Applied Artificial Intelligence
An International Journal

ISSN: 0883-9514 (Print) 1087-6545 (Online) Journal homepage: https://www.tandfonline.com/loi/uaai20

Nonlinear Short-term Prediction of Aluminum Foil Thickness via Global Regressor Combination

Ali Ozturk & Rifat Seherli

To cite this article: Ali Ozturk & Rifat Seherli (2017) Nonlinear Short-term Prediction of Aluminum Foil Thickness via Global Regressor Combination, Applied Artificial Intelligence, 31:7-8, 568-592, DOI: 10.1080/08839514.2017.1412815

To link to this article: https://doi.org/10.1080/08839514.2017.1412815

APPLIED ARTIFICIAL INTELLIGENCE
2017, VOL. 31, NOS. 7–8, 568–592
https://doi.org/10.1080/08839514.2017.1412815

Taylor & Francis
Taylor & Francis Group

Nonlinear Short-term Prediction of Aluminum Foil Thickness via Global Regressor Combination

Ali Ozturk[a] and Rifat Seherli[b]

[a]Computer Engineering Department, KTO Karatay University, Konya, Turkey; [b]Research Department, Panda Aluminium Co., Kazan, Ankara, Turkey

ABSTRACT

In this study, short-term prediction of aluminum foil thickness time-series data recorded during cold-rolling process was investigated. The locally projective nonlinear noise reduction was applied in order to improve the predictability of the time series. The higher-order statistics methods (bispectrum and bicoherence) were used to detect the nonlinearity. The embedding vectors with appropriate embedding dimension and time delay were obtained via the false nearest neighbors and mutual information methods, respectively. The maximum prediction horizon was determined depending on the maximal Lyapunov exponent. For various prediction horizons, the embedding vector and corresponding thickness value pairs were used as the dataset to assess the prediction performance of various machine learning algorithms (i.e., multilayer perceptron neural network, support vector machines with Pearson VII function-based kernel, and radial basis function network). The *n*-step ahead prediction outputs of the machine learning algorithms were globally combined with simple voting in favor of the one having minimum absolute error. The accuracy of our proposed method was compared with nonlinear autoregressive exogenous model for various thickness time-series data using mean absolute percentage error measure.

Introduction

The fundamental principle of cold-rolling process is the tension produced by the coiling and uncoiling motors of the rolling machine. The tension and its regulation are very important factors for maintaining the stability of the desired thickness over the whole surface of the aluminum foil. The elastic deformation values of the rollers and other mechanical systems of the rolling machine change accordingly depending on the tension. If the tension is not properly regulated, rupture may occur on the aluminum foil. This will stop the rolling process, large amount of aluminum metal will go to the scrap, and even rollers will be damaged. Therefore, short-term prediction of the

CONTACT Ali Ozturk ali.ozturk@karatay.edu.tr Computer Engineering Department, KTO Karatay University, Konya, Turkey.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/UAAI.

aluminum foil thickness plays a very important role to control the cold-rolling process. The angular velocity of the uncoiling and coiling motors determines the amount of the tension applied to foil. The angular velocity of the uncoiling motor must decrease and the angular velocity of the coiling motor must increase accordingly to keep the tension constant during cold-rolling process. Depending on the radius of the aluminum roll, the angular velocity of the motors must change. This value is determined by the armature and field currents applied to the motors. The existing programmable logic controller (PLC) system includes one thickness measurement device on the output side of the cold-rolling machine. The thickness value measured by this device is used to regulate the tension applied to the aluminum foil every second in when 5 m of aluminum foil passes through the mills. This instantaneous manner of regulation affects the precision of the thickness due to the mentioned latency. Therefore, short-term prediction is necessary for more effective regulation of the velocity of the rollers. This will improve the effectivity of the overall process by preventing ruptures.

There are many studies in the literature on the prediction of observed time-series data. In Nesreen et al. (2010), several machine learning algorithms and preprocessing methods were compared against monthly M3 time-series competition data for only 1-step ahead prediction. The authors found that multilayer perceptron and Gaussian processes have the best performance. Self-organizing map (SOM) and support vector machine (SVM) models were used to forecast day-ahead electricity prices in Niu, Liu, and Wu (2010). SOM was used to cluster the data automatically to avoid the problem of insufficient training data. Different SVM models were built on the SOM clustered categories. Bontempi, Ben Taieb, and Le Borgne (2013) investigated the machine learning and mutistep time-series forecasting strategies. They divided machine learning approaches into supervised and local learning settings for modeling the time dependencies of the observed data. The local learning approach utilizes either nearest neighbor or lazy learning techniques. Unlike n-step ahead (direct) prediction where actual past observations are used for the prediction, in iterated prediction strategy, the predicted output is fed back as input to the next prediction. In some cases, direct prediction strategy gives better results than iterated method (Sorjamaa et al. 2007). Iterated or direct methods have a common feature that they model a multiinput single-output mapping from historical data. There are also studies for long-term prediction in the literature where multiple-input multiple-output strategy was used (Bao, Xiong, and Hu 2014; Ben Taieb et al. 2012; Mao, Tian, and Yan 2012). However, in our study, we are interested in short-term prediction with higher precision. According to chaos theory, an underlying dynamical process which generates a chaotic time series depends on a certain deterministic nonlinear equation. Although long-term prediction is meaningless due to nonlinearity and sensitive dependency on initial

conditions, very accurate short-term prediction is possible on a chaotic attractor using neighborhood information of embedding vectors. Therefore, until the deterministic causality is lost, accurate short-term prediction can be made from an observed data. The key point here is reconstructing the time-series data in an n-dimensional embedding state space to obtain the trajectories of the chaotic attractor which corresponds to the observed one-dimensional time-series data. Basharat and Shah (2009) applied chaos theory methods to human action modelling and dynamic texture synthesis by using the embedding vectors obtained from the original multivariate time series of image data. In Tran, Yang, and Tan (2009), multistep ahead prediction of the operating conditions of machine was investigated. Regression trees and adaptive neuro fuzzy inference system were compared with direct prediction strategy. The training data were the embedding vectors obtained from the original time series with appropriate time-delay and embedding dimension. In Yang and Duan (2003), the chaotic characteristics of electricity price were investigated and a price forecasting model with recurrent neural network was proposed. In Iokibe and Fujimoto (2001), a local fuzzy reconstruction method of time-series data was proposed for nonlinear short-term prediction of peak injection pressure for the motor cylinders. In Wang, Chen, and Lee (2004), the daily change and prediction of the stocks in Shenzen and Shangai stock markets were investigated. The maximum prediction horizon (T_{m}) where the accuracy has been lost was found as 156 days using the maximum Lyapunov exponent as $T_{\mathrm{m}} = 1/\lambda_1$. The prediction errors were less than 0.05 for $T_{\mathrm{m}} \leq 156$ and 0.30 for $T_{\mathrm{m}} \geq 220$. The nonlinear short-term prediction of the earthquake magnitude using artificial neural networks was studied in Plagianakos and Tzanaki (2001) using 20 years of earthquake time-series data. In Xie, Liu, and Huang (2008), a prediction model of short-time traffic flow based on chaotic time-series analysis methods was presented and applied to predict the real traffic flow.

There are also studies in the literature about neural network-based estimation of aluminum strip thickness using electrical current variables of the cold-rolling machine (Marcellos, Denti, and Saousa 2009). In Zarate (2005), a neural network was trained with the data of the rolling process during the deformation of the material in order to control the output thickness of the aluminum strip. The short-term prediction of aluminum strip thickness was first introduced in Ozturk and Seherli (2015) where the performance of SVMs with various kernel functions was compared for different prediction horizons. In this study, a short-term prediction system was proposed which globally combined multiple regressors to improve the prediction accuracy after applying nonlinear noise reduction to the original time-series data. The experimental results indicated that the proposed system was very promising in short-term prediction of aluminum foil output thickness.

The NARX model

The autoregressive exogenous (ARX) models are linear and they are an adaptation of discrete-time filtering methods developed by Wiener (1949) and then applied to business and economic data by Box, Jenkins, and Reinsel (1994). These models yield successful results especially in prediction of stationary and seasonal time series. However, nonstationary signals are very common in particular when observing natural and cultural phenomena (Kantz and Schreiber 2005).

Nonlinear ARX models are an extension to linear models and define the predicted output as a nonlinear function of past inputs and outputs. A linear single-input-single-output ARX model predicts the current output $y_{\mathrm{p}}(t)$ as a weighted sum of its regressors. In the simplest case, regressors are delayed inputs and outputs and called standard regressors. We used two standard regressors in this study, namely $y(t-1)$ and $y(t-2)$.

$$y_p(t) = [-a_1, -a_2, \ldots, -a_{na}, b_1, b_2, \ldots, b_{nb}] \times [y(t-1), y(t-2), \ldots, y(t-na), u(t), u(t-1), \ldots, u(t-nb-1)]^T \quad (1)$$

where $y(t-1), y(t-2), \ldots, y(t-na), u(t), u(t-1), \ldots, u(t-nb-1)$ are delayed input and output variables, called regressors.

Nonlinear ARX models have more flexible nonlinear function instead of the weighted sum, as in the following

$$y_p(t) = f(y(t-1), y(t-2), y(t-3), \ldots, u(t), u(t-1), u(t-2), \ldots) \quad (2)$$

where f is a nonlinear mapping function and inputs to f are model regressors. Nonlinearity estimators in a nonlinear ARX model structure map the regressors to the model output using a combination of both nonlinear and linear functions. There are various nonlinearity estimators such as tree-partition networks, wavelet networks, or sigmoid networks. In this study, we used sigmoid network with 15 units as the nonlinearity estimator.

Nonlinear time-series analysis

If a time series generated by a nonlinear dynamical system is handled by standard linear methods like power spectrum analysis, linear transformations, or parametric linear modeling, then some critical features of this time series will be undetected or most of the time series will be considered as noise (Tong 1990). Since linear equations can only lead to exponentially changing or periodically oscillating data, all irregular behaviors of the system are because of some random external input to the system (Kantz and Schreiber 2005). However, chaos theory says that the random input is not the only possible source of irregularity in a system's output. A noise-like time series

with random appearance can be generated by a deterministic equation. Furthermore, properly embedding an observed time-series system into a higher-dimensional phase space can provide information about the underlying dynamics. By exploiting this information, we can make precise short-term prediction using local vicinity of embedding vectors in the phase space.

Nonlinearity test with higher-order statistics

The nonlinearity in an observed time series must be verified before applying nonlinear methods. Hinich (1982) has developed algorithms to test for Gaussianity and linearity. The basic idea is that if the third-order cumulants of a process are zero, then its bispectrum is zero, and hence its bicoherence is also zero. If the bispectrum is not zero, then the process is non-Gaussian; if the process is linear and non-Gaussian, then the bicoherence is a nonzero constant. The hypothesis testing problem for non-Gaussianity (nonzero bispectrum) is given below:

H1: the bispectrum of $y(n)$ is nonzero
H0: the bispectrum of $y(n)$ is zero
If hypothesis H1 holds, we can test for linearity, that is, we have a second hypothesis testing problem
H1′: the bicoherence of $y(n)$ is not constant
H0′: the bicoherence of $y(n)$ is a constant

If hypothesis H0′ holds, the process is linear.

If an observed S is consistent with a central chi-squared distribution, this is revealed as probability-of-false alarm (Pfa) value which is the probability of being wrong in assuming that the data have a nonzero bispectrum. If this probability is high, say 0.95, the assumption of zero bispectrum is accepted, that is, the Gaussianity assumption cannot be rejected. In this case, the results of the linearity test should be ignored since the data are Gaussian and hence also linear.

If the data are non-Gaussian where Pfa is very small or zero, an estimate of the constant λ (lambda) value is obtained by computing the mean value of the bicoherence over the points in the nonredundant region. The squared bicoherence is chi-squared distributed with two degrees of freedom and noncentrality parameter λ. Then, the estimated sample interquartile range (R-estimated) of the squared bicoherence can be compared with the theoretical interquartile range (R-theory) of a chi-squared distribution with two degrees of freedom and noncentrality parameter λ. If the estimated interquartile range is much larger or much smaller than the theoretical value, then the linearity hypothesis is rejected.

Nonlinear noise reduction and stationarity

The chaotic time series as well as the time series obtained from natural processes exhibit randomness in between pure deterministic signals like the output of sinus function and pure stochastic signals like the output of white Gaussian noise. For a fully random process with a slowly decaying autocorrelation function, making a precise prediction is impossible even with an absolute knowledge of the present value due to the fact that the underlying Autoregressive Moving Average (ARMA) models are stochastic. On the other hand, if there is a strong correlation in a time series, then the future values will be a linear combination of the preceding observations. The nonlinear deterministic dynamical systems represent another kind of temporal correlation. In contrast to the correlations extracted with the autocorrelation function for the linear signals, the ones for nonlinear deterministic dynamical systems may be visible only by using nonlinear statistics like nonlinear cross-prediction errors (Schreiber 1997). According to the nonlinear cross-prediction errors method, the time series is broken into segments *Si*, where $i = 1, \ldots, N$ and the root mean squared cross-prediction error is computed for segments *Si* and *Sj*. The cross-prediction error as a function of i and j reveals which segments differ in their dynamics (Kantz and Schreiber 2005) and gives a general idea about the stationarity of the time series.

In traditional signal processing, noise reduction means decomposing a time-series value into two components, one of which contains the signal value and the other contains random fluctuations called noise. This approach is not valid for nonlinear time series because they generally have broad-band power spectra and have spectral attributes that generally exhibit random noise behavior. Nonlinear noise reduction methods exploit the structure in the reconstructed phase space instead of the frequency information of the time series. The curved structures formed by the nonlinear signals in delayed phase space are taken into account. Nonlinear phase space filtering focuses on contaminated lower-dimensional manifolds formed by the noisy deterministic signals and project onto those parts in order to reduce noise. There are several nonlinear noise reduction methods discussed in the literature (Davies 1994; Grassberger et al. 1993; Kostelich and Schreiber 1993; Schreiber 1993). In this study, locally projective nonlinear noise reduction method (Kantz et al. 1993) was used for the thickness time-series data. This method assumes that the deterministic part of the data would lie on a low-dimensional attractor while the effect of noise is to spread the data off this attractor. The method tries to identify the attractor and to project the data onto it.

The idea behind the locally projective noise reduction method is that for each vector s_n embedded in the attractor A, there exists a correction θ_n,where θ_n is small, in such a way that $s_n - \theta_n \in A$ and θ_n is orthogonal on A. In

order to apply projection to the attractor, the vectors must be embedded in a phase space which has higher-dimensionality than the attractor A.

The chosen orthogonality metric is important because the delay vectors contain only temporal information. Euclidean distance is not the best choice in this situation. Because the boundaries of the delay vectors will eventually diverge due to the effect of Lyapunov exponents. Therefore, only the middle parts of the delay vectors are corrected and the other parts are left unchanged. This can be expressed with a weight matrix P as in the following

$$P_{ij} = \begin{cases} 1: & i = j \text{ and } 1 < i,j < m \\ 0: & \text{else where} \end{cases} \tag{3}$$

where m is the dimension of the over-embedded delay vectors.

Thus, the following minimization problem is solved for nonlinear noise reduction

$$\sum_i (\theta_i P^{-1} \theta_i) \stackrel{\Delta}{=} min \tag{4}$$

with the constraints $a_n^i(s_n - \theta_n) + b_n^i = 0$, where $i = q+1, \ldots, m$

And $a_n^i P a_n^j = \delta_{ij}$, where a_n^i are the normal vectors of attractor A at the points $s_n - \theta_n$.

The simple nonlinear prediction algorithm (Hegger, Kantz, and Schreiber 1999) is applied to the cross-segments of the time series in order to find the effect of the nonlinear noise reduction. The algorithm calculates the 1-step ahead prediction error on the respective embedding vectors with appropriate embedding dimension and time delay. In a delay embedding space, all neighbors of s_n are taken into account in order to make a prediction at time $n + k$ as in the following.

$$s_{n+\Delta k} = \frac{1}{|\mathcal{U}_\epsilon(s_n)|} \sum_{s_n \in \mathcal{U}_\in(s_n)} s_{n+\Delta k} \tag{5}$$

where $|\mathcal{U}_\epsilon(s_n)|$ is the number of elements in the neighbourhood $\mathcal{U}_\epsilon(s_n)$.

Reconstructing the time-series data

According to the Takens embedding theorem (Takens 1981), if the real dimension of the attractor is D_A, then the intersection of the trajectories can be avoided by choosing the embedding dimension as $D_E > 2D_A$. The embedded time-delay vectors are topologically equivalent to the original time series. The inputs to the prediction system on time t can be constructed as a D_E dimension vector space $Y(n)$, which includes D_E observed points with the same τ time-delay intervals on the observed time series $x(t)$ (Kuremoto et al. 2003)

$$Y(n) = (x(n), x(n+\tau), ..., x(n+(D_E-1)\tau)) \tag{6}$$

The variation of the embedding vectors in time can be shown as $Y(n) \rightarrow Y(n+1)$. In order to reconstruct the attractor correctly, calculation of the time-delay (τ) and minimum embedding dimension (D_E) is very important.

The time-delay value is a multiple of the sampling times of the observed data. If it is too small, then $x(n)$ and $x(n+\tau)$ coordinates in the embedding vectors will be very close to each other and the information gain between delay vectors will be less which leads to data redundancy. If it is too big, $x(n)$ and $x(n+\tau)$ coordinates will be totally irrelevant and the attractor will not exhibit the dynamics of the underlying system. In this study, mutual information (MI) method proposed in Fraser and Swinney (1986) was used. To find the optimum time delay, MI S is calculated for different τ values as in the following

$$S = -\sum_{ij} p_{ij}(\tau) \ln \frac{p_{ij}(\tau)}{p_i p_j} \tag{7}$$

where p_i is the probability of finding an observation value in the ith interval, p_j is the probability of finding an observation value in the jth interval, and $p_{ij}(\tau)$ is the joint probability of finding an observation value in the ith interval and τ times later observation value in the jth interval. The first τ value on which S becomes minimum is the optimum time delay.

The method proposed in Hegger at al. (1999) was used to find the minimum embedding dimension. This method made some small changes to the false nearest neighbors (FNNs) algorithm proposed by Kennel, Brown, and Abarbanel (1992) to avoid the wrong results due to the noise in the time series. Assuming that the standard deviation of the time series is σ, the threshold of FNNs is r, and the distance between the vectors of the phase space is found according to maximum difference, the FNNs statistics is calculated as in the following

$$X_{SYK}(r) = \frac{\sum_{n=1}^{N-m-1} \Theta\left(\frac{\left|S_n^{(m+1)} - S_{k(n)}^{(m+1)}\right|}{\left|S_n^{(m)} - S_{k(n)}^{(m)}\right|} - r\right) \Theta\left(\frac{\sigma}{r} - \left|S_n^{(m)} - S_{k(n)}^{(m)}\right|\right)}{\sum_{n=1}^{N-m-1} \Theta\left(\frac{\sigma}{r} - \left|S_n^{(m)} - S_{k(n)}^{(m)}\right|\right)} \tag{8}$$

where $S_{k(n)}^{(m)}$ is the nearest neighbor of the vector S_n and $k(n)$ is the index of the time series which is different than n and supplying the condition of $|Sn - Sk|$ being minimum. The second Heaviside function in the nominator is used to eliminate the vectors of which initial distances are higher than σ/r. The same function also exists in the denominator for the same reason.

The maximal Lyapunov exponent

A positive maximal Lyapunov (λ_1) exponent means that long-term prediction is impossible because of the exponential divergence of the nearby trajectories on the chaotic attractor. However, especially a small Lyapunov exponent means that short-term prediction is possible because of the local stability of the trajectories. The algorithm developed by Rosenstein at al. (1993) computes the local divergence rates of the state space distances over the whole time-series data. The algorithm is fast, easy to implement, robust to changes in embedding dimension, time delay, noise level, and independent of length of the time-series data. It uses the following formula to find the stretching factor S.

$$S = \frac{1}{N}\sum_{n_0=1}^{N} \ln\left(\frac{1}{|\vartheta_{Xn_0}|}\sum |X_{n_o} - X_n|\right) \tag{9}$$

where X_{n_o} is an embedding vector on the attractor, X_n are the neighboring vectors within diameter ϵ, and ϑ_{Xn_0} is the number of these neighbors. The average distances are summed up for N vectors and the stretching factor is obtained as the average of this value. The first slope of the curve obtained by plotting S values for various N values on x–y coordinate system gives the maximal Lyapunov exponent. In this study, the maximal Lyapunov exponent (λ_1) was used to find the prediction horizon where the accuracy has been lost in the embedded time-series data.

Short-term prediction

It is difficult to model complex, irregular signals by traditional nonlinear analysis methods, because of the large quantities of parameters and their complexity of characteristics. So, machine learning and soft-computing algorithms, such as neural networks, radial basis function (RBF) networks, reinforcement learning, fuzzy logic etc., were considered as effective nonlinear predictors (Kodogiannis and Lolis 2002; Kuremoto et al. 2003; Oliveira, Vannucci, and Da Silva 1996).

After embedding vectors are obtained, they are used as training data for various machine learning algorithms. The output value depends on the prediction horizon where the range is found using maximal Lyapunov exponent.

Neural networks are commonly used for prediction, matching, identification, pattern recognition, optimization, and classification problems (Marcellos, Denti, and Saousa 2009). They are easy to program, well suited to nonlinear systems, and robust against noise (Chaudhuri and Bhattacharya 2000). The multilayer perceptron neural network (MLPNN) used in this study is a feed-forward network using back propagation with stochastic

gradient descent algorithm. The number of hidden nodes in the built network model depends on the time series. The network which was built for training 27 μm time-series data included five neurons in a single hidden layer. On the other hand, a network of three neurons in a single hidden layer was built for the 20 μm time-series data. The optimal momentum and learning rate values were empirically found as 0.2 and 0.1, respectively. Sigmoid was used as the transfer function in the nodes. The MLPNN was trained for 500 epochs for each of the time series. Using longer epochs did not improve the accuracy significantly.

SVMs has strong nonlinear approximation ability by using kernel tricks (Zhang et al. 2013). SVM is based on statistical learning theory (Mohandes and Halawani 2004) and the basic idea is to map the input dataset X into a high dimensional feature space F via a nonlinear mapping function by constructing an optimal hyper plane in this new space (Osowski and Garanty 2007). The particular choice of the kernel function $K(X_i, X_j) = \emptyset(X_i)\emptyset(X_j)$ which transforms the input space to high-dimensional feature space, depends on the nature of the data which is usually unknown. Therefore, the best input-output mapping function can be determined by applying various kernel functions and setting best parameters which will yield the highest generalization performance. In this study, the Pearson VII function-based kernel (PUK) which was proposed by Ustun B et al. (2006) was used after comparing it with polynomial and RBF kernels.

RBF networks are feed-forward neural networks which have a single hidden layer of nonlinear units whose activation function are Gaussian or some other basis kernel function. The training of the RBF network is formulated as a nonlinear unconstrained optimization problem. They are trained with supervised training algorithm but much faster than back propagation networks. Because of the RBF hidden units, they are less susceptible to nonstationary inputs. The basis kernel function was Gaussian and the clustering algorithm was *k*-means in this study.

Global multiexpert combination

According to the *No Free Lunch Theorem*, no single algorithm in a domain always induces the most accurate learner (Alpaydin 2010) and by combining multiple prediction algorithms suitably, accuracy can be improved (Kuncheva 2004). The multiexpert combination method proposed in this study assumes that the prediction algorithms work in parallel and given an input, the best output generated by the regressors is chosen. The accuracy level will always be better, because the method chooses the best prediction with minimum absolute error for each instance found by each algorithm individually. In Figure 1, the proposed global multiexpert combination

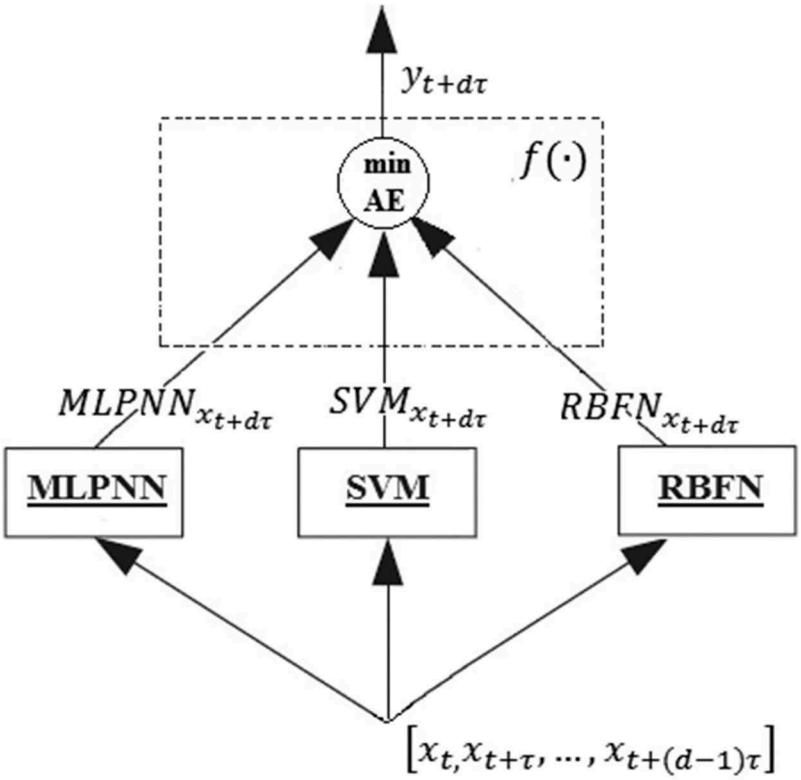

Figure 1. The model of global multiexpert combination (GMEC) with simple voting.

(GMEC) method is given. For each $[x_t, x_{t+\tau}, \ldots, x_{t+(d-1)\tau}]$ input vector, the output of each regressor is found as the predicted value at time $t + d\tau$. The global function $f(\cdot)$ linearly combines the outputs of the prediction algorithms by selecting the one with the minimum absolute error (min AE) and generates the output $y_{t+d\tau}$. Let $MLPNN_{AE} = \left|MLPNN_{x_{t+d\tau}} - x_{t+d\tau}\right|$, $SVM_{AE} = \left|SVM_{x_{t+d\tau}} - x_{t+d\tau}\right|$ and $RBF_{AE} = \left|RBF_{x_{t+d\tau}} - x_{t+d\tau}\right|$, then the $f(\cdot)$ selects the output of the regressor with minimum AE as $y_{t+d\tau}$.

Aluminum foil thickness prediction system model

The sequential steps of the system proposed in this study for aluminum foil thickness prediction are shown in Figure 2 diagrammatically. The details of each step are given below.

Step 1 Data acquisition: The aluminum foil thickness data are obtained from the PC connected to the PLC device for a certain period of time.

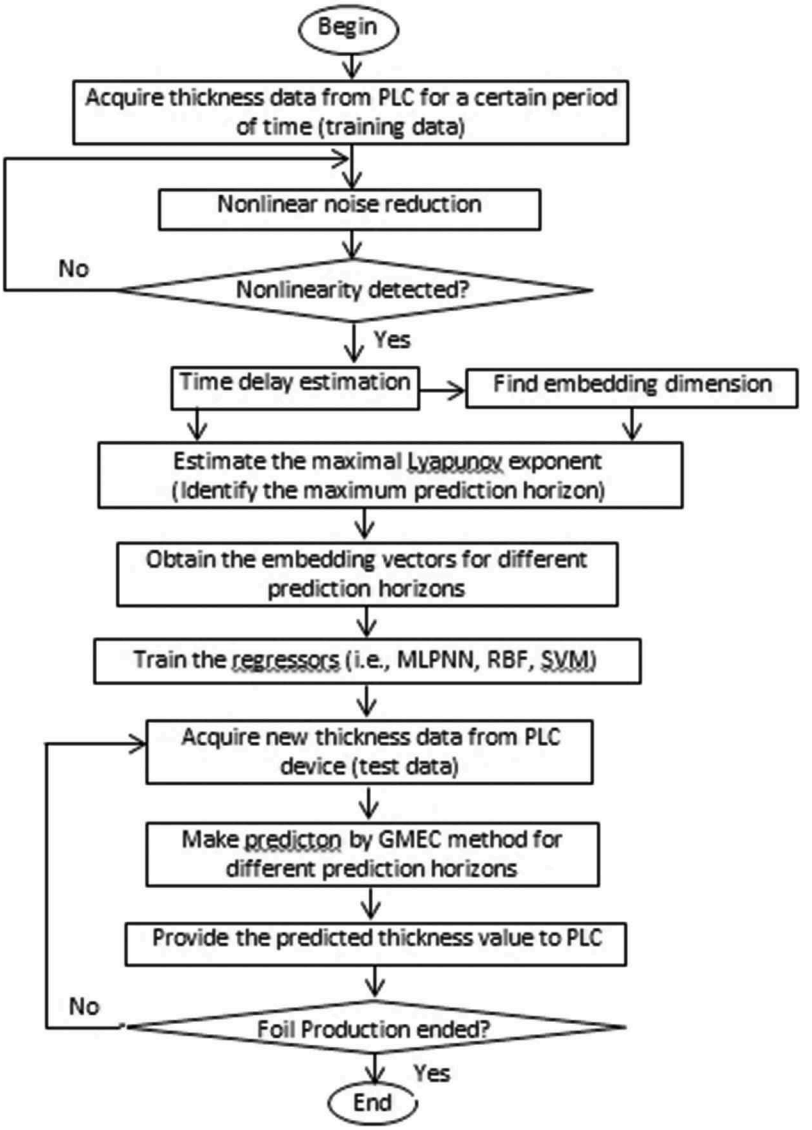

Figure 2. Aluminum foil thickness prediction system model.

Step 2 Nonlinear noise reduction: Identify the deterministic part of the attractor and project the data onto it.

Step 3 Nonlinearity detection: Verify that the time-series data is nonlinear or not.

Step 4 Find the time-delay and embedding dimension: The time-delay and embedding dimension information are crucial for the remaining steps of the system.

Step 5 Estimate the maximal Lyapunov exponent: The maximal Lyapunov exponent gives the time horizon where prediction accuracy is lost.

Step 6 Obtain the time-delay vectors: The time-delay vectors are used for training and testing the proposed system for different prediction horizons.

Step 7 Train the regressors: MLPNN, RBFN, and SVM are trained using training subset.

Step 8 Make prediction on the test subset for different time horizons: The output of the regressor with minimum absolute error is chosen as the prediction of the GMEC for the test subset.

Experimental results and discussion

The output thickness values of the aluminum foil are continuously recorded every second on the computer where PLC software resides. The thickness is measured via an x-ray device, while the foil is passing through the rolling mill.

The experimental results were obtained for two different thickness time series of 20 μm and 27 μm, respectively. In the study, no normalization was applied to the data because of the online nature of the proposed system, but the thickness difference values were used instead. Figure 3 shows the thickness difference values obtained from the recorded data for 12,629 s during a production process of 20 μm aluminum foil. The spikes which correspond to the points where rupture has occurred were eliminated from the signals. The production stops at

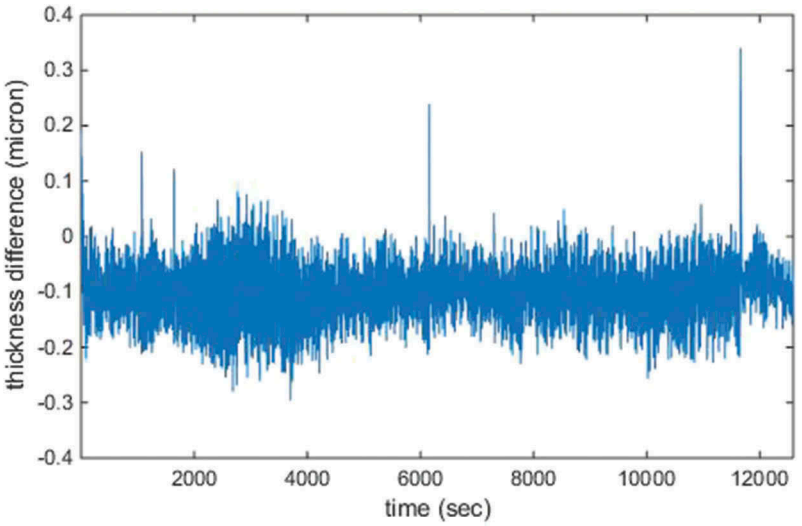

Figure 3. The thickness difference values of 20 μm aluminum foil production.

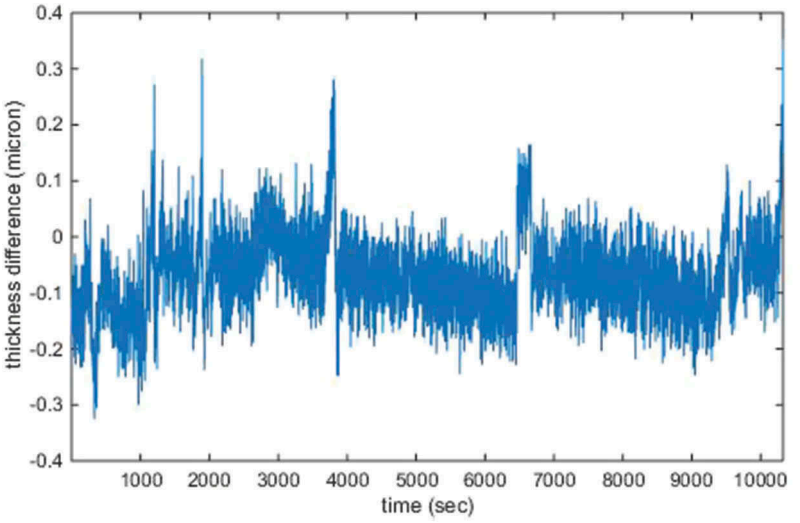

Figure 4. The thickness difference values of 27 μm aluminum foil production.

those points and these data have no meaning for the evaluation of the prediction performance of the proposed method.

The thickness difference values of a 27 μm aluminum foil production recorded for 10,323 s are shown in Figure 4.

The bispectrum and bicoherence methods were applied in order to detect the nonlinearity in the aluminum foil thickness time-series data. According to the Gaussianity test results, the Pfa value for the 20 μm time-series data before nonlinear noise reduction is 0. So the assumption of zero bispectrum is not accepted, which means the Gaussianity assumption is rejected. We continued to the nonlinearity test in this case. The R (estimated) and R (theory) were found as 23.0288 and 10.3963, respectively. The R values are not close, so we cannot accept the linearity hypothesis for 20 μm time-series data. After nonlinear noise reduction, the Pfa value still remained as 0 which indicated non-Gaussianity. The R (estimated) and R (theory) were found as 463.4785 and 46.6494, respectively, which indicated nonlinearity more evidently.

For the original 27 μm time series before nonlinear noise reduction, the Pfa value was found 0.932 which means that the Gaussianity cannot be rejected. So, the results of linearity test were ignored in this case. After nonlinear noise reduction, the Pfa value became 0 which means the Gaussianity assumption is rejected. The R (estimated) and R (theory) were found as 21.4375 and 9.9718, respectively. The R values were not close which indicated nonlinearity after nonlinear noise reduction.

The contour plots of the estimates of the bispectrum for the thickness difference time series of 20 μm and 27 μm aluminum foils are shown in Figure 5 and Figure 6, respectively. The presence of pronounced peaks in the bispectrum is

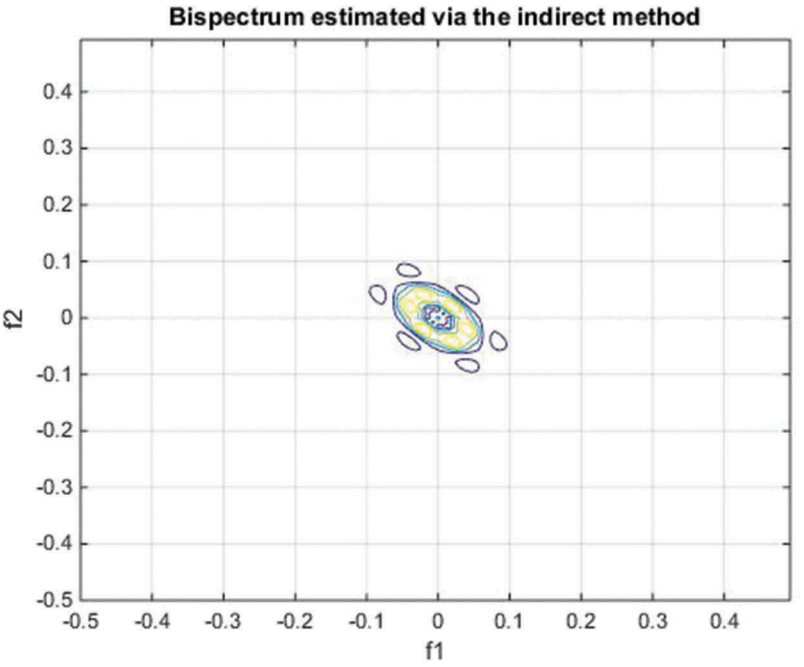

Figure 5. Bispectrum plot for 20 μm time-series data.

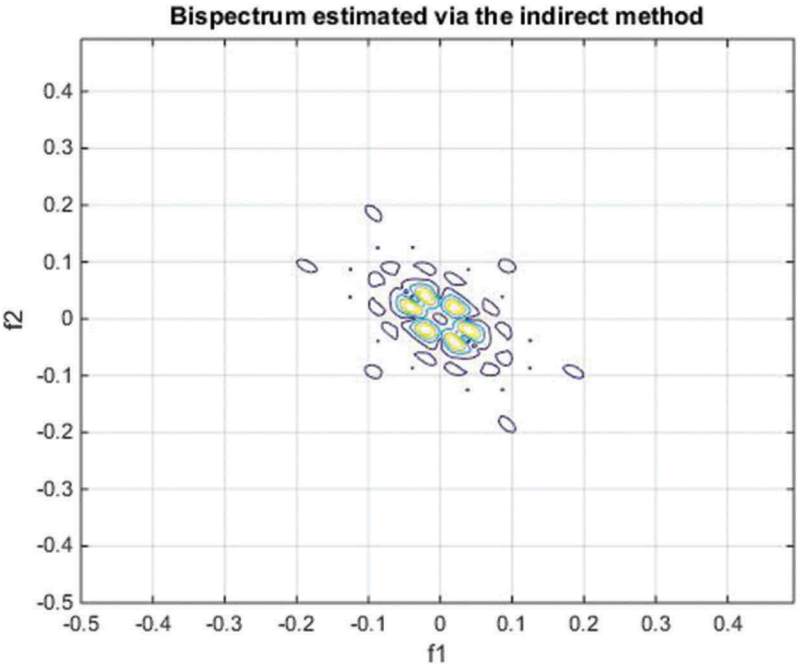

Figure 6. Bispectrum plot for 27 μm time-series data.

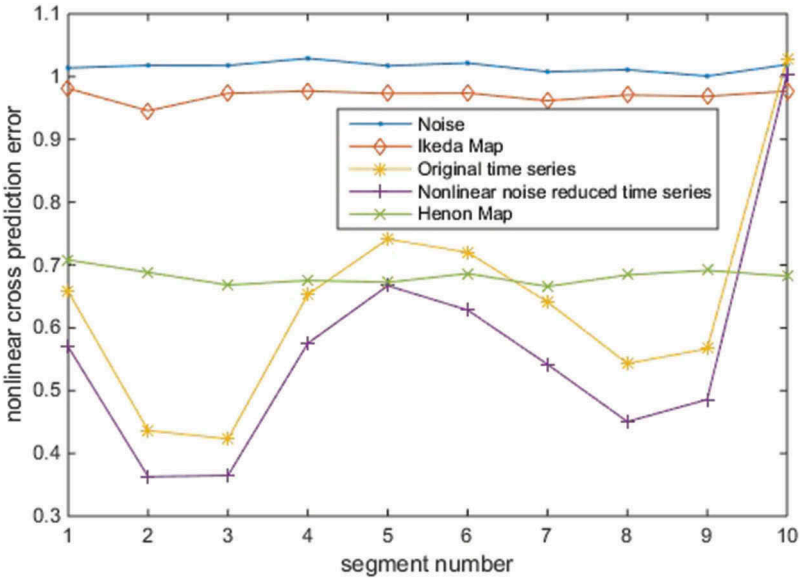

Figure 7. Effect of the nonlinear noise reduction of the nonlinear cross-predictability of the 20 μm thickness time-series data.

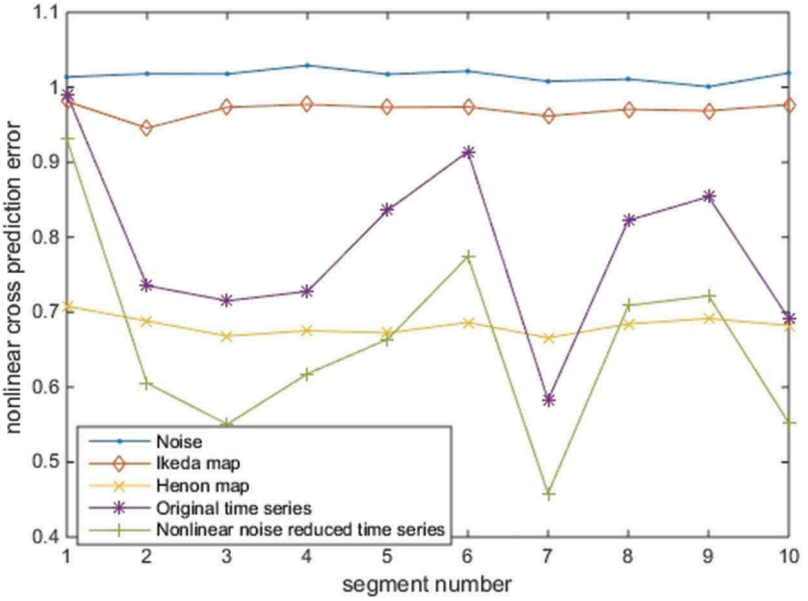

Figure 8. Effect of the nonlinear noise reduction of the nonlinear cross-predictability of the 27 μm thickness time-series data.

indicative of nonlinear phenomena. The effect of nonlinear noise reduction on nonlinear cross prediction errors is shown in Figure 7 and Figure 8, respectively.

The time-delay and embedding dimension are crucial for correctly reconstructing the time series in the phase space. The time-delay vectors are obtained

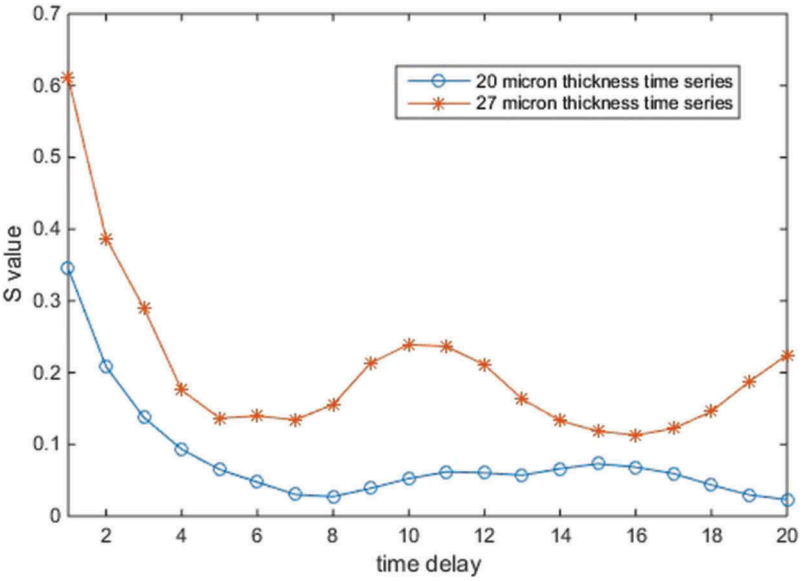

Figure 9. Finding time delay (τ) for each of the thickness time-series data.

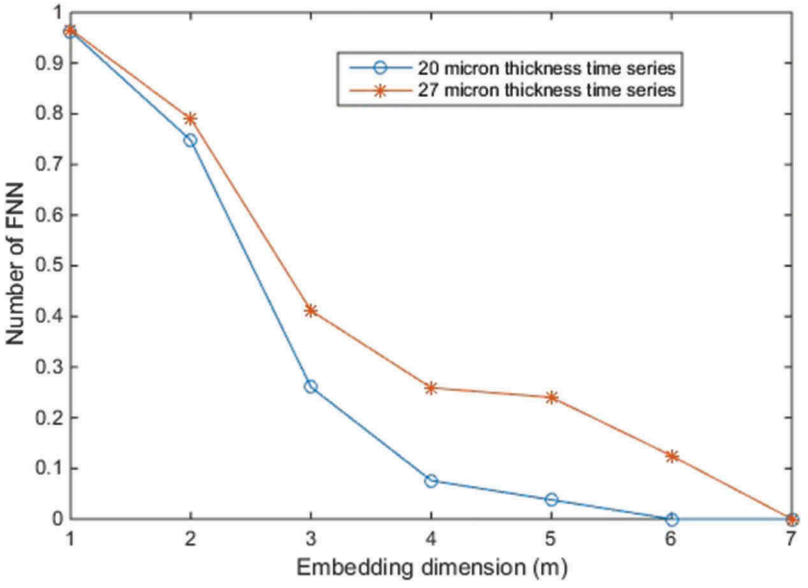

Figure 10. Finding embedding dimension (DE) for each of the thickness time-series data.

by using the time-delay and embedding dimension values. Figure 9 shows how the time-delay value τ is found for each of the thickness time-series data. The first τ value where S is minimum is chosen as the optimum time delay. Here, S takes

the first minimum when τ is 5 for 20 μm time-series data and when τ is 8 for 27 μm time-series data.

Figure 10 shows how the embedding dimension D_E is found for each of the time-series data. The dimension where FNN becomes zero is chosen as the minimum embedding dimension for the thickness time-series data. According to the figure, D_E is taken 9 to reconstruct the time-series time in phase space. The D_E is the found as 9 for any value of the Theiler window t greater than 0.

The maximum Lyapunov exponents (λ_1) were estimated for the 20 μm and 27 μm thickness time series as 0.0473 and 0.0359, respectively. The maximum prediction horizons where accuracy is lost were found approximately as 21 and 28, respectively, which corresponds to $1/\lambda_1$.

The mean absolute percentage error (MAPE) which is used for the evaluation of the machine learning algorithms is given in the following.

$$MAPE = \frac{1}{N}\sum_{i=1}^{n}\left|\frac{T_i - O_i}{T_i}\right| \tag{10}$$

where O_j is the observed value and T_j is the target value.

Each of the time series was divided into training and test subsets. The first two-third of the time series is the training subset, while the remaining one-third is the test subset. The MAPE values for the training subset of 20 μm thickness time-series data are given in Figure 11 for different prediction horizons. According to

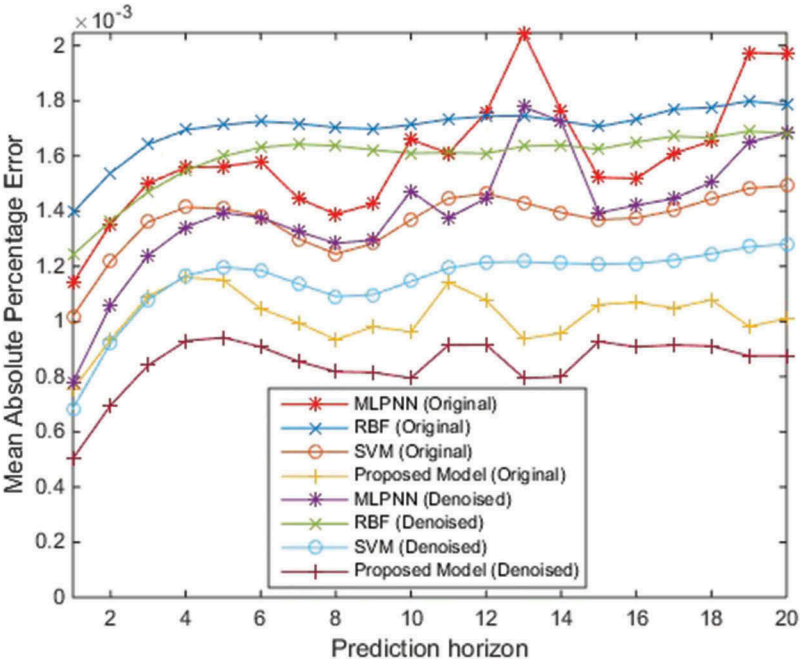

Figure 11. Comparison of the machine learning algorithms for the training subset of 20 μm time-series data.

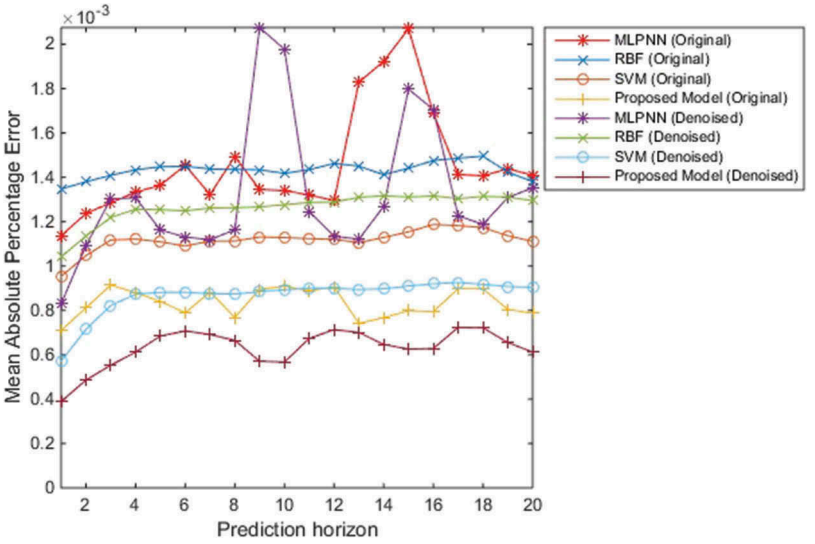

Figure 12. Comparison of the machine learning algorithms for the training subset of 27 μm time-series data.

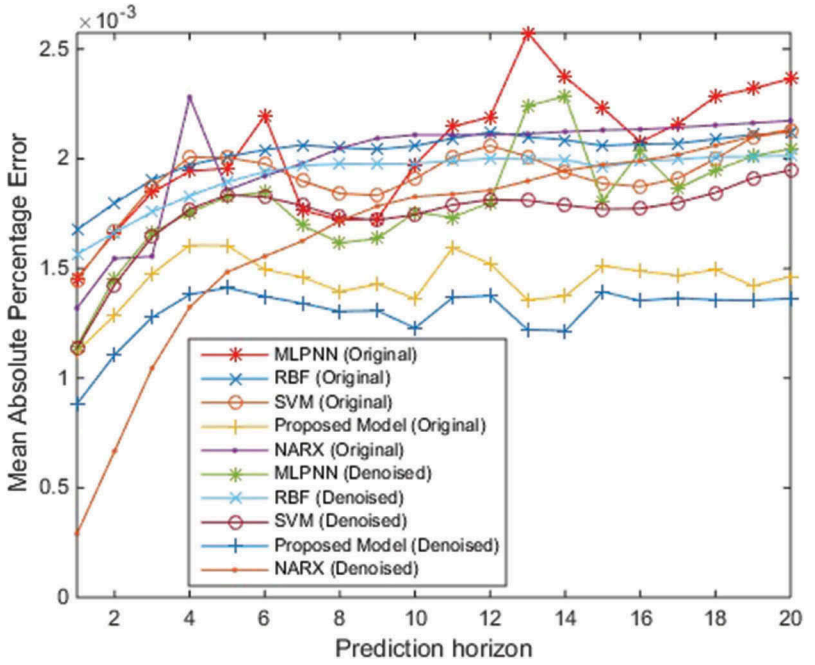

Figure 13. Comparison of the machine learning algorithms for the testing subset of 20 μm time-series data.

the results shown in Figure 11 and Figure 12, SVM model with PUK had the best accuracy among the learners. As expected, the proposed model of GMEC gave better accuracy for all prediction horizons.

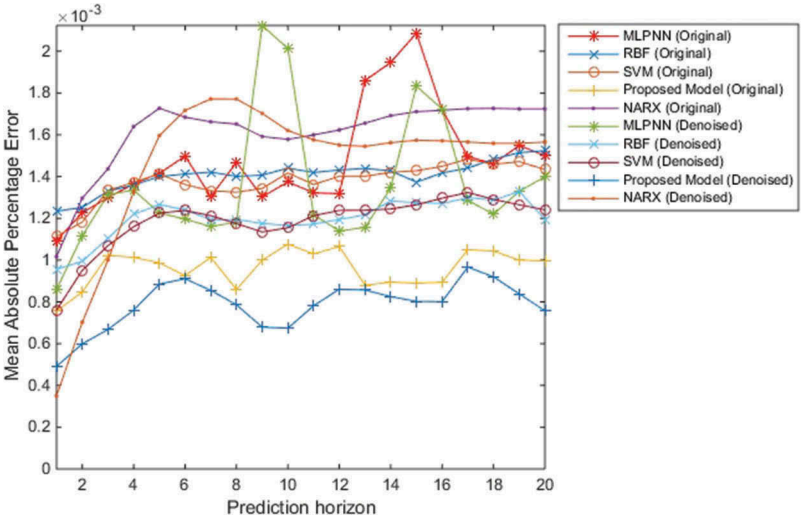

Figure 14. Comparison of the machine learning algorithms for the testing subset of 27 μm time-series data.

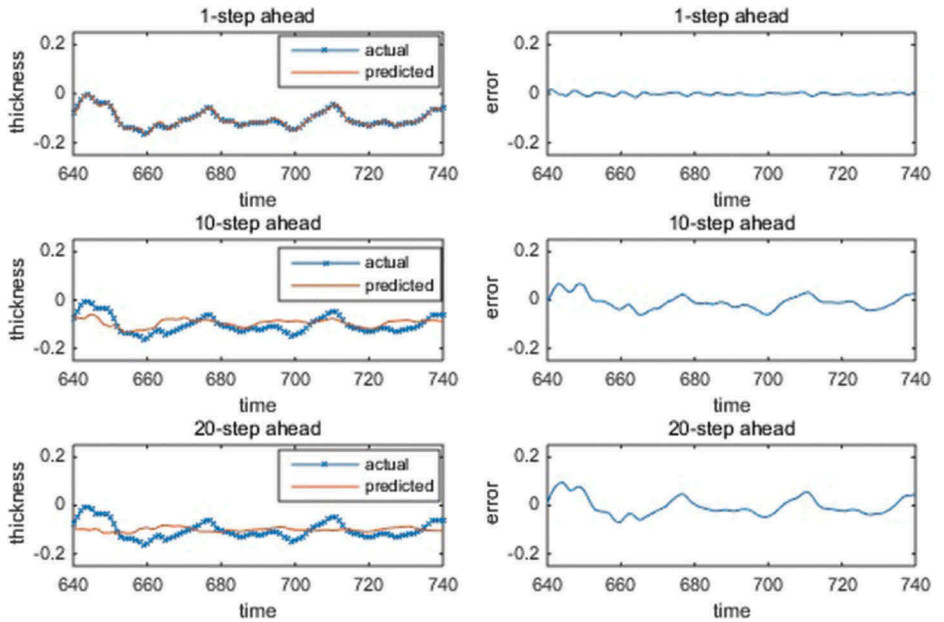

Figure 15. Actual and predicted values of NARX model for 20 μm time-series data.

The MAPE values for the training subset of 27 μm thickness time-series data are given in Figure 12 for different prediction horizons.

The MAPE values for the testing subset of 20 μm and 27 μm thickness time-series data are given in Figure 13 and Figure 14, respectively, for

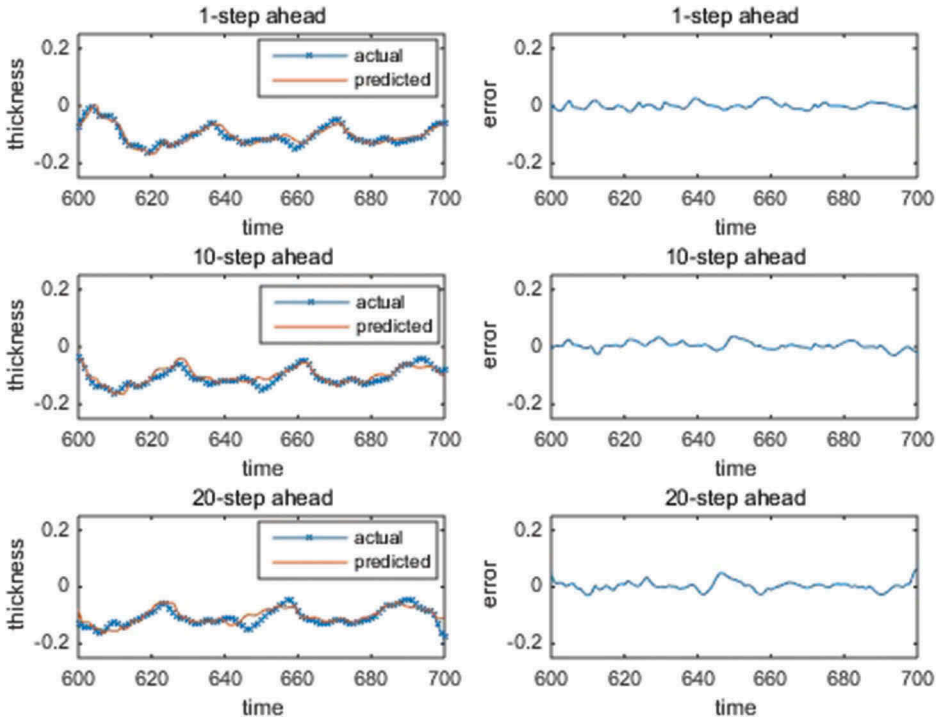

Figure 16. Actual and predicted values of GMEC model for 20 μm time-series data.

different prediction horizons. These figures also include the MAPE values of nonlinear autoregressive exogenous (NARX) method. In Figure 13, NARX has lower MAPE on the nonlinear denoised time series for the first four prediction horizons. After that point, the MAPE for NARX linearly increases, while it is remaining approximately constant for the proposed (GMEC) model. The MAPE values of NARX on the original time series were higher than GMEC for all prediction horizons.

In Figure 14, NARX is better than GMEC model only for 1-step ahead prediction. After then, the MAPE of NARX linearly increases until 8-step ahead prediction and then remains constant but always higher than GMEC for the following horizons. The MAPE values of NARX on the original 27 μm time-series data were higher than GMEC for all prediction horizons.

The actual and predicted thickness difference values of 20 μm time-series data are given in Figure 15 and Figure 16, for NARX and our GMEC model, respectively. The plots are for the first 100 values of the test subset. There is a $\tau \times m$ difference at the beginning of x axis between two figures which corresponds to the embedding vector size. The accuracy for NARX model is better than GMEC model for 1-step ahead prediction horizon as expected, but terribly deteriorates for 10-ahead and 20-ahead prediction horizons.

The actual and predicted thickness difference values of test subset of 27 μm time-series data are given in Figure 17 and Figure 18, for NARX

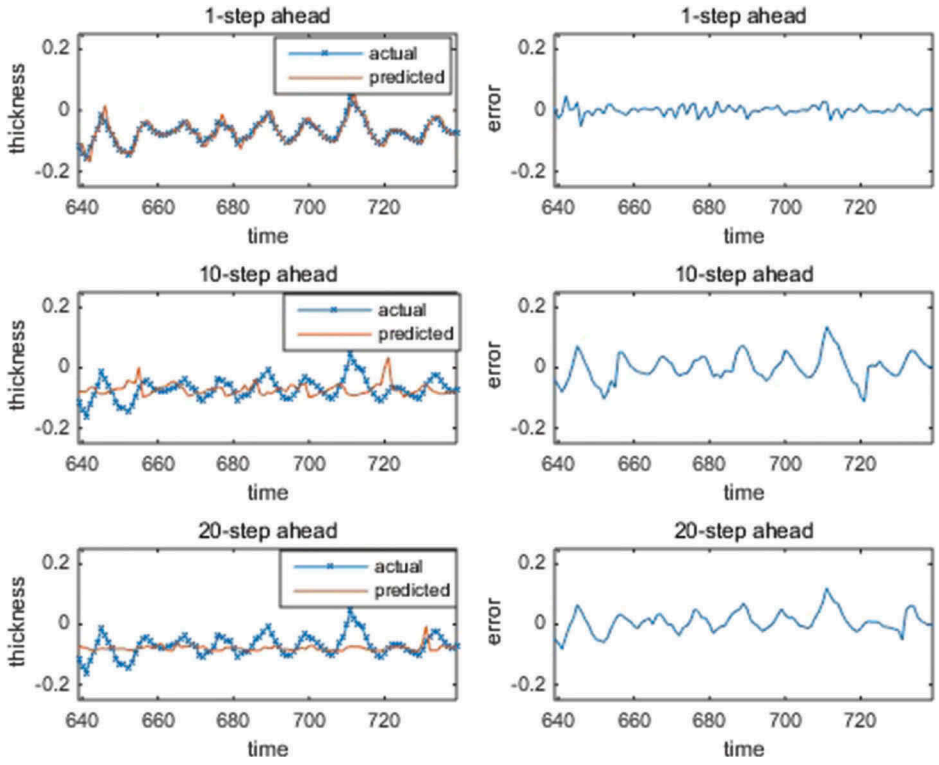

Figure 17. Actual and predicted values of NARX model for 27 μm time-series data.

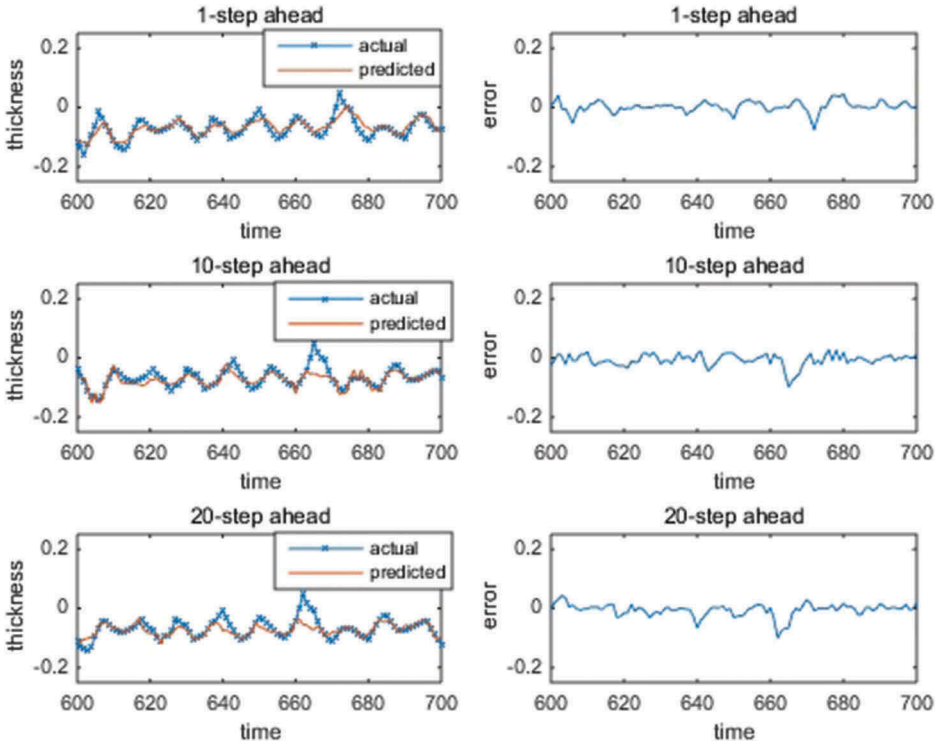

Figure 18. Actual and predicted values of GMEC model for 27 μm time-series data.

and our GMEC model, respectively. Again, the accuracy of NARX model is higher than GMEC model for 1-step ahead prediction, but deteriorates later on.

Conclusions and future work

The thickness predictions obtained from the proposed model will be used to effectively regulate the tension applied to the aluminum foil. In the current instantaneous manner of regulation approach, 5 m of aluminum foil passes through the mills in each second when the thickness value is measured. This latency affects the precision of the aluminum foil thickness. By means of short-term prediction, the velocity of the rollers will be more effectively regulated using the *n*-step ahead thickness prediction value and thus avoiding the latency in roller velocity regulation. The solution proposed in this study can be used in practice by integrating the *n*-step ahead predicted thickness value into the PLC system. The cold-rolling process itself is a closed-loop system where the measured thickness value is used as feedback to regulate the armature and field currents applied to the motors. The process is currently running in online mode without human intervention. Since the prediction in this study is *n*-step ahead, there will be enough time to change the parameters accordingly. Some cold-rolling process experts state that it is sufficient even if the system could tell the change in the thickness will be positive or negative. Therefore, the proposed system has promising chance of application in real production.

References

Alpaydın, E. 2010. *Introduction to machine learning*, 2nd ed. Cambridge, Massachusetts, London, England: The MIT Press.

Bao, Y., T. Xiong, and Z. Hu. 2014. Multi-step-ahead time series prediction using multiple-output support vector regression. *Neurocomputing* 129:482–93.

Basharat, A., and M. Shah 2009. Time series prediction by chaotic modeling of nonlinear dynamical systems. IEEE 12th International Conference on Computer Vision 1941 – 1948, Kyoto, Japan.

Ben Taieb, S., G. Bontempi, A. F. Atiya, and A. Sorjamaa. 2012. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Systems with Applications* 39 (8):7067–83.

Bontempi, G., S. Ben Taieb, and Y. Le Borgne. 2013. Machine learning strategies for time series forecasting. *Lecture Notes in Business Information Processing* 138:62–77.

Box, G. E. P., G. M. Jenkins, and G. C. Reinsel. 1994. *Time series analysis: forecasting and control*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall.

Chaudhuri, B. B., and U. Bhattacharya. 2000. Efficient training and improved performance of multilayer perceptron in pattern classification. *Neurocomputing* 34:11–27.

Davies, M. E. 1994. Noise reduction schemes for chaotic time series. *Physica D. Nonlinear Phenomena* 79:174.

Fraser, A. M., and H. L. Swinney. 1986. Independent coordinates for strange attractors from mutual information. *Physical Review A* 33 (22):1134–40.

Grassberger, P., R. Hegger, H. Kantz, C. Schaffrath, and T. Schreiber. 1993. On noise reduction methods for chaotic data. *Chaos (Woodbury, N.Y.)* 3:127.

Hegger, R., H. Kantz, and T. Schreiber. 1999. Practical implementation of nonlinear time series methods: The TISEAN package. *Chaos (Woodbury, N.Y.)* 9:413.

Hinich, M. J. 1982. Testing for gaussianity and linearity of a stationary time series. J*ournal of Time Series Analysis* 3:169–176.

Iokibe, T., and Y. Fujimoto 2001. Predicting combustion pressure of automobile engine employing chaos theory. Proc. of 2001 IEEE Int. Symp. On Comp Intell. In Robotics and Automation 511–16.

Kantz, H., and T. Schreiber. 2005. *Nonlinear time series analysis*. Cambridge, London, England: Cambridge University Press.

Kantz, H., T. Schreiber, I. Hoffmann, T. Buzug, G. Pfister, L. G. Flepp, J. Simonet, R. Badii, and E. Brun. 1993. Nonlinear noise reduction: A case study on experimental data. *Physical Reviews E* 48:1529.

Kennel, M. B., R. Brown, and H. D. I. Abarbanel. 1992. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Physical Review A* 45:340–53.

Kodogiannis, V., and A. Lolis. 2002. Forecasting financial time series using neural network and fuzzy system-based techniques. *Neural Computing & Applications* 11:90–102.

Kostelich, E. J., and T. Schreiber. 1993. Noise reduction in chaotic time series data: A survey of common methods. *Physical Reviews E* 48:1752.

Kuncheva, L. I. 2004. *Combining pattern classifiers: methods and algorithms*. Hoboken, NJ: Wiley.

Kuremoto, T., M. Obayashi, A. Yamamoto, and K. Kobayashi 2003. Predicting chaotic time series by reinforcement learning. Proc. of The 2nd Intern. Conf. on Computational Intelligence, Robotics and Autonomous Systems.

Mao, W., M. Tian, and G. Yan. 2012. Research of load identification based on multiple-input multiple-output SVM model selection. *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science* 226:1395–409.

Marcellos, N. S., J. F. Denti, and G. C. Saousa. 2009. Strip thickness estimation in rolling mills from electrical variables in AC drives. *Latin American Applied Research* 39:353–59.

Mohandes, M. A., and T. O. Halawani. 2004. Support vector machines for wind speed prediction. *Renewable Energy* 29:939–47.

Nesreen, K. A., A. F. Atiya, N. E. Gayar, and H. El-Shishiny. 2010. An empirical comparison of machine learning models for time series forecasting. *Econometric Reviews* 29 (5):594–621.

Niu, D., D. Liu, and D. D. Wu. 2010. A soft computing system for day-ahead electricity price forecasting. *Applied Soft Computing* 10:868–75.

Oliveira, K. A., A. Vannucci, and E. C. Da Silva. 1996. Using artificial neural networks to forecast chaotic time series. *Physica A* 284:393–404.

Osowski, S., and K. Garanty. 2007. Forecasting of daily meteorological pollution using wavelets and support vector machine. *Engineering Applications of Artificial Intelligence* 20:745–55.

Ozturk, A., and R. Seherli. 2015. Short term prediction of aluminium strip thickness via Support Vector Machines. *23th Signal Processing and Communications Applications Conference* 1:304–07.

Plagianakos, V. P., and E. Tzanaki. 2001. Chaotic analysis of seismic time series and short term forecasting using neural networks. *IEEE International Joint Conference on Neural Networks* 3:1598–602.

Rosenstein, M. T., J. J. Collins, and C. J. De Luca. 1993. A practical method for calculating largest Lyapunov exponents from small data sets. *Physica D. Nonlinear Phenomena* 65:117–34.

Schreiber, T. 1993. Extremely simple nonlinear noise reduction method. *Physical Reviews E* 47:2401.

Schreiber, T. 1997. Detecting and analyzing non-stationarity in a time series with nonlinear cross-predictions. *Physical Review Letters* 78:843–46.

Sorjamaa, A., J. Hao, N. Reyhani, Y. Ji., and A. Lendasse. 2007. Methodology for long-term prediction of time series. *Neurocomputing* 70:2861–69.

Takens, F. 1981. Detecting strange attractors in turbulence. In Rand D., Young LS. (eds) Dynamical Systems and Turbulence, Warwick 1980. Lecture Notes in Mathematics, Vol. 898. Berlin, Heidelberg: Springer.

Tong, H. 1990. *Non-linear time series – A dynamical system approach*. Oxford, England: Oxford University Press.

Tran, V. T., B. Yang, and A. C. C. Tan. 2009. Multi-step ahead direct prediction for the machine condition prognosis using regression trees and neuro-fuzzy systems. *Expert Systems with Applications* 36:9378–87.

Ustun, B., W. J. Melssen, and L. M. C. Buydens. 2006. Facilitating the application of support vector regression by using a universal Pearson VII function based kernel. *Chemometrics and Intelligent Laboratory Systems* 81:29–20.

Wang, H., G. Chen, and J. Lee. 2004. Complex dynamical behaviours of daily data series in stock exchange. *Physics Letters A* 333:246–55.

Wiener, N. 1949. *Extrapolation, interpolation, and smoothing of stationary time series*. New York: Wiley.

Xie, H., Z. Liu, and H. Huang. 2008. Nonlinear forecasting of daily traffic flow based on optimal embedding phase-space. *IEEE International Conference on Machine Learning and Cybernetics* 3:1341–46.

Yang, H., and X. Duan. 2003. Chaotic characteristics of electricity price and its forecasting model. *Canadian Conference on Electrical and Computer Engineering* 1:659–62.

Zarate, L. E. 2005. A predictive thickness control structure and decision about the better control parameter for the cold rolling process through sensitivity factors via neural networks. *IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications* 1:17–23.

Zhang, L., W. Zhou., P. Chang, J. Yang, and F. Li. 2013. Iterated time series prediction with multiple support vector regression models. *Neurocomputing* 99:411–22.