sft} = \big\{\boldsymbol{X}_{sft}, \boldsymbol{y}_{sft}\big\}$ is sequentially obtained by moving the window one step ahead, that is, $\mathcal{W}_{sft}$ contains the samples $\{\boldsymbol{x}(t), y(t)\}_{t=t_{ini}+1}^{t_{ini}+1+W}$. If the two local regions $\mathcal{W}_{ini}$ and $\mathcal{W}_{sft}$ are not significantly different, it can be considered that the process data within $\mathcal{W}_{sft}$ follow the same distribution as in $\mathcal{W}_{ini}$ and the window is continued to be shifted forward. Otherwise, $\mathcal{W}_{sft}$ is considered to represent a new operating mode different from the previous mode, and a new local linear model f_{new} should be developed based on $\mathcal{W}_{sft}$. Let the estimation error vector produced by f_{ini} on $\mathcal{W}_{sft}$ be

$$\boldsymbol{e}_{sft} = \boldsymbol{y}_{sft} - f_{ini}\big(\boldsymbol{X}_{sft}\big). \tag{4}$$

Whether $\mathcal{W}_{ini}$ and $\mathcal{W}_{sft}$ are similar or not can then be turned into the equivalent testing that tests whether $\boldsymbol{e}_{ini}$ and $\boldsymbol{e}_{sft}$ are significantly different or not. Since f_{ini} is a linear model, $\boldsymbol{e}_{ini}$ and $\boldsymbol{e}_{sft}$ are considered not significantly different when both their means, μ_{ini} and μ_{sft}, and variances, σ_{ini}^2 and σ_{sft}^2, are the same. Therefore, the two null hypotheses can be set to

$$H_0^{\mu} : \mu_{ini} = \mu_{sft}, \tag{5}$$

$$H_0^{\sigma^2} : \sigma_{sft}^2 = \sigma_{ini}^2. \tag{6}$$

The mean μ_{ini} and variance σ_{ini}^2 are estimated based on $\boldsymbol{e}_{ini}$, while μ_{sft} and σ_{sft}^2 are estimated based on $\boldsymbol{e}_{sft}$. Since f_{ini} is an unbiased estimator, we have $\mu_{ini} = 0$ and $\sigma_{ini}^2 = \frac{1}{W-1}\boldsymbol{e}_{ini}^{\mathrm{T}}\boldsymbol{e}_{ini}$. Assuming that $\boldsymbol{e}_{ini}$ and $\boldsymbol{e}_{sft}$ follow normal distribution, the T and χ^2 statistics are constructed as

$$T = \sqrt{W}\big(\mu_{sft} - \mu_{ini}\big)/\sigma_{sft}, \tag{7}$$

$$\chi^2 = (W-1)\sigma_{sft}^2/\sigma_{ini}^2. \tag{8}$$

According to the statistical theory, if the hypotheses H_0^{μ} and $H_0^{\sigma^2}$ are both valid, the T statistic (7) and χ^2 statistic (8) follow the t distribution and χ^2 distribution with the degree of freedom $W-1$, respectively. Thus, the t-test and χ^2-test can be utilized to test the above two hypotheses. Specifically, the conditions of accepting H_0^{μ} and $H_0^{\sigma^2}$ are

$$|T| < \lambda_t \text{ and } \chi^2 < \lambda_{\chi}, \tag{9}$$

where λ_t is the threshold of the T statistic for the given significance level α_t which satisfies $\Pr\{|T| < \lambda_t\} = 1 - \alpha_t$, while λ_{χ} is the threshold of the χ^2 statistic for the given significance level α_{χ}, which satisfies $\Pr\{\chi^2 < \lambda_{\chi}\} = 1 - \alpha_{\chi}$.

Let the local model set contain $L > 1$ independent local linear models $\{f_l\}_{l=1}^{L}$, and $f_{ini} = f_L$. When one or both conditions of (9) are violated, $\mathcal{W}_{ini}$ and $\mathcal{W}_{sft}$ are significantly different, and the new local linear model $f_{new} = f_{sft}$ is identified, which is different from f_L. We need to test whether f_{new} is different from the other models f_l for $1 \le l \le L-1$. This task can also be fulfilled based on the similar statistic hypothesis testing. Let the predicted errors of $\big\{\boldsymbol{X}_{sft}, \boldsymbol{y}_{sft}\big\}$ based on f_{new} and f_l be defined respectively by

$$\boldsymbol{e}_{new} = \boldsymbol{y}_{sft} - f_{new}\big(\boldsymbol{X}_{sft}\big), \tag{10}$$

$$\boldsymbol{e}_{l} = \boldsymbol{y}_{sft} - f_{l}\big(\boldsymbol{X}_{sft}\big), \, 1 \le l \le L-1. \tag{11}$$

With the assumption that $\boldsymbol{e}_{new}$ and $\boldsymbol{e}_l$ follow normal distribution, the T and χ^2 statistics are constructed according to

$$T_l = \sqrt{W}\big(\mu_l - \mu_{new}\big)/\sigma_l, \tag{12}$$

$$\chi_l^2 = (W-1)\sigma_l^2/\sigma_{new}^2, \tag{13}$$

where μ_{new} and σ_{new}^2 are the mean and variance of $\boldsymbol{e}_{new}$, which can be estimated using $\boldsymbol{e}_{new}$, while μ_l and σ_l^2 are the mean and variance of $\boldsymbol{e}_l$, which can be estimated in the same way. Based on the statistical theory, if the null hypotheses

$$H_l^{\mu}: \ \mu_l = \mu_{new}, \tag{14}$$

$$H_l^{\sigma^2}: \ \sigma_l^2 = \sigma_{new}^2, \tag{15}$$

are both valid, the T_l statistic in (12) and χ_l^2 statistic in (13) follow the t distribution and χ^2 distribution with the degree of freedom $W-1$, respectively. Therefore, if there exist an $l \in \{1, 2 \ldots L-1\}$ such that

$$|T_l| < \lambda_t \text{ and } \chi_l^2 < \lambda_{\chi}, \tag{16}$$

the hypotheses (14) and (15) are both valid, and $\boldsymbol{e}_{new}$ and $\boldsymbol{e}_l$ are regarded to be identical. Consequently, f_{new} and f_l are the same model, and one of them should be removed. Since f_l is 'older' than f_{new}, we keep the local model f_{new} and delete f_l. On the other hand, if one or both conditions are violated $\forall l \in \{1, 2 \ldots L-1\}$, f_{new} is different from f_l for $1 \le l \le L$. Thus, we have identified a new process state, and we add f_{new} to the local model set by setting $L = L+1$ and $f_L = f_{new}$.

The proposed adaptive local model set development procedure is summarized in Algorithm 1. A small widow size W may lead to large number of local models, which will increase online operating time, but it may result in better nonstationary adaptation capability. A large W has the opposite efforts. The significance levels in the statistical testings are typically set to $\alpha_t = 0.05$ and $\alpha_{\chi} = 0.05$.

Remark 1: This local learning strategy can operate both offline and online.

B. Adaptation of model prediction

After the online operation at time sample t, Algorithm 1 produces the local model set of $\{f_l\}_{l=1}^{L}$. At the next time sample of $t_{next} = t+1$, the task of online modeling update is to produce the model prediction $\widehat{y}(t_{next})$ for the process's true output $y(t_{next})$, given the process input $\boldsymbol{x}(t_{next})$ and the available local model set $\{f_l\}_{l=1}^{L}$. We adopt a selective ensemble of local linear models from $\{f_l\}_{l=1}^{L}$ based on the $p(>1)$ latest labeled data $\{\boldsymbol{x}(t-i), y(t-i)\}_{i=0}^{p-1}$.

Let $\boldsymbol{e}_l(t) = [e_l(t) \ e_l(t-1) \cdots e_l(t-p+1)]^{\mathrm{T}}$ be the modeling error vector of the lth local linear model f_l on the available data set $\{\boldsymbol{x}(t-i), y(t-i)\}_{i=0}^{p-1}$, which is given by

$$e_l(t-i) = y(t-i) - f_l(\boldsymbol{x}(t-i)), \ 0 \le i \le p-1. \tag{17}$$

The performance metric of the lth local model is defined as

$$J_l(t) = \left\| \boldsymbol{e}_l(t) \right\|^2. \tag{18}$$

Algorithm 1 Adaptive local model set development

```
Initialization
Collect $\mathcal{W}_{ini}$ with $W$ consecutive samples from historical data, and construct the LS linear model $f_{ini}$ on $\mathcal{W}_{ini}$.
Calculate $\boldsymbol{e}_{ini}$, and estimate $\mu_{ini}$ and $\sigma_{ini}^2$.
Set $L = 1$, $\{\mathcal{W}_L, f_L\} = \{\mathcal{W}_{ini}, f_{ini}\}$ and $\mathcal{W}_{sft} = \mathcal{W}_L$.
Step 1: New local model detection
When a new data sample is available, shift $\mathcal{W}_{sft}$ one sample ahead and construct $f_{sft}$ on $\mathcal{W}_{sft}$.
Calculate $\boldsymbol{e}_{sft}$, and estimate $\mu_{sft}$ and $\sigma_{sft}^2$.
Construct $T$ and $\chi^2$ statistics using (7) and (8).
If both conditions of (9) are satisfied
  Go to Step 1.
End if
Set $\mathcal{W}_{new} = \mathcal{W}_{sft}$, $f_{new} = f_{sft}$, $\boldsymbol{e}_{new} = \boldsymbol{e}_{sft}$, as well as $\mu_{new} = \mu_{sft}$ and $\sigma_{new}^2 = \sigma_{sft}^2$.
Step 2: Redundant local model deletion
For $l = 1, 2, \ldots, L-1$
  Compute $\boldsymbol{e}_l$, and estimate $\mu_l$ and $\sigma_l^2$.
  Construct $T_l$ and $\chi_l^2$ statistics using (12) and (13).
  If both conditions of (16) are satisfied
    Delete $f_l$, set $f_i = f_{i+1}$ for $i = l, l+1, \cdots, L-1$, set $L = L-1$, then go to Step 3.
  End if
End for
Step 3: Add new local model
Set $L = L+1$, $\mathcal{W}_L = \mathcal{W}_{new}$ and $f_L = f_{new}$.
Return to Step 1.
```

By defining

$$J_{l_{\max}}(t) = \max_{1 \le l \le L} J_l(t), \tag{19}$$

we can normalize the performance metrics of (18) to

$$\bar{J}_l(t) = \frac{J_l(t)}{J_{l_{\max}}(t)}, \ 1 \le l \le L. \tag{20}$$

Obviously, $\bar{J}_l(t) \in (0, \ 1]$. Clearly, the best local model, whose index $l_1 = l_{\min}$ is given by

$$l_{\min} = \arg \min_{1 \le l \le L} \bar{J}_l(t), \tag{21}$$

should be selected. Moreover, other local models whose performance metrics (20) are below a given threshold $0 < \varepsilon \le 1$ are also selected. Note that if $\varepsilon = 1$, all the L local models are selected, while if $\varepsilon \le J_{l_{\min}}(t)$, only the best local model f_{l_1} is selected. Assume that $M(\ge 1)$ local models are selected at time t, and the indexes of the selected local models are represented by the index set Γ as

$$\Gamma = \{l_1, l_m | 2 \le m \le M, J_{l_m}(t) \le \varepsilon, 2 \le l_m \le L\}. \tag{22}$$

This selection procedure yields the M local model outputs

$$\widehat{y}_{l_m}(t-i) = f_{l_m}(\boldsymbol{x}(t-i)), \ 1 \le m \le M, \tag{23}$$

for $0 \le i \le p-1$. The estimate $\widehat{y}(t-i)$ of the process output

$y(t-i)$ is given as the weighted summation of the M selected subset models, which is computed by

$$\widehat{y}(t-i) = \sum_{m=1}^{M} \theta_m(t)\widehat{y}_{l_m}(t-i),\ 0 \le i \le p-1, \tag{24}$$

where nonnegative $\theta_m(t)$ is the combining coefficient for the mth selected local model, and the combining coefficients must satisfy the constraint

$$\sum_{m=1}^{M} \theta_m(t) =1. \tag{25}$$

The estimation errors

$$e(t-i) =y(t-i) - \widehat{y}(t-i),\ 0 \le i \le p-1, \tag{26}$$

are utilized to determine the combining coefficients.

Specifically, the optimal combining coefficients can be obtained by minimizing the LS cost function

$$V(t) =\frac{1}{2}\sum_{i=0}^{p-1} e^2(t-i), \tag{27}$$

subject to the constraint (25). Because of $\sum_{m=1}^{M} \theta_m(t) = 1$,

$$\begin{aligned} V(t) =&\frac{1}{2}\sum_{i=0}^{p-1}\left(y(t-i) - \sum_{m=1}^{M}\theta_m(t)\widehat{y}_{l_m}(t-i)\right)^2 \\ =&\frac{1}{2}\sum_{i=0}^{p-1}\left(\sum_{m=1}^{M}\theta_m(t)y(t-i) - \sum_{m=1}^{M}\theta_m(t)\widehat{y}_{l_m}(t-i)\right)^2 \\ =&\frac{1}{2}\sum_{i=0}^{p-1}\left(\sum_{m=1}^{M}\theta_m(t)e_{l_m}(t-i)\right)^2 \\ =&\frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}(t)\bar{\boldsymbol{E}}(t)\boldsymbol{\theta}(t), \end{aligned} \tag{28}$$

where $\boldsymbol{\theta}(t) = \big[\theta_1(t)\cdots\boldsymbol{\theta}_M(t)\big]^{\mathrm{T}}$ and $\bar{\boldsymbol{E}}(t)$ is the estimated error covariance matrix which is given as

$$\bar{\boldsymbol{E}}(t) = \sum_{i=0}^{p-1}\begin{bmatrix} e_{l_1}^2(t-i) & \cdots & e_{l_1}(t-i)e_{l_M}(t-i) \\ \vdots & \ddots & \vdots \\ e_{l_1}(t-i)e_{L_M}(t-i) & \cdots & e_{l_M}^2(t-i)\end{bmatrix}. \tag{29}$$

The problem of determining the optimal $\boldsymbol{\theta}(t)$ can then be formulated as the following optimization

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & \tfrac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}(t)\bar{\boldsymbol{E}}(t)\boldsymbol{\theta}(t), \\ \text{s.t.} \quad & \sum_{m=1}^{M}\theta_m(t) = 1. \end{aligned} \tag{30}$$

The Lagrangian function for the optimization (30) is given by

$$L\big(\boldsymbol{\theta}(t);\gamma\big) = \frac{1}{2}\boldsymbol{\theta}^{\mathrm{T}}(t)\bar{\boldsymbol{E}}(t)\boldsymbol{\theta}(t) + \gamma\big(\mathbf{1}_M^{\mathrm{T}}\boldsymbol{\theta}(t) - 1\big), \tag{31}$$

where $\gamma > 0$ is Lagrange multiplier, and $\mathbf{1}_M = [1\cdots1]^{\mathrm{T}} \in \mathbb{R}^M$. Letting $\frac{\partial}{\partial\boldsymbol{\theta}(t)}L = \mathbf{0}_M$ yields

$$\bar{\boldsymbol{E}}(t)\boldsymbol{\theta}(t) + \gamma\mathbf{1}_M = \mathbf{0}_M, \tag{32}$$

where $\mathbf{0}_M = [0\cdots0]^{\mathrm{T}} \in \mathbb{R}^M$. This suggests that the optimal combining vector $\widehat{\boldsymbol{\theta}}$ can be obtained as follows. First, calculate

$$\widetilde{\boldsymbol{\theta}}(t) = \bar{\boldsymbol{E}}^{-1}(t)\mathbf{1}_M, \tag{33}$$

which is followed by the normalization

$$\widehat{\theta}_m(t) =\frac{1}{\sum_{j=1}^{M}\widetilde{\theta}_j(t)}\widetilde{\theta}_m(t),\ 1 \le m \le M. \tag{34}$$

The prediction $\widehat{y}(t_{next})$ for the process's true output $y(t_{next})$ is produced as the selected ensemble

$$\widehat{y}(t_{next}) = \sum_{m=1}^{M}\widehat{\theta}_m(t)f_{l_m}\big(\boldsymbol{x}(t_{next})\big) \tag{35}$$

Algorithm 2 Online prediction and adaptive modeling

```
Initialization
At the beginning of online operation, the local model set $\{\mathcal{W}_l, f_l\}_{l=1}^{L}$ has been constructed.
Set $\{\mathcal{W}_L, f_L\} = \{\mathcal{W}_{ini}, f_{ini}\}$ and $\mathcal{W}_{sft} = \mathcal{W}_L$.
Step 1: Online prediction
Give input $\boldsymbol{x}(t_{next})$ at new sample time $t_{next} = t + 1$.
Calculate the performance metrics $\bar{J}_l(t)$ using (20) for $1 \le l \le L$ on past $p$ data points.
Select the subset models with the index set $\Gamma$ of (22).
Calculate the error covariance matrix $\bar{\boldsymbol{E}}(t)$ using (29).
Calculate the optimal combining coefficients $\widehat{\boldsymbol{\theta}}(t)$ using (33) and (34).
Predict true process output $y(t_{next})$ with the selective ensemble (35).
Carry out other unrelated online operations.
Step 2: Online model adaptation
When the observation $y(t_{next})$ is available, add $\{\boldsymbol{x}(t_{next}), y(t_{next})\}$ to the dataset with $t = t + 1$.
Shift $\mathcal{W}_{sft}$ one sample ahead, and perform relavent local model set adaptation.
Set $t_{next} = t_{next} + 1$, and go to Step 1.
```

Algorithm 2 summarizes the online prediction and adaptive modeling operations. The choice of p trades off the computational complexity and the robustness against noise.

III. Microwave Heating Process Case Study

A. Process description

MHP is a complex thermal process with nonlinear dynamics and nonstationary characteristics. Unlike conventional heat transfer and heat radiation, microwave heating not only involves thermal dynamic variation but also coupled with conversion of microwave energy. Temperature of heated material is a crucial measurement during MHP, as thermal runaway often occurs due to the time-varying physicochemical properties of material. With the increase of the material temperature, its dielectric loss increases dramatically, which conversely poses a positive feedback to temperature increase [18]. Therefore, accurate online temperature estimation is vital to detect thermal runaway in advance.

A real-world distributed microwave heating system [19] is used in this case study, which consists of five microwave generators and waveguides. Microwave generated by each microwave generator is transmitted through the corresponding waveguide, fed into the cavity and absorbed by the heated material. The material is continuously transported through cavity by the conveyor belt, whose speed can be adjusted by a motor driver. Three fiber optical sensors (FOSs), denoted as FOS1 to FOS3, are placed at three different locations to online record multiple-points of temperature. During the real-time operation of this MHP, the control center receives the measured temperature values from the FOSs, and sends control commends, including the five microwave powers $u_{p_i}(t)$, $1 \le i \le 5$, for the five microwave generators as well as the conveyor speed $v(t)$ to the cavity. Thus, the control inputs to this MHP are given by

$$\boldsymbol{u}(t) = \big[u_{p_1}(t)\ u_{p_2}(t)\ u_{p_3}(t)\ u_{p_4}(t)\ u_{p_5}(t)\ v(t)\big]^{\mathrm{T}}. \tag{36}$$

Each FOS measures the temperature, which is the MHP's output $y(t)$ at the FOS's location. For notational simplification and without causing ambiguity, we have dropped the index for the FOS from the output $y(t)$. Because of near instantaneous response of MHP, the temperature $y(t)$ can be adequately represented by [18]–[20]

$$y(t) = f_{\mathrm{nl-ns}}(\boldsymbol{x}(t);t), \tag{37}$$

where $f_{\mathrm{nl-ns}}(\cdot;t)$ represents the unknown nonlinear and time-varying system mapping with the input

$$\boldsymbol{x}(t) = \big[y(t-1)\ \boldsymbol{u}^{\mathrm{T}}(t)\big]^{\mathrm{T}} \in \mathbb{R}^7. \tag{38}$$

From large amount of data collected from this distributed microwave heating system [20], we use three datasets from the three FOSs, and each data set contains 3,000 data samples. We first normalize the five microwave power inputs and the temperature measurement according to

$$\bar{u}_{p_i}(t) = \frac{u_{p_i}(t)}{1000},\ 1 \le i \le 5, \tag{39}$$

$$\bar{y}(t) = \frac{y(t) - y_{\min}}{y_{\max} - y_{\min}}, \tag{40}$$

where $y_{\min}$ and $y_{\max}$ are the minimum and maximum temperatures for each FOS, respectively. For each FOS's dataset, we use the first 1,000 samples for model training, and the last 2,000 samples for online prediction and adaptive modeling.

B. Experimental results

The performance of the proposed multiple local model learning approach are compared with those of the SO-ELM with sigmoid hidden nodes (SO-ELM (sigmoid)) and the SO-ELM with RBF hidden nodes (SO-ELM (RBF)) [8]–[10]. For the dataset of each FOS, the 1,000 training samples are employed for the initial model training, and the 2,000 testing samples are used for online prediction and adaptive modeling. Note that our proposed multiple local model learning method does not really need a large number of training samples, but the OS-ELM needs such a large number of training samples, as the ELM model must contain a large number of hidden nodes. Two performance indexes, the root mean square error (RMSE)

$$\mathrm{RMSE}(t) = \sqrt{\frac{1}{t}\sum_{i=1}^{t}\big(y(i) - \widehat{y}(i)\big)^2}, \tag{41}$$

and the mean absolute error (MAE)

$$\mathrm{MAE}(t) = \frac{1}{t}\sum_{i=1}^{t}\big|y(i) - \widehat{y}(i)\big|, \tag{42}$$

are used to evaluate the online prediction performance, where $\widehat{y}(i)$ denotes the model prediction for $y(i)$.

1) Impacts of algorithmic parameters: For our proposed method, we first investigate the impacts of the window size W for adaptive local modeling, the number of the latest data samples p and the threshold ε for selective ensemble.

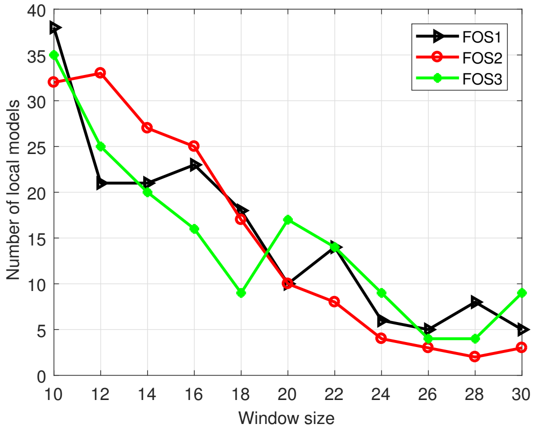
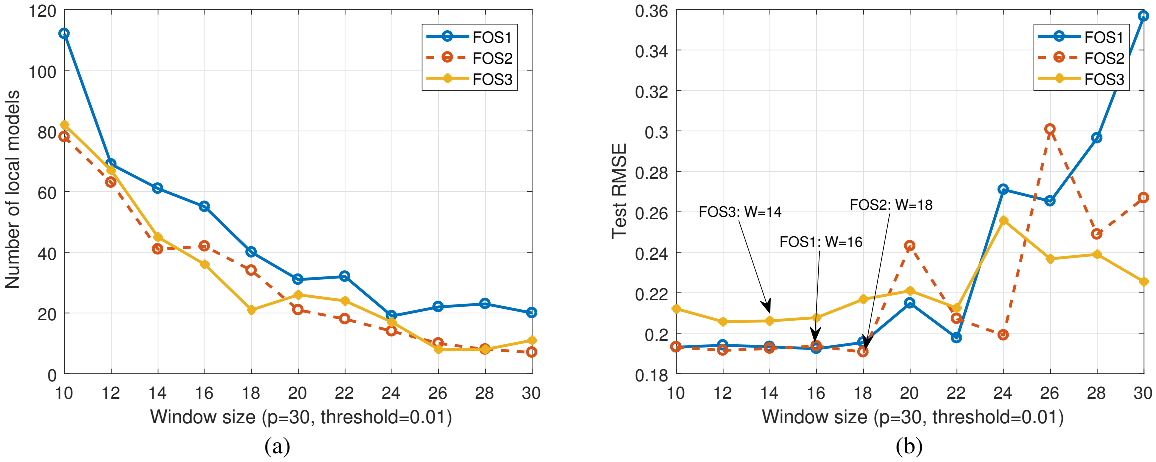


Fig. 1. Influence of window size W on number of local models obtained for three training data sets.

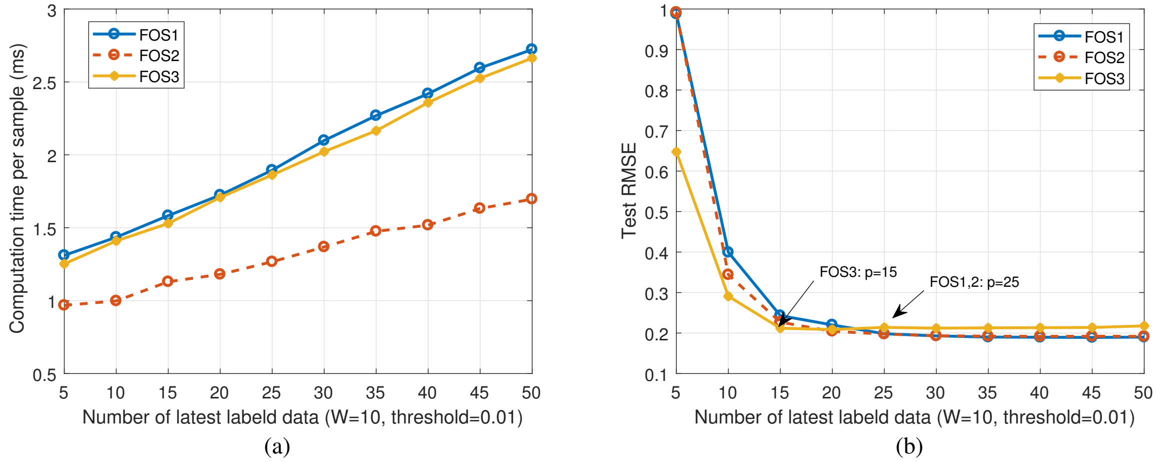
We apply Algorithm 1 to the training data sets of the three FOSs. Fig. 1 shows the numbers of local linear models obtained as the functions of the window size W. As expected, small W leads to large number of local models identified, and vice versa. For eaxmple, for FOS1, 38 local linear models are constructed given $W = 10$ but only 5 local linear models are identified given $W = 30$. With the initial local model sets identified in training, we then apply Algorithm 2 to the three testing data sets.

With the parameters of selective ensemble set to $p = 30$ and $\varepsilon = 0.01$, Fig. 2 (a) and (b) depict the number of total local models and the prediction RMSE as the functions of window size W, respectively. As expected, small W results in better prediction accuracy but leads to large local model set obtained which has adverse effort on online computation complexity. It can be seen from Fig. 2 (b) that $W \le 18$ is appropriate. More specifically, $W = 16$ for FOS1, $W = 18$ for FOS2, and $W = 14$ for FOS3 are appropriate, in terms of achievable prediction accuracy. Compared Fig. 2 (a) with Fig. 1, it can be seen that the proposed learning approach have identified



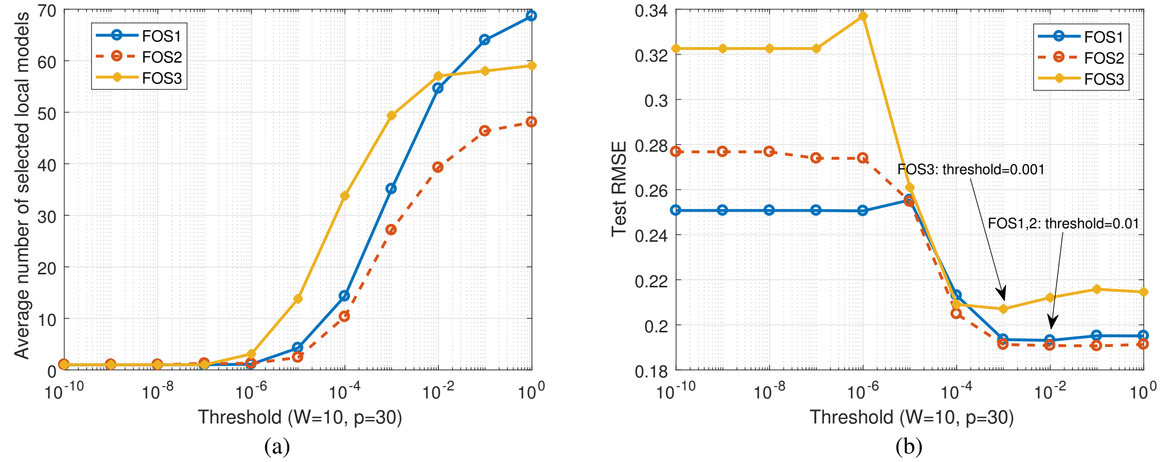
(a) (b)

Fig. 2. Influence of window size W on online prediction and adaptive modeling given $p = 30$ and $\varepsilon = 0.01$ for three testing data sets: (a) number of total local models and (b) online prediction accuracy, both obtained after online adaptation.



(a) (b)

Fig. 3. Influence of number of latest labeled data samples p on online prediction and adaptive modeling given $W = 10$ and $\varepsilon = 0.01$ for three testing data sets: (a) online computation time per sample, and (b) online prediction accuracy obtained after online adaptation.



(a) (b)

Fig. 4. Influence of threshold ε on online prediction and adaptive modeling given $W = 10$ and $p = 30$ for three testing data sets: (a) Average selected ensemble size, and (b) online prediction accuracy obtained after online adaptation.

the new process's states with the associated new local linear models which occur during the online operation of the process.

Impacts of the number of of latest labeled data samples p for selective ensemble on the achievable performance of online prediction and adaptive modeling are investigated in Fig. 3, given $W = 10$ and $\varepsilon = 0.01$. As expected, Fig. 3 (a) shows that online computation complexity increases linearly with p. It can be seen from Fig. 3 (b) that the prediction RMSEs reach

TABLE I

Comparison of online prediction and adaptive modeling performance for the SO-ELM and the proposed multiple local model learning approach

Sensor	Model	RMSE	MAE	Computation time per sample (ms)	Models/Nodes	
					Initial	Final
FOS1	SO-ELM (sigmoid)	3.0724	0.2014	0.18	100	100
		0.2122	0.1236	**1.38**	300	**300**
		0.2387	0.1652	7.68	500	500
	SO-ELM (RBF)	0.7598	0.2437	0.57	100	100
		0.6477	0.1642	3.04	300	300
		0.3744	0.2201	12.05	500	500
	Proposed ($W = 16$, $\varepsilon = 0.01$, $p = 25$)	**0.1978**	0.1346	**0.75**	23	**55**
FOS2	SO-ELM (sigmoid)	2.9911	0.2058	0.18	100	100
		0.2520	0.1427	1.39	300	300
		0.2370	0.1427	**5.98**	500	**500**
	SO-ELM (RBF)	1.3952	0.3476	0.56	100	100
		0.9209	0.2006	2.98	300	300
		0.4353	0.1676	10.69	500	500
	Proposed ($W = 18$, $\varepsilon = 0.01$, $p = 25$)	**0.1953**	0.1411	**0.66**	16	**34**
FOS3	SO-ELM (sigmoid)	2.2194	0.2233	0.18	100	100
		0.2409	0.1657	1.38	300	300
		0.2316	0.1653	**6.33**	500	**500**
	SO-ELM (RBF)	1.4670	0.4087	0.55	100	100
		0.7763	0.2135	3.25	300	300
		0.3487	0.1932	12.84	500	500
	Proposed ($W = 14$, $\varepsilon = 0.001$, $p = 15$)	**0.2019**	0.1476	**0.74**	20	**45**

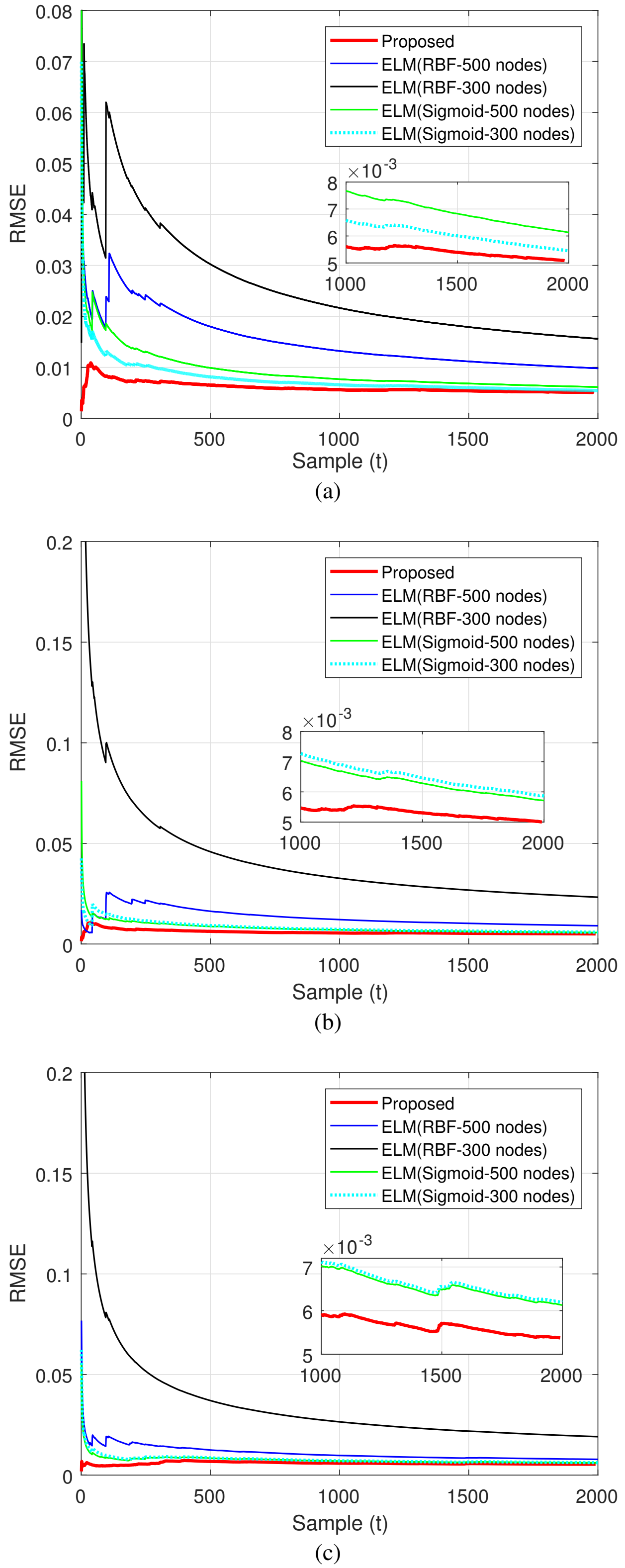
the minimum values when $p \geq 25$ for FOS1 and FOS2 as well as when $p \geq 15$ for FOS3.

Fig. 4 shows how the threshold ε impacts on online prediction and adaptive modeling, given $W = 10$ and $p = 30$. Specifically, observe from Fig. 4 (a) that when ε is smaller than certain value, only one (the best) local linear model is selected. When ε is larger than this value, the average size of selected ensemble increases with ε. Also when $\varepsilon = 1$, all the local models are selected and the size of selected ensemble reaches the maximum value. Fig. 4 (b) indicates that the best prediction RMSEs are achieved with $\varepsilon = 0.01$ for FOS1 and FOS2 as well as with $\varepsilon = 0.001$ for FOS3.

2) Test performance comparison: We now compare the online prediction and adaptive modeling performance of the proposed multiple local model learning approach with those of the SO-ELM (sigmoid) and SO-ELM (RBF) in Table I. It can be seen that our proposed method not only achieves a better online prediction accuracy but also imposes significantly lower average online computation complexity per sample, compared with the SO-ELM. Specifically, for FOS1, our proposed method attains the final prediction RMSE of 0.1978 at the cost of 0.75 ms computation time per sample, while the best SO-ELM with 300 sigmoid hidden nodes achieves the final prediction RMSE of 0.2122 at the cost of 1.38 ms computation time per sample. For FOS2, our method achieves the final prediction RMSE of 0.1953 with the complexity of 0.66 ms computation time per sample, compared with the final prediction RMSE of 0.2370 and the complexity of 5.98 ms computation time per sample attained by the best SO-ELM having 500 sigmoid hidden nodes. Similar performance gains of our proposed method over the SO-ELM can be observed for FOS3, in terms of prediction accuracy and online computation complexity. Finally, Fig. 5 (a) to (c) compare the online prediction RMSE(t) performance of our proposed method and the SO-ELM for the three FOSs. The results of Fig. 5 further demonstrate that our multiple local model learning approach can much better track the nonlinear and fast time-varying characteristics of the underlying system.

IV. Conclusions

Industrial microwave heating processes are highly nonlinear and nonstationary. In this paper, a novel online modeling approach has been proposed. Our proposed multiple local model learning approach automatically identifies the newly emerging process state during online operation and fits a local linear model to the newly identified process state. Adaptive modeling is achieved by a selective ensemble strategy which selects a number of best local linear models from the local model set and optimally combines them to produce the online prediction. In the application to a real-world distributed microwave heating system, our proposed multiple local model learning approach has been demonstrated to be capable of fast tracking the nonlinear and time-varying characteristics of the underlying system. In particular, it has been shown that our proposed method not only achieves better online prediction accuracy but also imposes significantly lower online



(a)

(b)

(c)

Fig. 5. Comparison of online prediction performance for the proposed multiple local model learning approach and the SO-ELM: (a) FOS1, (b) FOS2, and (c) FOS3.

computation complexity per sample, compared with the SO-ELM for for nonlinear and nonstationary modeling. Although we derive this adaptive multiple local model learning in the context of industrial microwave heating processes, it is self-evident that our approach is applicable to generic nonlinear and nonstationary systems.

References

[1] S. Chandrasekaran, S. Ramanathan, and T. Basak, “Microwave material processing - a review,” *AIChE Journal*, vol. 58, no. 2, pp. 330–363, Feb. 2012.

[2] C. A. Vriezinga, S. Sánchez-Pedreño, and J. Grasman, “Thermal runaway in microwave heating: a mathematical analysis,” *Applied Mathematical Modelling*, vol. 26, no. 11, pp. 1029–1038, Nov 2002.

[3] J. Zhong, S. Liang, Y. Yuan, and Q. Xiong, “Coupled electromagnetic and heat transfer ODE model for microwave heating with temperature-dependent permittivity,” *IEEE Trans. Microwave Theory & Techniques*, vol. 64, no. 8, pp. 2467–2477, Aug. 2016.

[4] S. Chen, S. A. Billings, and P. M. Grant, “Non-linear systems identification using neural networks,” *Int. J. Control*, vol. 51, no. 6, pp. 1191–1214, 1990.

[5] S. Chen, C. F N. Cowan, and P. M. Grant, “Orthogonal least squares learning algorithm for radial basis function networks,” *IEEE Trans. Neural Networks*, vol. 2, no. 2, pp. 302–309, Mar. 1991.

[6] S. Chen, X. X. Wang, and C. J. Harris, “NARX-based nonlinear system identification using orthogonal least squares basis hunting,” *IEEE Trans. Control Systems Technology*, vol. 16, no. 1, pp. 78–84, Jan. 2008.

[7] S. Chen, “Nonlinear time series modelling and prediction using Gaussian RBF networks with enhanced clustering and RLS learning,” *Electronics Letters*, vol. 31, no. 2, pp. 117–118, 1995.

[8] G. B. Huang, Q. Y. Zhu, and C. K. Siew, “Extreme learning machine: Theory and applications,” *Neurocomputing*, vol. 70, nos. 1-3, pp. 489–501, Dec. 2006.

[9] N. Liang, G. Huang, P. Saratchandran, and N. Sundararajan, “A fast and accurate online sequential learning algorithm for feedforward networks,” *IEEE Trans. Neural Networks*, vol. 17, no. 6, pp. 1411–1423, Nov. 2006.

[10] G.-B. Huang and L. Chen, “Enhanced random search based incremental extreme learning machine,” *Neurocomputing*, vol. 71, nos. 16-18, pp. 3460–3468, Oct. 2008.

[11] S. A. Billings and S. Chen, “Extended model set, global data and threshold model identification of severely non-linear systems,” *Int. J. Control*, vol. 50, no. 5, pp. 1897–1923, 1989.

[12] H. Tong and K. S. Lim, “Threshold autoregression, limit cycles and cyclical data,” *J. Royal Statistical Society*, vol. 42, no. 3, pp. 245–292, 1980.

[13] H. Tong, *Threshold Models in Non-linear Time Series Analysis*. Springer-Verlag: New York, 1983.

[14] H. Chen, Y. Gong, and X. Hong, “A new adaptive multiple modelling approach for non-linear and non-stationary systems,” *Int. J. Systems Science*, vol. 47, no. 9, pp. 2100–2110, 2016.

[15] X. Hong and Y. Gong, “A constrained recursive least squares algorithm for adaptive combination of multiple models,” in *Proc. IJCNN 2015* (Killarney, Ireland), Jul. 11-16, 2015, pp. 1–6.

[16] W. Shao, X. Tian, P. Wang, X. Deng, and S. Chen, “Online soft sensor design using local partial least squares models with adaptive process state partition,” *Chemometrics and Intelligent Laboratory Systems*, vol. 144, pp. 108–121, May 2015.

[17] W. Shao, S. Chen, and C. J. Harris, “Adaptive soft sensor development for multi-output industrial processes based on selective ensemble learning,” *IEEE Access*, vol. 6, pp. 55628–55642, Oct. 2018.

[18] X. Shi, J. Li, Q. Xiong, Y. Wu, and Y. Yuan, “Research of uniformity evaluation model based on entropy clustering in the microwave heating processes,” *Neurocomputing*, vol. 173, pp. 562–572, Jan. 2016.

[19] T. Liu, S. Liang, Q. Xiong, and K. Wang, “Adaptive critic based optimal neurocontrol of a distributed microwave heating system using diagonal recurrent network,” *IEEE Access*, early access.

[20] K. Wang, L. Ma, Q. Xiong, S. Liang, G. Sun, X. Yu, Z. Yao, and T. Liu, “Learning to detect local overheating of the high-power microwave heating process with deep learning,” *IEEE Access*, vol. 6, pp. 10288–10296, Feb. 2018.