Journal of Siberian Federal University. Engineering & Technologies 2021 14(8): 966–985
~ ~ ~

DOI: 10.17516/1999-494X-0366

УДК 669.712.111.2;004.891.2

Machine Learning Approach to Simulation of Continuous Seeded Crystallization of Gibbsite

Vladimir O. Golubev*[a], Iliya V. Blednykh[a], Matvey V. Filinkov[b], Oleg G. Zharkov[c] and Tatiyana N. Shchelkonogova[c]

[a]RUSAL Engineering and Research Center
Department of Mathematical Modeling
St. Petersburg, Russian Federation
[b]JSC «RUSAL URAL» in Kamensk-Uralsky
Production Department
Kamensk-Uralsky, Russian Federation
[c]RUSAL Engineering and Research Center
Department for Technology and Technical Development of Alumina Production
Kamensk-Uralsky, Russian Federation

Received 15.06.2021, received in revised form 03.08.2021, accepted 15.10.2021

Abstract. Continuous seeded crystallization is characterized by oscillations of particle size distribution (PSD) and liquor productivity. To describe these oscillations using analytical methods is a complicated task due to non-linearity and slow response of the process. This paper uses a statistical approach to the preparation of initial data, determination of the significant factors and arrangement of the said factors by their impact on the dynamics of crystal population development. Various methods of machine learning were analyzed to develop a model capable of forecasting the time series of particle size distribution and composition of the final solution. The paper proposes to use deep learning methods for predicting the distribution of crystals by grades and liquor productivity. Such approach has never been used for these purposes before. The study shows that models based on long short-term memory (LSTM) cells provide for better accuracy with less trainable parameters as compared with other multilayer neural networks. Training of the models and the assessment of their quality are performed using the historical data collected in the hydrate crystallization area at the operating alumina refinery.

Keywords: seeded crystallization, oscillation process, prediction of time series, deep learning, alumina production, long short-term memory, convolutional network.

* Corresponding author E-mail address: vladimir.golubev2@rusal.com

Citation: Golubev, V. O., Blednykh, I. V., Filinkov, M. V., Zharkov, O. G., Shchelkonogova, T. N. Machine learning approach to simulation of continuous seeded crystallization of gibbsite, J. Sib. Fed. Univ. Eng. & Technol., 2021, 14(8), 966–985. DOI: 10.17516/1999-494X-0366

Моделирование непрерывной затравочной кристаллизации гиббсита методом машинного обучения

В. О. Голубев[а], И. В. Бледных[а], М. В. Филинков[б], О. Г. Жарков[в], Т. Н. Щелконогова[в]

[а]РУСАЛ Инженерно-технологический центр
Отдел математического моделирования
Российская Федерация, Санкт-Петербург
[б]АО «РУСАЛ УРАЛ» в Каменске-Уральском
Производственный отдел
Российская Федерация, Каменск-Уральский
[в]РУСАЛ Инженерно-технологический центр
Департамент по технологии и техническому развитию глиноземного производства
Российская Федерация, Каменск-Уральский

Аннотация. Непрерывной затравочной кристаллизации характерны осцилляции фракционного состава и продуктивности раствора, которые трудно описать аналитическими методами из-за существенной нелинейности и высокой инерционности процесса. В работе использован статистический подход к подготовке исходных данных, определению значимых факторов и их ранжированию по степени влияния на динамику развития популяции кристаллов. Выполнен анализ эффективности различных методов машинного обучения для построения модели, прогнозирующей временные ряды классов крупности частиц и состав конечного раствора. Предложен способ прогнозирования распределения популяции кристаллов по размерам и продуктивности раствора с использованием методов глубокого обучения, который для решения этой задачи в мировой практике еще не применялся. Показано, что модели на основе ячеек с долгой краткосрочной памятью (LSTM) обеспечивают более высокую точность при меньшем числе обучаемых параметров в сравнении с другими архитектурами многослойных нейронных сетей. Обучение моделей и оценка их качества выполнены на основе архива исторических данных, собранных на участках кристаллизации гидроксида алюминия на действующем глиноземном заводе.

Ключевые слова: затравочная кристаллизация, осцилляционный процесс, прогнозирование временных рядов, глубокое обучение, производство глинозема, сеть с долгой краткосрочной памятью, сверточная сеть.

Цитирование: Голубев, В. О. Моделирование непрерывной затравочной кристаллизации гиббсита методом машинного обучения / В. О. Голубев, И. В. Бледных, М. В. Филинков, О. Г. Жарков, Т. Н. Щелконогова // Журн. Сиб. федер. ун-та. Техника и технологии, 2021, 14(8). С. 966–985. DOI: 10.17516/1999-494X-0366

Introduction

The seed improves the coarseness of the solids particles in course of the crystallization of metal salts. Continuous seeded crystallization is used for large-scale production of aluminum hydroxide, potassium chloride, ammonium sulfate. Distribution of crystals influences chemical purity, porosity, bulk density, an angle of repose and other properties of a crystalline product, thus it is one of its crucial characteristics. Though industry requires crystals of a regular size at maximum extraction from the liquor, continuous seeded crystallization usually results in dynamic fluctuations of these properties characterized by a very complex nature (Fig. 1).

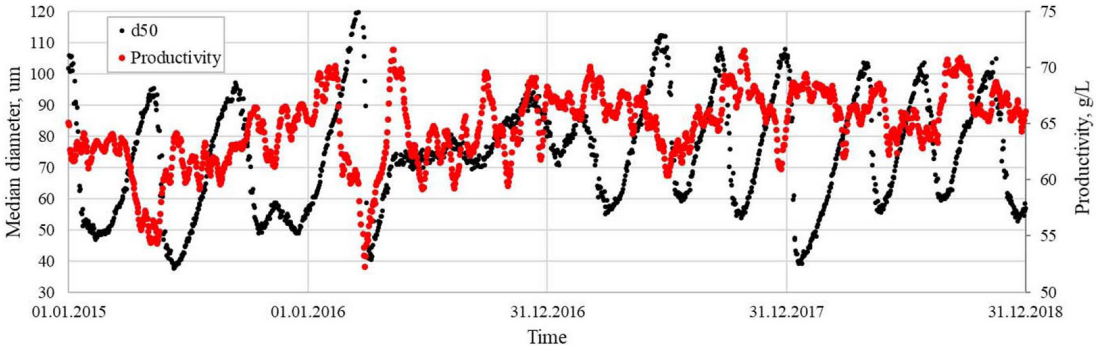


Fig. 1. Oscillations of a crystal median size and liquor productivity at the alumina refinery during the continuous seeded crystallization

The best example of seeded crystallization is the precipitation of gibbsite from the supersaturated sodium aluminate liquor during the alumina production by Bayer process [1-3]. Coarse alumina is more valuable in terms of aluminum smelting due to lower mechanical losses, better transportation properties; besides it enhances cell operation, improves current efficiency and consumption of cryolite and anodes [4]. To meet the smelters' requirements alumina refineries aim to achieve stable particle size distribution (PSD) of their product. Excessive crystal coarseness reduces the process efficiency and increases the coefficients of consumption of raw materials and energy supplies, while excessive fineness does not comply with customers' requirements to the product quality.

Currently the size of crystal powder obtained in course of continuous seeded crystallization is controlled manually. The process engineer takes decisions based on the PSD, number and appearance of crystals, current productivity, and personal experience. Due to the complexity of the process such approach often results in excessive corrections or loss of control. Application of a predictive model can improve the control efficiency.

The dynamics of changes of PSD and productivity of the liquor is dependent on multiple disturbing factors, unbalanced control and complexity of the seeded crystallization process that is described by a differential equation with nonlinear right-hand side [5]:

$$\frac{\partial n_i}{\partial t} + G\frac{\partial n_i}{\partial u} = B_i - D_i, \quad (1)$$

where n_i is a number of i^{th} grade particles in unit volume of a crystallization vessel, $1/m^4$; t is time; $\partial n_i/\partial t$ is the rate of change of the number of i^{th} grade particles over the time in the slurry unit volume, $1/(m^4 \times s)$; G is the linear rate of crystal growth, m/s; B_i and D_i are functions expressing the birth and death rates of i^{th} grade particles due to their formation and transfer into other grades respectively, $1/(m^4 \times s)$.

Computational complexity of a numerical solution of equation (1) along with the need for its accurate tuning do not promote its use for in-process control of industrial crystallization cycles [6]. The available scientific papers on crystallization dynamics are mainly academic and based on laboratory research [7-9]. For now, no industrially applied model based on the population balance equation has been developed to describe and predict changes in crystal PSD and liquor productivity.

The succession of the values of registered crystallization parameters reflects the dynamics of changes in the population of crystals circulating with the seed via the crystallization tanks (Fig. 1). As per equation (1) changes of a crystal size occur in chronological sequence and cause the changes in properties and behavior of all population. Therefore, it can be inferred that the analyzed time series have recursion and connectiveness. Regular alternation of maximum and minimum values indicates the repeatability and seasonality in these data. Due to these properties, the prediction of the dynamics of continuous seeded crystallization can be referred to as a classical machine learning task similar to the prediction of weather, financial risks, recourse requirement, epidemics [10-13]. Multivariate time series forecasting methods are used as tools for addressing and solving such tasks [14, 15].

1. Literature Review

To analyze and predict multivariate time series, linear and non-linear methods are applied. The most commonly encountered methods are regression and correlation methods, as well as classical, recurrent and convolutional neural networks.

1.1. Regression and correlation methods

In autoregression models every variable is a linear function of the previous states of itself and other variables. Multivariate time series are predicted with the use of Vector Autoregressive Models (VAR), Vector Moving-Average Models (Vector ARMA). Correlation methods are based on the application of parameter linear combinations instead of their actual true values: Cross-Correlation Matrices, Threshold Cointegration and Arbitrage [16-18].

A decision tree refers to structural adjustment methods with unknown regression function [18]. Due to its adaptability, this method provides for quick adjustment to non-linear changes of time series, but it is prone to overfitting. A random forest algorithm averages possible prediction errors of some decision trees thus recovering the compatible computational output for the ensemble of decision trees [19]. Performance of the random forest method decreases in proportion to the number of used trees.

1.2. Classical Artificial Neural Networks

An Artificial Neural Network (ANN) is an interconnected group of nodes called neurons. Each neuron calculates the linear sum of the input signals taken with specific weights. The neuron output is used to calculate the non-linear threshold function [15]. Unlike linear models, neural networks have no restrictions on taking into account non-linear relationships in data. They are also more resistant to

stochastic changes in the input data as compared with linear models that is essential for handling the industrial process data [20].

Artificial Neural Networks based on Multilayer Perceptron (MLP) were among the first methods used for the prediction of time sequences. In this case, the values of the input variables at points in time preceding the current moment are fed to the input MLP layer in the form of unrelated parameters [21]. Borovikov V. and Milkov M. [22] present the case study of using the perceptron model with one hidden layer to predict the changes in the content of one size grade during the continuous crystallization of gibbsite. Unfortunately, when dividing the data into training and test samples, the authors placed the test series not at the edge of the time interval of the analyzed data, but interspersed with the training series, without indents. This interspersion of the test and training data resulted in the overestimation of the model quality.

1.3. Recurrent Neural Networks

A recurrent neural network (RNN) is an artificial neural network that uses the data of previous points of the time series to predict the values in the future prospective points. Such networks consist of sequence of nodes where each node is connected with similar adjacent nodes and transfers the information along the entire series. Connections between the nodes are of variable weight that enables to adjust the influence of the network previous conditions on the results of prediction of current and upcoming values of the time series. The main disadvantage of fully recurrent RNNs is the problem of disappearing or explosive gradients in the process of training using the backpropagation method [23].

A long short-term memory (LSTM) network architecture differs from RNN by the presence of memory cells in the nodes. Each LSTM cell has links to one or more neighboring cells and three multiplicative parameters: entry, exit and forget gates [24]. Due to such architecture, the cells control the significance of the information in the memory of the adjacent cells thus overcoming the RNN limitations. LSTM networks are used for developing the predictive models in hydrogeology, climatology, medicine, mechanics, economics, social sciences, natural language recognition, image recognition [25-28]. Networks with Gated Recurring Unit (GRU) are a simplified version of LSTM networks that do not have the input gate, so they require less memory and time for training, besides they are more productive, but less accurate on long time series [29].

1.4. Convolutional Neural Networks

Convolutional Neural Networks (CNN) are mainly designed to recognize patterns (images of objects, handwritten texts, events). Unlike MLP networks that have interlayer connections among all neurons, a CNN contains only a small set of a weight matrix - a convolution kernel that encodes a specific feature. The set of these kernels forms a feature map. Changing the scale of the time series allows you to encode, and in the future, predict its small and large features. Finally, all features arrive to the fully connected neural network or LSTM network with the much smaller size as compared with standard MLP network [30, 31]. The research proves the validity of using a CNN network independently or jointly as CNN-LSTM networks for forecasting of time series. Brownlee J. [32] performed the comparative analysis of MLP, LSTM, CNN, CNN-LSTM networks using such time series as auto sales and power consumption by a city.

A Temporal Convolutional Network (TCN) was introduced in 2016 to recognize the moving objects in the video footage [33]. This type of a CNN network uses random convolutions and extensions so it is more suitable for the sequences that can alter over the course of time. When used for prediction of very long sequences, this method proves to be more effective as compared with RNN and LSTM networks [29, 34].

1.5. Methods for improving the initial data

Along with useful information, industrial data inevitably contain all kinds of noises, outliers and data gaps, systematic changes in the signal level (disconnection, zero offset, amplitude change). The initial data might contain variables that are not significant for the specific prediction task, or alternatively, some features might contain similar information. Data preprocessing is required to isolate useful information against the background of interference and weak signals.

A preprocessing method usually differs depending on each specific dataset. Outlier detection and gap filling in long data sequences can be done in a sliding window. One of the simplest methods of outlier detection is Z-score, which determines how many standard deviations from the sample mean are needed to classify a value as outlier [35]. A K-nearest neighbor method (KNN), that can also be used for multivariate data, allows detecting single outliers and filling the gaps [36, 37]. For oscillating features, it can be done with Discrete Fourier or Wavelet transformation [38, 39].

Relative significance of certain variables of a non-linear model can be evaluated by application with the use of random forest or gradient boosting. In this case, for a multidimensional random variable, a space of lower dimension is sought, in which the correlations between its individual coordinates tend to zero. At the same time, the corresponding variance in the initial data is preserved, which allows performing inverse transformations without significant loss of information. To reduce dimensionality of the time series, various methods are used, e.g. Principal Component Analysis (PCA) which uses the matrix factorization technique; Singular Value Decomposition (SVD); Randomized PCA which uses a stochastic algorithm that accelerates convergence; Kernel PCA which applies decision boundaries; Local Linear Embedding (LLE) which is similar to K-nearest neighbor method, etc. [14, 40]. In the present paper a standard Principal Component Method is applied.

1.6. Selection of methods

Prediction of changes of seed PSD is characterized by the time series of middle length (we consider the prediction for 60 points) that are also predominantly steady (among several consecutive oscillations we are interested in the seasonal oscillation pattern without explicit shift in average) and non-linear (as per the known theoretical description (1)). Based on this description and literature review the most prospective methods for prediction of continuous seeded crystallization parameters are deep learning networks. The quality of forecasts can be enhanced by eliminating systematic bias and rough errors and reducing random noise by filtering or smoothing the original data. Reducing the cross-correlation of input variables by regularizing l1 and/or l2 can also be useful.

2. Research Methodology

2.1. Collection of initial data

The initial data that are used for the present study are taken from the historical database of the parameters of precipitation of the pregnant liquor (Bayer circuit) collected at the Urals alumina

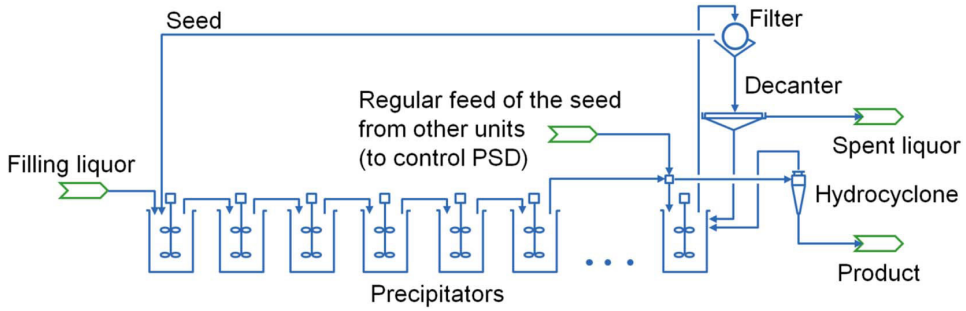


Fig. 2. PFD of continuous crystallization of gibbsite

refinery (Kamensk Uralsky, Russia) within the period from 2011 to 2020. Fig. 2 shows the PFD of the precipitation area. There are three precipitation areas at the refinery. Each comprises precipitation trains (of deferent capacity), slurry classification and filtration unit.

The records were made on a daily basis both manually and automatically. There are over 3600 records for each of the areas. Every record comprises such parameters as the flow rate of the pregnant liquor fed to the area (*Q*); alumina concentration in the pregnant liquor (*A*), total (*S*) and caustic (*C*) alkali; A/C ratio in the spent liquor (*A_C*); solids in the samples from the first precipitator (*Cs*); content of minus size grades in the seed of the final precipitators (*F5, F10, F20, F30, F45, F50, F63, F100, F125* and *F150* – the numbers correspond to the sieve sizes, μm); slurry temperature in the first (*T1*) and final (*Tn*) precipitators. The seed slurry can be redirected to other areas so the following seed flows were registered, i.e., from area 10 to area 3 (*Qs3)* and from area 10 to area 6 (*Qs10).* That data set includes also other parameters, i.e., period of feeding the liquor after chemical cleaning of precipitators, number of hydrocyclones in operation, diameters of spigots of hydrocyclones, PSD of overflows and sands of hydrocyclones, etc.

2.2. Data sampling

The process data which were used for the present study are registered and collected by various methods. Some of the data are the readings from the automatic sensors (flow rate and temperature); some values were calculated indirectly and recorded in the logbooks (the amount of the slurry transferred among the areas was determined on the basis of the pump operation period); and some data are the results of measurements/analysis of samples of slurries and liquors (chemical composition, solids concentration, PSD).

The initial data are characterized by a high level of the noise due to errors during the sampling, sample preparation and measurements. Data outliers are associated with faults of the sensors, errors caused by data manual input, rough errors in the measurements. Data gaps result from faults of the sensors, failures to perform analysis/measurements (days off, failures of the analytical instruments).

All parameters were preprocessed using the same method. Outliers were detected and excluded using Z-score in a sliding window with a width of 10 records. Values in the 2.2-3.0 standard deviation range were found to be acceptable for cutting off the most part of coarse outliers (Fig. 3). Eliminated outliers amounted to 0.5-6% for different parameters. Data gaps were filled with their previous values. To reduce the noise level, the average values in a sliding window with a width of 3 points were used.

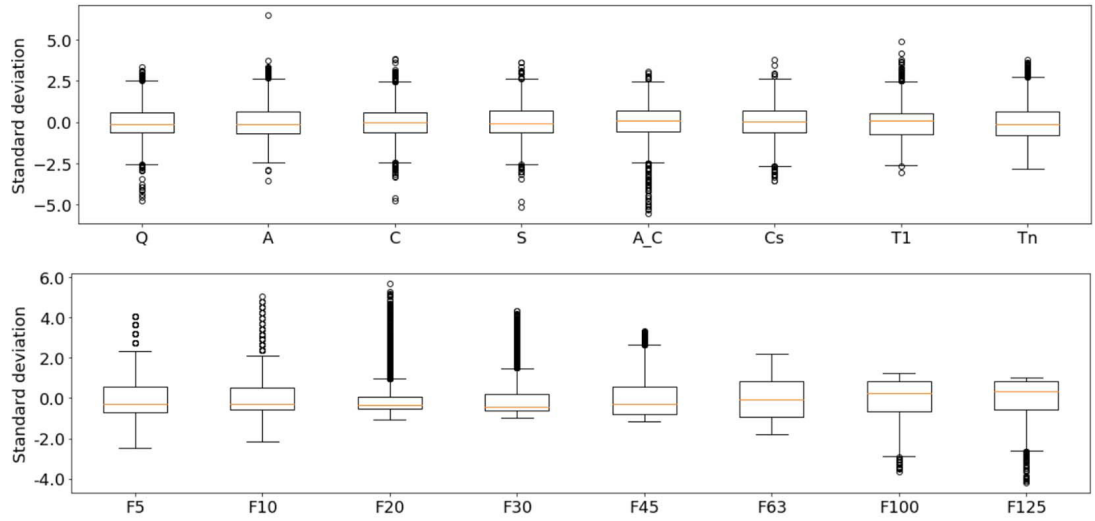


Fig. 3. Box plots before elimination of coarse outliers for some parameters

2.3. Conditionality assessment

Presence of cross correlations in the introduced set of parameters is easily detected. Strong cross correlation is observed between the content of certain grades of particles in the population and the parameters of the chemical composition, *A*, *C* and *S*. Pair correlation coefficients of all key target parameters with other factors are sufficiently close in all precipitation areas (Fig. 4).

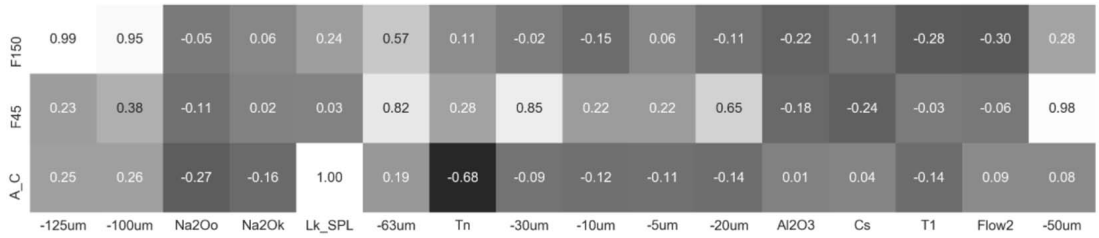
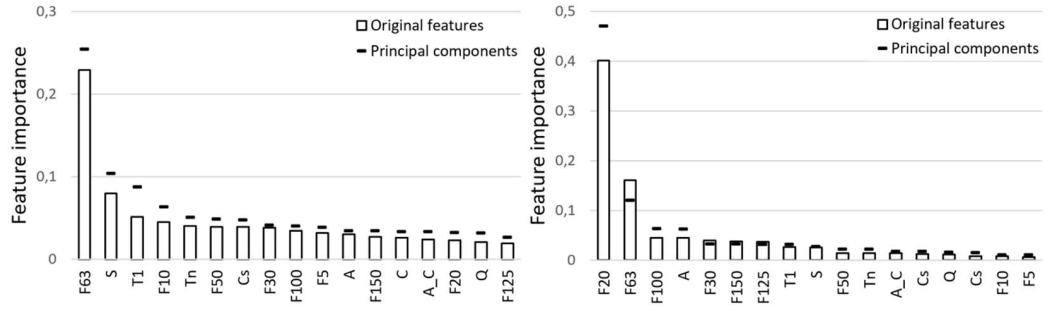


Fig. 4. Heat map of correlation coefficients of main target parameters with other factors in precipitation area No. 6

Individual significance of parameters was assessed using a random forest algorithm to generate a relevant data sample, the number of trees was assumed equal to 50. The present study showed that the variation of target functions is described to the fullest extent by the following parameters: -10, -20, -63 μm grade content, temperature in the first precipitator, concentration of soda and alumina in the pregnant liquor (Fig. 5).

To access the possibility of dimensionality reduction and improvement of model quality, a Principal Component Method was applied. It was determined that for the same dimensionality as the original features the reduction of correlation between its coordinates promotes to some extent the total importance of the features in relation of -45 μm fraction in the slurry (Fig. 5-*a*) but it does not influence A/C ratio of the spent liquor (Fig. 5-*b*). Thus, the original parameters were applied.



a b

Fig. 5. The importance of original parameters and principal components in relation of: a) -45 μm content; b) spent liquor ratio

2.4. Model data structure

Regardless of the architecture of a Recurrent Neural Network the structure of the model data is the same. The example in Fig. 6 shows 10 archive data points for 5 parameters as inlet data of the network (5x10 array highlighted in green), assumptions for expected values for 3 parameters in 5 data points (3x5 array highlighted in blue) and data array 2x5 (highlighted in red) indicating the forecast data for two outlet parameters for 5 downstream data points.

Target parameters that the user wants to predict using the model are output parameters. Such output parameters include the following: A/C ratio of the spent liquor indicating the liquor productivity (*A_C*); the content of the size grades used for product certification (in our case it is *F45*), the content of the coarse crystals that might cause slurry segregation in the precipitators (*F150*). The parameters that are used for the process control are input parameters for the neural network. In the present case the set of control parameters comprises the flow rate of the pregnant liquor (*Q*), temperature in the first and final precipitators (*T1* and *Tn*), solids content in the first precipitator (*Cs*). The division of the remaining parameters into input and output ones is a key aspect of network structure development. Other parameters might contain information on disturbances of the process and serve as additional features characterizing the process development and having significance for the forecast. If the assumption of the future values of such factor can be obtained from the user or outside model, it should be used as

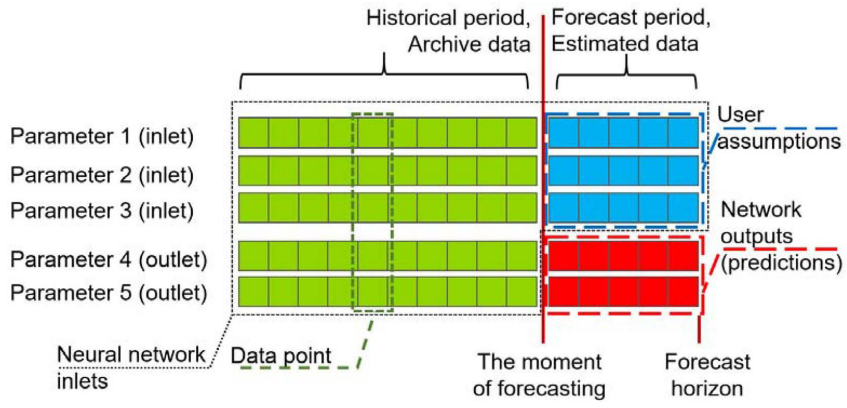


Fig. 6. Structure of Recurrent Neural Network data

an input parameter. If the assumption of the future values of such factor is unavailable and the said feature is a secondary one and has a strong correlation with output parameters of the developed model, it should be treated as an output parameter. In this study such additional input parameter for the model is the chemical composition of the pregnant liquor (*A*, *C*, *S*), and non-target size grades of the crystals (*F5*, *F10*, *F20*, *F30*, *F50*, *F63*, *F100*, *F125*) are referred as output parameters of the model.

The data set for each precipitation area is divided into training and test samples in equal proportion 0.8/0.2. The resulted data samples were equal for all three precipitation areas, i.e., 1671 batches in the training data sample and 418 batches in the test data sample. It is important to avoid perturbation of the batches prior to the division as in such a case the data over the same periods with different offset are included into the both data samples, that results in significant overestimating of training quality. In the present study, the test data are indicated to the right from the training data on the time axis so their intersection is impossible. Besides, the model cannot learn the changes of the simulated object over the latest period of time, i.e., last 418 days in this case.

2.5. Performance indices

To assess the quality of trained neural networks, two standard performance metrics were used, i.e., mean absolute deviation (MAE) and root mean square error (RMSE):

$$MAE(X,h)=\frac{1}{N}\sum_{i=1}^{N}\left|h(x_i)-y_i\right|; \tag{2}$$

$$RMSE(X,h)=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(h(x_i)-y_i\right)^2}, \tag{3}$$

where X is a matrix comprising values of all features for all records in the initial data; h is a model prediction function; N is a number of training or test cases; x_i is a vector of values of all features of *i-th* records; y_i is a target value in the *i-th* record.

RMSE metric was used at this training stage (calculation of weight coefficients of the neural network) as it computes large weights for large errors, and MAE metric was used for information.

3. Selection of Network Architecture

3.1. General settings

For this study the data from three areas of seeded crystallization are available. The set of parameters and the amount of these three data samples are roughly the same so the network architecture was selected with the use of the data from area No. 6 but the resulted conclusions will be applicable to any of three areas.

In the present study *Keras* deep learning library was used for the development of the models [15, 41]. Out of its standard methods for forecasting the multivariate time series of continuous crystallization, the multilayer fully connected MLP-based networks, multilayer LSTM networks, and combinations of CNN and TCN convolutional layers with LSTM layers are considered to be the most prospective. Fig. 7 shows the architectures of the said models.

The amount of data and complexity of dependences are not too large, so the number of main hidden layers are limited to two. In the structure with two networks an addition hidden layer, i.e., convolution layer is used before LSTM layers. *DropOut* regularizers are used after each hidden layer.

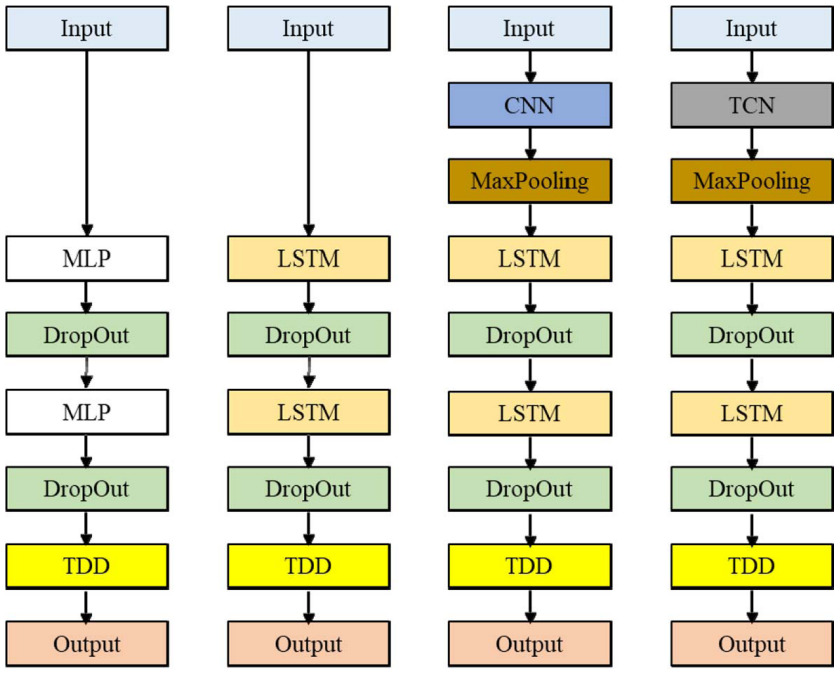


Fig. 7. Architectures of deep learning neural networks: *a* – double layer MLP, *b* – double layer LSTM, *c* – normal convolution layer and two LSTM layers; *d* – temporary convolution layer and two LSTM layers

TDD output layer is of *Dense* type with linear activation function; its data are vectorized with the use of *TimeDistribution*. The *RMSprop* optimizer is utilized within this model with a fixed learning rate of 0.001 and batch size of 128.

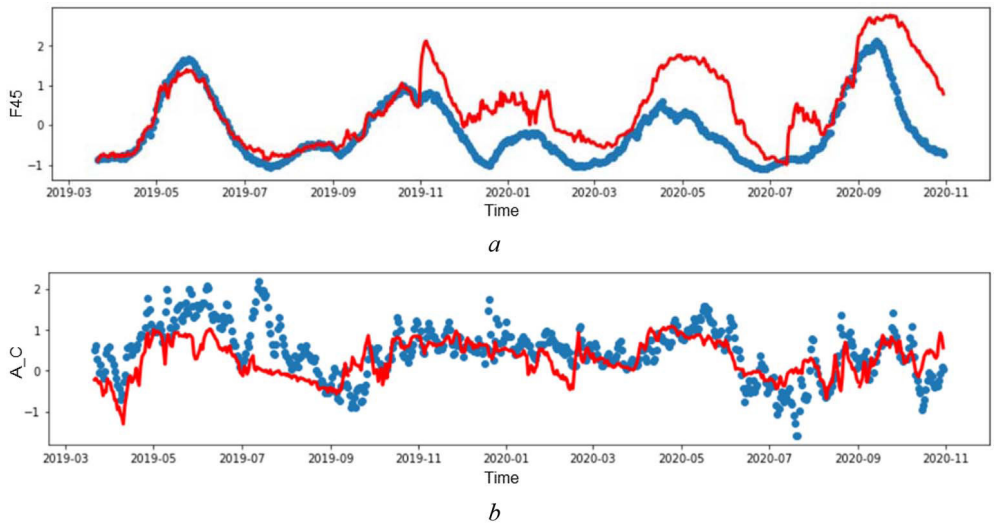
3.2. Fully connected models

It is advisable to begin the study of benefits of deep learning networks as compared with one-layer networks and applicability of advanced network architectures to predict the multiparameter time series using the available data with development of classical fully connected models based on multilayer perceptron. In the present paper the quality of fully connected models is assessed for MLP networks with different number of layers and 32 neurons on each layer as per the PFD in Fig. 7-*a*. Table 1 presents the results of comparison of the quality of prediction made by these networks for two key parameters of precipitation area No. 6, i.e., -45 μm content in the slurry of the final precipitators (*F45*) and A/C ratio in the spent liquor (*A_C*).

All tests on training MPL-based networks proved the overfitting of such networks because as the number of epochs increases the error for the training sample decreases and the error for the test sample increases or remains at the constant level. Use of regularization affects the rate of learning but it does not improve the generalization quality. It should be noted that by the end of 2019 a new precipitation train had been put into operation in precipitation area No. 6 so, the capacity of the area increased. These data were not used in the training sample that resulted in the sharp decrease of the forecast quality for *F45* parameter (Fig. 8). So, MPL-based networks are capable to predict

Table 1. Quality of fully connected machine learning models

Number of layers	Number of trainable parameters	Performance index	Training sample		Test sample	
			F45	A_C	F45	A_C
1	1483	MAE	2.026	0.040	6.371	0.095
		RMSE	2.628	0.050	5.302	0.075
2	1864	MAE	1.527	0.033	6.154	0.096
		RMSE	1.979	0.041	5.059	0.077
3	2920	MAE	1.947	0.038	6.760	0.072
		RMSE	1.594	0.091	5.514	0.030



a

b

Fig. 8. Results of the forecast of the crystallization with the use of double layer MPL networks (normalized values): *a* – *F45*; *b* – *A_C*

qualitatively the tendency towards the increasing or decreasing of the crystal size but they are not very accurate in predicting the values. However, the model showed good results for extrapolation of *A_C* time series.

3.3. LSTM-based models

LSTM networks enable to study and simulate horizontal links in the time series, moreover due to the built-in filter, they allow managing the depth of long-term dependencies. This feature provides for better learning under conditions of noise, reduces their sensitivity to the changes in a simulated object over the time and thus improves their generalization ability.

An important step in LSTM setting is the selection of a sufficient number of cells and regularization level to provide for steady learning. To determine the optimum number of neurons and epochs of learning, the numerical experiment with a double layer LSTM network (Fig. 7-*b* shows the network structure) was performed. Under otherwise equal conditions 10, 20, 30, 40, and 50 neurons were placed

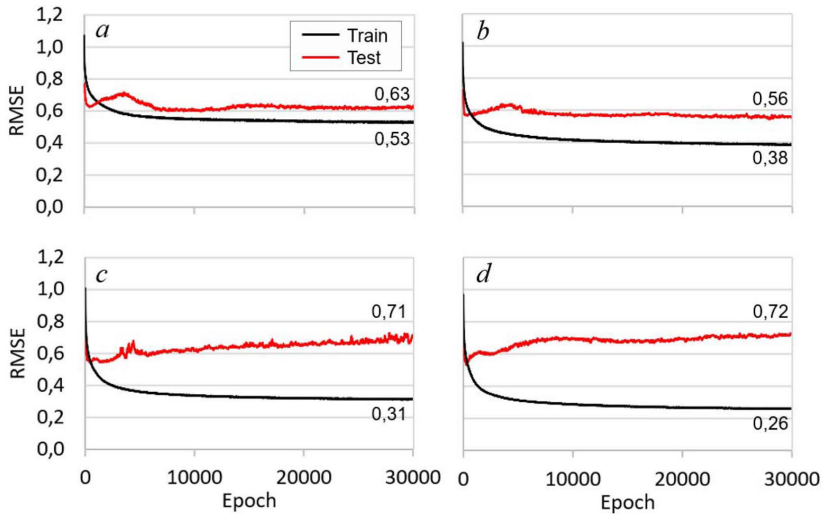


Fig. 9. Changes of the error in course of training of double layer LSTM networks: *a, b, c, and d* are the number of LSTM cells on each layer, i.e., 10, 20, 30, and 40 respectively

on the both hidden layers and taught without tolerance for 30 000 epochs. For the training the data from precipitation area No. 6 were used.

Fig. 9 shows the change of an error for training and validation samples during the training process. In all cases the error for the validation sample decreases during 150-200 epochs, then overfitting was observed. Deep double descent that sometimes occur during the training of deep networks [42, 43] was not observed in our case. The critical point that can be seen in Figures 9-*a*, 9-*b* within 1000-2000 epochs disappears if different regularization coefficients are used for the training; the plateau after this point is not always better than the first incline minimum. Consequently, it is practical to limit the learning to 200 epochs with 20 epoch tolerance.

Based on the training results the network with 30 neurons on the both hidden layers (13 208 learned parameters) showed the best training performance, i.e., it achieved RMSE error of 0.53 for the validation sample at 175 epochs. The networks with 10 and 20 neurons on their hidden layers (2008 and 6408 parameters respectively) demonstrated somewhat lower forecasting performance for F45 and *A_C* (Table 2), due to insufficient number of configurable parameters. The networks with 40 and 50 neurons (22 408 and 34 008 parameters respectively) overfitted quickly and were less accurate, though they showed a lower error for A/C ratio in the spent liquor (*A_C).*

3.4. Combination of CNN-LSTM and TCN-LSTM

Based on the theory the observed variability of the simulated time series of the content of particle size grades in the population is defined by the regulation law of population balance that describes the main mode of oscillations [6] and the alterations of this mode for *A_C* parameter are more dynamic along with the change of the supersaturation of the liquor and phase interface area. Other disturbances are of occasional nature. The use of CNN convolutional layers combined with LSTM might facilitate the detection of regular dependences in the data and the use of a TCN layer might help to prevent the diffusion of these dependences that inevitably occurs due to gradual change of the equipment in the continuous crystallization area over the years of observations.

Table 2. Change of the quality of a LSTM neural network depending on the number of neurons (Fig. 7-*b*, precipitation area No. 6) - MAE index values.

Number of neurons	10		20		30		40		50	
Sample	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Content of size grade, %:										
-10	0.24	0.24	0.23	0.24	0.22	0.23	0.22	0.25	0.21	0.23
-20	0.90	0.73	0.81	0.81	0.74	0.70	0.72	0.54	0.62	0.82
-30	2.39	1.95	1.97	2.16	1.80	1.84	1.79	1.54	1.55	2.70
-45	5.58	4.62	4.60	5.13	4.19	4.57	4.23	4.94	3.77	7.30
-63	8.07	6.96	6.98	7.45	6.25	6.47	6.20	8.33	5.28	9.70
-100	6.53	6.94	5.38	6.17	5.10	5.39	4.83	6.44	3.66	5.11
+150	2.13	2.24	1.74	1.95	1.64	1.68	1.58	1.99	1.20	1.51
A/C ratio in the spent liquor	0.103	0.087	0.090	0.071	0.087	0.069	0.084	0.061	0.078	0.061

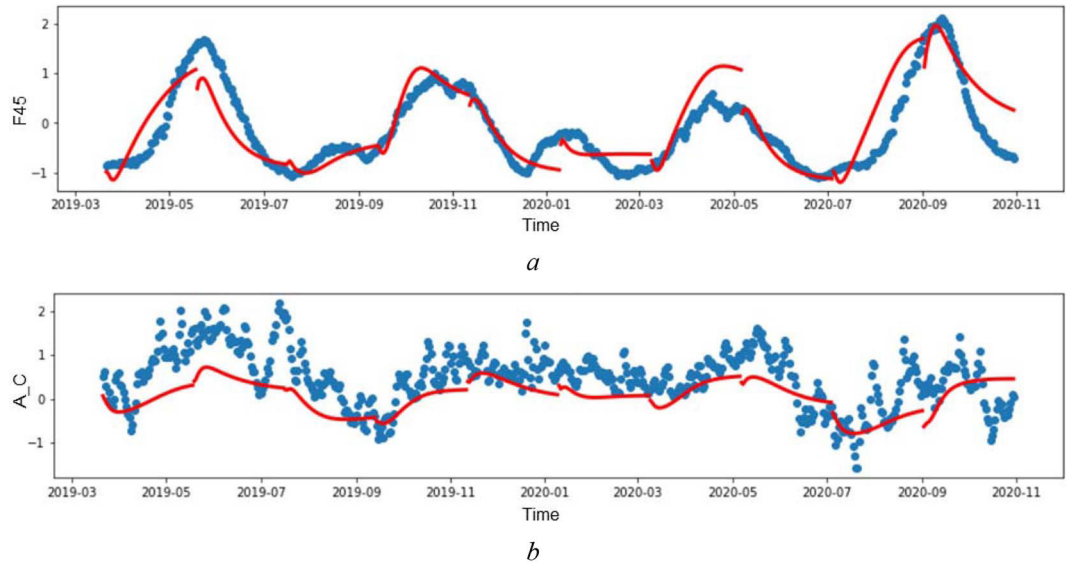
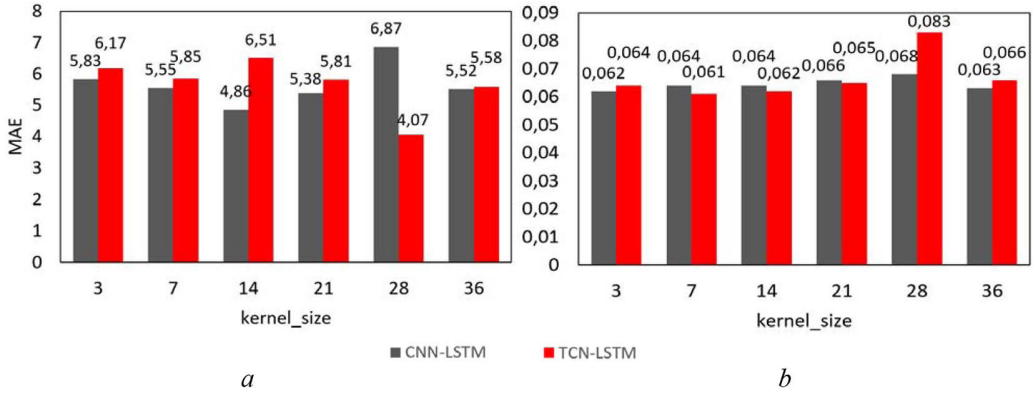


Fig. 10. Results of the forecast of the crystallization with the use of double layer LSTM networks: *a – F45*; *b – A_C*

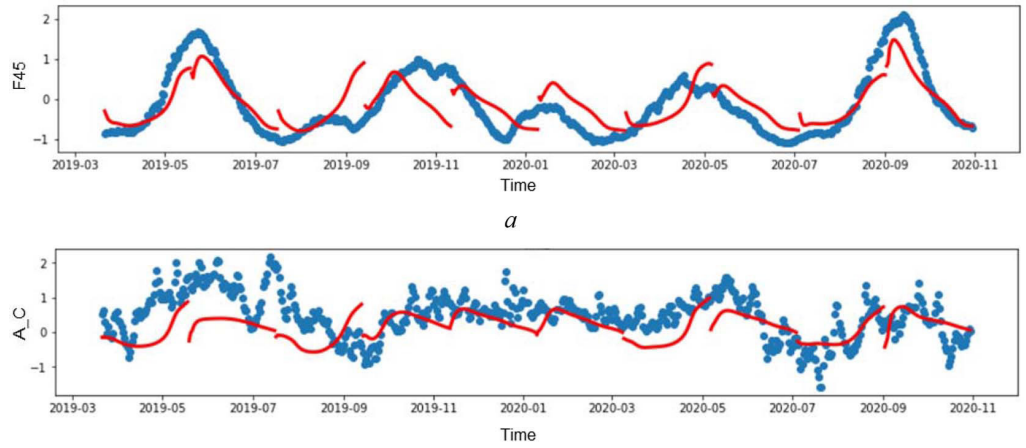
In the analyzed models CNN and TCN layers were used successively with LSTM layers (Figures 7-*c* and 7-*d*). The number of neurons was accepted as 50 on the convolutional layer and 10 on each of LSTM layers. Kernel size was measured within the range from 3 to 36 days. The number of trainable parameters in the developed models amounted to 10 000-90 000. Upon testing of the training rate using the procedure similar to the one used for LSTM networks, the training was conducted at 2000 epochs with the tolerance of 200 epochs.

Numerical experiments proved that the use of convolution is efficient only on the first hidden layer. The use of two hidden layers combined with one LSTM layer degraded the quality of the network. The



a *b*

Fig. 11. Results of the training of the models based on CNN-LSTM and TCN-LSTM networks (Fig. *7-c* and *7-d*) at various kernel sizes: *a* – forecast error of *F45*, *b* – forecast error of *A_C*



a

b

Fig. 12. Results of the forecast of the crystallization with the use of triple layer TCN-LSTM networks: *a* – *F45*; *b* – *A_C*

tests of the kernel size showed that the residual error of the forecasting is randomly distributed within kernel size range from 3 to 36 for both classical and temporal convolution layer (Fig. 11). The quality of the forecasts obtained for one of the best TCN-LSTM networks can be seen in Fig. 12.

3.5. Discussion

The present study proved the possibility to generate medium-range forecast for the dynamics of continuous seeded crystallization with the duration of approximately half of PSD oscillation period, i.e., 2 months in our case. The optimum number of hidden layers for such network amounts to 2 or 3 depending on the learning data volume and specifics of a simulation object. Table 3 presents one model of each type demonstrating the best result for the test samples.

The fully connected models showed good results for the prediction of A/C ratio in the spent liquor. The residence time of the pregnant liquor in the precipitation is approximately 40 hours, i.e., its composition

Table 3. Results of the training of neural candidate-networks. MAE of the test sample.

Network	Number of trainable parameters	*F45*	*A_C*	Options
MLP (Fig. 7–*a*)	1 864	6.15	0.096	32 neurons on the layer
LSTM (Fig. 7–*b*)	13 208	4.57	0.069	30 neurons on the layer
CNN-LSTM (Fig. 7–*c*)	72 378	4.86	0.064	kernel_size=14
TCN-LSTM (Fig. 7–*d*)	50 426	4.07	0.083	kernel_size=28

is influenced by the data for the previous two days. As the deep recursion in the archive data is not required for *A_C* calculation and the influence of the fluctuations of seed coarseness on *A_C* is significantly less as compared with the composition and temperature of the liquor (non-predictable parameters), so fully connected networks show better performance for *A_C* as compared with other models.

For prediction of the particle size distribution LSTM, CNN-LSTM and TCN-LSTM networks showed almost the same accuracy. The table presents only the best results but the analysis of the best 10 options does not reveal one top option in terms of generalization quality. In terms of the learning rate LSTM networks show the best performance as their training required fewer parameters and significantly fewer epochs. So, the use of layers of normal and temporal convolution layers does not improve the result, probably due to the lack of short regular patterns in the data.

Numerical experiments with double layer LSTM networks enable to determine that the optimum number of network parameters that do not promote the overfitting ranges from 2 to 20 000. Besides, double layer LSTM networks are the most prospective ones in terms of improvement, as they demonstrate the best balance between the accuracy and computational efforts.

4. Optimization of Network Structure

Further optimization of LSTM network structure consists of determination of a number and type of neurons on separate layers for qualitative representation of non-linear connections and depth of archive data required for efficient extrapolation for the specified forecast horizon in all events recorded in the training sample.

To optimize the network configuration, the stochastic search was used. The following parameters were specified in a random manner: number of LSTM neurons on each layer ranging from 10 to 50; type of activation function (sigmoid, tanh, relu); number of inputs ranging from 60 to 200. The variants with the number of adjustable parameters less than 3 000 and more than 20 000 were eliminated without training as not promising. The candidate-networks were trained at 200 epochs, the sensitivity to the lack of improvement (tolerance) was equal to 20 epochs. The best network was determined using RMSE assessment. A small number of epochs enabled to analyze several dozens of neural network configurations for each of three areas of continuous crystallization.

Table 4 shows the structure of the best double layer LSTM models trained using the initial data from three different precipitation areas of the Urals alumina refinery. Table 5 presents the results of testing of these models.

The analysis of the structure of the best models in Table 4 shows one common feature, i.e., depth of the historical data required for the forecasting ranges from 90 to 122 days; the models that used

Table 4. Structure of deep learning models for precipitation areas at the Urals alumina refinery

Area No.	3	6	10
Input layer dimensions	122x8	100x8	111x8
The 1st LSTM layer parameters:			
Number of neurons	8	30	20
Activation function	sigmoid	sigmoid	sigmoid
Dropout	0.014	0.15	0.1
Recurrent Dropout	0.44	0.050	0.38
The 2nd LSTM layer parameters:			
Number of neurons	20	30	15
Activation function	tanh	sigmoid	tanh
Dropout	0.014	0.15	0.1
Recurrent Dropout	0.44	0.05	0.38
TDD output layer dimensions	59x8	59x8	59x8
Output layer activation function	linear	linear	linear
Number of trainable parameters			
1st LSTM layer	832	5640	3040
2nd LSTM layer	2320	7320	2160
TDD output layer	168	248	128
Total	3320	13208	5328

Table 5. MAE index of the precipitation neural networks

Area No.	3		6		10	
Sample	Train	Test	Train	Test	Train	Test
Size grade content, %:						
-10	0.22	0.31	0.22	0.23	0.36	0.35
-20	0.80	1.32	0.74	0.70	1.37	1.54
-30	2.08	2.90	1.80	1.84	2.79	3.16
-45	4.47	4.77	4.19	4.57	4.45	4.99
-63	5.86	5.47	6.25	6.47	5.08	7.29
-100	3.95	4.25	5.10	5.39	3.44	5.67
+150	1.13	1.50	1.64	1.68	1.09	1.59
A/C ratio in the spent liquor	0.058	0.062	0.087	0.069	0.083	0.078

72 or more archive points demonstrated the similar quality. However, during the training none of the selected networks had the depth of the archive data less than the forecast horizon. Nothing specific was detected in relation to the type or number of neurons on the specific layers.

The quality of the model forecasts for the precipitation areas proved to be rather close (Table 5). In all three cases mean arithmetic error for *F45* was less than 5%, but in no case this MAE value of

less than 4.5% was achieved; MAE for *A_C* parameter ranges from 0.062 to 0.078. Such constant value most likely indicates that further improvement of the model quality by optimization of neural network structure is not possible.

Conclusions

The present study has established that deep neural networks provide for a medium-range forecast of changes in PSD and liquor decomposition degree during the continuous seeded crystallization. In the present paper three main types of the neural networks that are the best known and approved in the academic circles have been studied and tested to forecast the multivariate time series, i.e., LSTM, CNN, and TCN. The use of convolution layers has not resulted in any significant improvement of the models, so the classical double layer LSTM model is considered to be the most suitable model for solution of this specified task.

Preprocessing of the initial data from three areas of Bayer precipitation circuit, determination and selection of network architecture, multifactor optimization of hidden layer structure and training settings allowed developing the models based on the multilayer neural networks with long short-term memory that ensure low error for the forecast horizon of 2 months or one half of PSD change cycle. It was sufficient under the conditions of the Urals alumina refinery and enabled to use the generated models for optimal multivariable control of continuous seeded crystallization.

All three predictive models are of similar quality without any prospects for further improvement by optimization of their structure. Enhancement of the model quality will require more comprehensive understanding of continuous seeded crystallization and reduction of noise of the measurements. Additional information can be obtained from the use of alternative observation methods based on other physical methods, i.e., crystal micrographs of different zoom factors (as inputs for convolutional network); results of intermediate processing of such micrographs (particle shape, fractality, agglomerate fraction); particle quantity data calculated with the use of Coulter counter. However, these data can be used only if they are collected/measured on a regular basis.

References

[1] Li M., Wu Y. Dynamic simulation of periodic attenuation in seeded precipitation of sodium aluminate solution. *Hydrometallurgy*, 2012, 113-114, 91–97.

[2] Bekker A.V., Li T.S., Livk I. Dynamic response of a plant-scale gibbsite precipitation circuit *Hydrometallurgy*, 2016, 170, 24-33.

[3] Bekker A.V., Li T.S., Livk I. Understanding oscillatory behavior of gibbsite precipitation circuits, *Chemical engineering research and design*,2015, 101, 113-124.

[4] Raahauge B.E. Smelter Grade Alumina Quality in 40+ Year Perspective: Where to from Here? *Light Metals*, 2015, 73-78.

[5] Ramkrishna D. *Population balances. Theory and Applications to Particulate Systems in Engineering*. London: Academic Press, 2000,p. 355.

[6] Golubev V.O., Litvinova T.E. Dynamical simulation of industrial scale gibbsite crystallization. *Journal of Mining Institute*, to appear.

[7] Balakin B.S., Hoffmann A.C., Kosinski P. Population Balance Model for Nucleation, Growth, Aggregation, and Breakage of Hydrate Particles in Turbulent Flow, *AIChE Journal*, 2010, 56(8), 2052-2062.

[8] Farhadi F., Babaheidary M.B. Mechanism and estimation of Al(OH)3 crystal growth, *Journal of Crystal Growth,* 2002,234, 721-730.

[9] Ilievski D. Development and application of a constant supersaturation, semi-batch crystallizer for investigating gibbsite agglomeration, *Journal of Crystal Growth*, 2001, 233, 846-862.

[10] Salman A. G., Kanigoro B., Heryadi Y. Weather forecasting using deep learning techniques, *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2015, 281-285.

[11] Tsay R.S. *Multivariate time series analysis: with R and financial applications*. Chicago: School of Business, 2014, 502 p. ISBN 978-1-118-61790-8.

[12] Long J. Sun Z., Pardalos P.M. et al. A robust dynamic scheduling approach based on release time series forecasting for the steelmaking-continuous casting production, *Applied Soft Computing Journal*, 2020, 92, 1-17.

[13] Petropoulos F., Makridakis S., Stylianou N. COVID-19: Forecasting confirmed cases and deaths with a simple time-series model, *International Journal of Forecasting,* 2020, to appear.

[14] Wei W.W.S. *Multivariate Time Series Analysis and Applications*. Oxford: Wiley, 2019, 518 p. ISBN: 978-1-119-50285-2.

[15] Chollet F. *Deep Learning with Python*. New York: Manning Publications Co, 2018, 440 p.

[16] Reinsel G.C. *Vector ARMA Time Series Models and Forecasting. Elements of Multivariate Time Series Analysis*, 1993, 21-51.

[17] Almqvist O. *A comparative study between algorithms for time series forecasting on customer prediction*. Thesis. University of Skövde, Sweden, June, 2019.

[18] Hastie T., Tibshirani R., Friedman J. *The Elements of Statistical Learning. Data Mining, Inference and Prediction. Second edition*. Springer, 2017, 764 p.

[19] Breiman L. Random Forest, *Machine Learning*, 2001, 45(1), 5-32.

[20] Zhang G.P. *Neural Networks for Time-Series Forecasting. Handbook of Natural Computing*, 2012, 461-477.

[21] Shiblee M., Kalra P.K., Chandra B. Time Series Prediction with Multilayer Perceptron (MLP): A New Generalized Error Based Approach. *International Conference on Neural Information Processing*, 2008, 37-44.

[22] Borovikov V., Milkov M. Statistical Analysis of Aluminate Liquor Precipitation Process with Statistica: Classic and Modern Data Mining Methods. *TRAVAUX 48, Proceedings of the 37th International ICSOBA Conference and XXV Conference «Aluminum of Siberia»*, Krasnoyarsk, Russia, 16-20 September, 2019, 295-303.

[23] Graves A., Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 2005, 18(5-6), 602-610.

[24] Hochreiter S., Schmidhuber J. Long Short-Term Memory, *Neural Computation,* 1997, 9(8), 1735-1780. doi: https://doi.org/10.1162/neco.1997.9.8.1735

[25] Zhang N., Shen S.-L., Zhou A., Jin Y.-F. Application of LSTM approach for modelling stress–strain behavior of Soil. *Applied Soft Computing Journal*,2021, 100, 1-11.

[26] Fang Z., Wang Y., Peng L., Hong H., Predicting flood susceptibility using longshort-term memory (LSTM) neural network model, *Journal of Hydrology*, 2020, 1-56.

[27] Farah S., Zameer A., Muneeb M. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons and Fractals* 2020,140, 1-9.

[28] Yan H., Qin Y., Xiang S., Wang Y. Chen H. Long-term gear life prediction based on ordered neurons LSTM neural networks. *Measurement*, 2020, 165, 1-11.

[29] Yan J., Mu L., Ranjan R., Zomaya A.Y. Temporal Convolutional Networks for the Advance Prediction of ENSO. *Scientific reports*, 2020, 10:8055. 1-15.

[30] LeCun Y., Boser B., Denker J.S., Henderson D., Howard R.E., Hubbard W., Jackel L.D.: Backpropagation Applied to Handwritten Zip Code Recognition, *Neural Computation*,1989, 1(4), 541-551.

[31] Yann L., Bottou L., Bengio Y., Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1988, 86(11), 2278–2324.

[32] Brownlee J. *Deep Learning for Time Series Forecasting. Predict the Future with MLPs, CNNs and LSTMs in Python* / Machine Learning Mastery, 2019, 555 p.

[33] Lea C., Flynn M.D., Vidal R. et al. Temporal Convolutional Networks for Action Segmentation and Detection. *arXiv*:1611.05267v1, 2016, 1-10.

[34] Bai S., Kolter J.Z., Koltun V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling.*arXiv*:1803.01271v2, 2018, 1-10.

[35] Smiti A. A critical overview of outlier detection methods. *Computer Science Review*, 2020, 38.

[36] Ramaswamy S.; Rastogi R.; Shim K. Efficient algorithms for mining outliers from large data sets. *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*.2000, p. 427. ISBN 1-58113-217-4.

[37] Gu X., Akoglu L., Rinaldo A. Statistical Analysis of Nearest Neighbor Methods for Anomaly Detection, *33rd Conference on Neural Information Processing Systems*, *Vancouver*, 2019, 1-11.

[38] Jackson A.C., Lacey S. The discrete Fourier transformation for seasonality and anomaly detection of an application to rare data, *Data Technologies and Applications,*2020, 54(2), 121-132.

[39] Plazas-Nossa L., Avila A., Miguel A., Torres A. Detection of Outliers and Imputing of Missing Values for Water Quality UV-VIS Absorbance Time Series, *Ingeniería*, 2017, 22(1),111-124.

[40] Scholz M., Fraunholz M., Selbig J. Nonlinear Principal Component Analysis: Neural Network Models and Applications. *Principal Manifolds for Data Visualization and Dimension Reduction*, 2008, 44–67.

[41] Keras. Simple. Flexible. Powerful. [Electronic resource] – Access: https://keras.io/guides/

[42] Belkin M. Hsu D., Xu J. Two Models of Double Descent for Weak Features / M. Belkin, *SIAM Journal on Mathematics of Data Science*, 2020, 2(4),1167-1180.

[43] Nakkiran P., Kaplun G., Bansal Y. et al. Deep Double Descent: Where Bigger Models and More Data Hurt / P. Nakkiran, *arXiv*:1912.02292, 2019, 1-24.