

Studies in Systems, Decision and Control 417

Alla G. Kravets  
Alexander A. Bolshakov  
Maxim Shcherbakov *Editors*

# Cyber-Physical Systems: Intelligent Models and Algorithms

 Springer

# **Studies in Systems, Decision and Control**

Volume 417

## **Series Editor**

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences,  
Warsaw, Poland

The series “Studies in Systems, Decision and Control” (SSDC) covers both new developments and advances, as well as the state of the art, in the various areas of broadly perceived systems, decision making and control—quickly, up to date and with a high quality. The intent is to cover the theory, applications, and perspectives on the state of the art and future developments relevant to systems, decision making, control, complex processes and related areas, as embedded in the fields of engineering, computer science, physics, economics, social and life sciences, as well as the paradigms and methodologies behind them. The series contains monographs, textbooks, lecture notes and edited volumes in systems, decision making and control spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

Indexed by SCOPUS, DBLP, WTI Frankfurt eG, zbMATH, SCImago.

All books published in the series are submitted for consideration in Web of Science.

More information about this series at <https://link.springer.com/bookseries/13304>

Alla G. Kravets · Alexander A. Bolshakov ·  
Maxim Shcherbakov  
Editors

# Cyber-Physical Systems: Intelligent Models and Algorithms

 Springer

*Editors*

Alla G. Kravets  
Volograd State Technical University  
Volograd, Russia

Maxim Shcherbakov   
Volograd State Technical University  
Volograd, Russia

Alexander A. Bolshakov  
Peter the Great St. Petersburg Polytechnic  
University  
St. Petersburg, Russia

ISSN 2198-4182

ISSN 2198-4190 (electronic)

Studies in Systems, Decision and Control

ISBN 978-3-030-95115-3

ISBN 978-3-030-95116-0 (eBook)

<https://doi.org/10.1007/978-3-030-95116-0>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

# Preface

Intelligent models and algorithms of their implementation are the core components of cyber-physical systems. The complexity of developing and deploying cyber-physical systems requires new approaches for model design. This book dwells with results in the modelling technology field for the exploitation of artificial intelligence, including artificial general intelligence (AGI) and weak artificial intelligence. The book continues a series of articles on the topic of modelling cyber-physical systems.

The book has four sections, including chapters with the results of original research on the proposed topic.

In the first section, the authors offer original results by using bio-inspired methods to improve cyber-physical systems operation. The authors study the activities that allow for developing new and perspective concepts of bio-inspired modelling combining various artificial neural networks, machine learning methods, uncertainty explanation approaches, biocomputation technologies including machine learning survival models and artificial immune systems by addressing a large scale of applications. Those methods suggest an extension of the neural additive model, artificial immune systems. The book also describes fuzzy models and algorithms as well as the implementation of these technologies in different domains such as cyber-physical management, management of security information and events, and advanced search machines. The section Predictive Modelling contains chapters dedicated to improving approaches for the remaining useful lifetime prediction, predictive maintenance of robotic agricultural vehicles, and predicting unstable stages of cyber-physical systems. The Computer Vision and Image Processing section includes chapters for enhanced computer vision approach for monitoring states of cyber-physical systems, explores vehicle trajectory anomalies, and analyses the content of business documents.

This book is directed to researchers, practitioners, engineers, software developers, professors, and students. We do hope the book will be useful for them.

This edition of the book is dedicated to 2021, the Year of Science and Technology in Russia, and technically supported by the Project Laboratory of Cyber-Physical Systems of Volgograd State Technical University.

Volgograd, Russia  
St. Petersburg, Russia  
Volgograd, Russia  
October 2021

Alla G. Kravets  
Alexander A. Bolshakov  
Maxim Shcherbakov

# Contents

## Bio-inspired Modelling

**An Extension of the Neural Additive Model for Uncertainty  
Explanation of Machine Learning Survival Models** ..... 3  
Lev Utkin and Andrei Konstantinov

**Biocomputation Using Molecular Agents Moving in Microfluidic  
Channel Networks: An Alternative Platform for Information  
Technology** ..... 15  
Thomas Blaudeck, Christoph R. Meinecke, Danny Reuter,  
Sönke Steenhusen, Archa Jain, Sascha Hermann, Stefan E. Schulz,  
Eduard I. Zenkevich, Till Korten, and Heiner Linke

**Artificial Immune Systems—Models and Applications** ..... 29  
Yuriy Skobtsov

**The Analysis of Technology Development Trends Based  
on the Network Semantic Structure “Subject-Action-Object”** ..... 43  
Veronika Kolesnikova, Dmitriy Korobkin, Sergey Fomenkov,  
Eduard Rayushkin, and Vsevolod Glushkin

**Machine Learning of Diagnostic Neural Network for Railway  
Track Monitoring** ..... 55  
Sergey Orlov, Antonina Piletskaya, Nadezhda Kusakina,  
and Andrey Tyugashev

**Neural Network Analysis of Electromagnetic Field Effect  
on the Thermofluctuation Characteristics Measurement  
in the Power Cable Insulating Materials** ..... 67  
N. K. Poluyanovich, M. N. Dubyago, N. V. Azarov,  
and A. V. Ogrenichev



## **Fuzzy Models and Algorithms**

<b>Algorithm for Configuring Sugeno-Type Fuzzy Inference Systems Based on the Nearest Neighbor Method for Use in Cyber-Physical Systems</b> .....	83
Mikhail Golosovskiy, Alexey Bogomolov, and Mikhail Balandov	
<b>Construction of Membership Functions for Fuzzy Management of Security Information and Events</b> .....	99
Igor Kotenko and Igor Parashchuk	
<b>Building a Fuzzy Model to Automate the Search for Product Names in Procurement Documentation for a Shipbuilding Plant</b> .....	111
Lilia Emaletdinova, Aigul Kabirova, and Ivan Kulagin	
<b>Synthesis of a Neuro-Fuzzy Model in the Structure of an Automatic Rolling Stock Control System</b> .....	121
Konstantin Yurenko, Pavel Kharchenko, Evgeniy Fandeev, and Ivan Yurenko	
<b>Predictive Modelling</b>	
<b>Predicting the Equipment Useful Lifetime Based on the Deep Neural Networks</b> .....	135
Maxim Dli, Andrey Puchkov, and Ekaterina Lobaneva	
<b>Operational Model Predictive Control on the Example of the Stabilization Process for Hydro Treatment of Oil Fractions</b> .....	145
I. V. Gogol, N. A. Kalashnikov, O. A. Remizova, V. V. Syrokvashin, and A. L. Fokin	
<b>The Models for Predictive Maintenance of Robotic Agricultural Vehicles</b> .....	157
Sergey Susarev, Sergey Orlov, Elizaveta Bizyukova, and Roman Uchaikin	
<b>Development of a Cyber-Physical System for Neurofuzzy Prediction of the Concentration of the Contained Prime During Transportation of Oil Wells Emulsion</b> .....	169
Artur Sagdatullin and Gennady Degtyarev	
<b>Non-life Insurance Reserve Prediction Using LightGBM Classification and Regression Models Ensemble</b> .....	181
Vladimir Soloviev and Vadim Feklin	
<b>Machine Learning-Based Cyber-Physical Systems for Forecasting Short-Term State of Unstable Systems</b> .....	189
Alexander Musaev and Dmitry Grigoriev	

**Computer Vision and Image Processing**

<b>Application of Computer Vision Tools to Create a System for Monitoring the Work of Ground Equipment in Open Pits of Gold Mining Enterprises</b> .....	203
--	-----

Anna Alekhina, Alexandr Gurenko, and Mikhail Dorrer

<b>Framework for Biometric User Authentication Based on a Dynamic Handwritten Signature</b> .....	219
---	-----

Igor Anikin and Ellina Anisimova

<b>Image Compression Based on the Significance Analysis of the Wavelet Transform Coefficients Using the Energy Feature Model</b> .....	233
--	-----

Maya M. Lyasheva, Stella A. Lyasheva, and Mikhail P. Shleymovich

<b>Construction of a Fuzzy Model for Contour Selection</b> .....	245
--	-----

L. Y. Emaletdinova and M. A. Nazarov

<b>Identification of Vehicle Trajectory Anomalies on Streaming Video</b> .....	255
--	-----

Igor Anikin and Aigul Mardanova

<b>Analyzing the Content of Business Documents Recognized with a Large Number of Errors Using Modified Levenshtein Distance</b> .....	267
---	-----

Oleg Slavin, Vera Farsobina, and Aleksei Myshev

# **Bio-inspired Modelling**

# An Extension of the Neural Additive Model for Uncertainty Explanation of Machine Learning Survival Models



Lev Utkin  and Andrei Konstantinov 

**Abstract** A method for the local and global uncertainty explanations of machine learning survival models by censored data in the framework of survival analysis is proposed. The method aims to select features of an object, which significantly impact on uncertainty of a survival model prediction corresponding to the object. The survival model is viewed as the black box, that is, only the input and the corresponding output are known. The first basic idea behind the method is to approximate the prediction uncertainty by the prediction uncertainty of the extended semi-parametric Cox proportional hazards model using the well-known generalized additive model. The second idea is to apply the neural additive model for getting the Cox approximation and to train the network in accordance with a specific loss function that takes into account the peculiarities of survival models. Results of the uncertainty explanation are represented in the form of shape functions which show the impact or contribution of every feature on the prediction provided by the black-box model. An algorithm implementing the method is presented. Numerical examples illustrate the proposed method by using real datasets including the German Breast Cancer Study Group 2 dataset and the Monoclonal Gammopathy dataset.

**Keywords** Interpretable model · XAI · Survival analysis · Generalized additive model · Cox model · Neural network · Uncertainty

## 1 Introduction

Machine learning models are successfully applied to survival analysis [1, 2] which is an important part of statistics dealing with censored and uncensored data in order to predict when an event of interest will happen. In survival analysis, there is a training set  $Z$  consisting of  $n$  instances  $(\mathbf{x}_i, \delta_i, T_i)$ ,  $i = 1, \dots, n$ , for example, patients, where  $\mathbf{x}_i^T = (x_{i1}, \dots, x_{im})$  is the instance feature vector;  $T_i$  is time to event of interest of the instance. If the event is observed at time  $T_i$ , then we have an uncensored observation,

---

L. Utkin (✉) · A. Konstantinov  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

and  $\delta_i = 1$ . If the event is not observed at  $T_i$ , then we have a censored observation which means that  $\delta_i = 0$ . Survival analysis aims to estimate probabilistic measures of time to the event of interest for a new instance with a feature vector  $\mathbf{x}$  on the basis of  $Z$  [2].

The survival models allow us to solve problems when there are non-linear relationships between covariates (features) of training and testing instances (objects, patients) and the hazard function. However, most machine learning survival models are black-boxes for their users, for example, random survival forests (RSFs) [3–5], the survival SVM [6] or survival neural networks [7, 8], i.e., the users do not know how the model predictions are obtained and which features of the input instance impact on the predictions. As a result, a user, for example, a doctor, cannot understand predictions and cannot believe whether the prediction is correct or not. This fact significantly restricts the machine model usage especially in such areas as medicine, safety, etc.

In order to avoid the problem, several methods for explaining the prediction and the black-box model have been developed and proposed [9–12]. The explanation methods can be divided into methods for local and global explanations. Local methods explain a black-box model locally around a test instance. Most local methods come to the linear approximation of a non-linear function implemented by the black-box model. One of the very popular methods is the Local Interpretable Model-Agnostic Explanation method (LIME) [13]. According to LIME and its modifications [14, 15], random instances are generated around a test instance and coefficients of the linear approximation obtained by using these instances quantify impacts of the corresponding features on the explained prediction. Global methods explain predictions for the whole training set.

However, most methods deal with machine learning models which do not take into account censored data. It should be pointed out that some methods explaining survival models have been proposed in the last time [16–18]. But these methods do not solve an important problem of uncertainty explanation when we would like to find which features of an instance impact on uncertainty of the corresponding prediction provided by the machine learning survival model. By having features contributing to the uncertainty of predictions, a user can change her or his decision concerning the prediction, for example, a doctor can correct a diagnosis in accordance with her or his knowledge about the impacts of features on the stated diagnosis.

To the best of our knowledge, there are only two papers devoted to this aspect [19, 20]. The first one [19] introduces the methods called the Counterfactual Latent Uncertainty Explanations. The second paper [20] derives a sensitivity analysis of input variables for predictive epistemic and aleatoric uncertainty. The above approaches do not consider survival machine learning models as black boxes. Moreover, in contrast to the survival models where the predictions are functions (the survival function or the cumulative hazard function), the above-considered approaches explain only point-valued predictions. Therefore, our aim is to propose a method for explaining the uncertainty of predictions provided by the machine learning survival models. In other words, we try to explain by means of the proposed method why the predicted survival function is uncertain.

An important question arising with respect to the uncertainty consideration is how to measure the prediction uncertainty (the survival function uncertainty). Another question is how to define the most uncertain survival function. The third question following the above ones is how to select features that significantly impact the survival function uncertainty. We propose a method that answers the above questions.

The basic idea underlying the proposed method is to apply an explanation method called the Neural Additive Model (NAM) [21] with a specific expected loss function that defines the uncertainty of survival functions in the framework of survival analysis. NAM is a very promising method that is implemented as a linear combination of neural subnetworks where a single feature is fed to each subnetwork, and the output of each subnetwork is a shape function of the feature. In fact, each subnetwork implements a function of the Generalized Additive Model (GAM) [22]. GAM is a well-known model which approximates a function of the black-box model by a sum of arbitrary shape functions of features. In NAM, shape functions are obtained by training the whole neural network.

Another idea behind the proposed method is to approximate the unknown survival function, produced by the black-box survival mode, by the semi-parametric Cox proportional hazards model [23]. Moreover, the survival function is approximated by the extended Cox model where the simple linear function of covariates is replaced with GAM. This idea jointly with the first one has been used for explaining predictions of machine learning survival models in the form of the SurvNAM method [24].

The third idea behind the proposed method is to derive the expected loss function of NAM, which aims to explain the prediction uncertainty instead of the prediction itself. To derive the corresponding loss function, we have to solve the problem of how to define the prediction uncertainty. This is another problem which is considered below.

In sum, our contribution can be formulated as follows:

1. The most uncertain survival function is defined. This function is derived to compute a measure of certainty as a dual measure to the uncertainty one.
2. A measure of certainty is introduced. We will show that the dual certainty measure can be used instead of the uncertainty one because this measure is simply defined for survival functions.
3. NAM is trained by using the expected loss function which approximates the network output uncertainty by the extended Cox model survival function uncertainty. The extension of the Cox model means that GAM is used in the model instead of the linear combination of covariates.
4. Numerical examples with real censored data illustrate the proposed method.

## 2 Survival Models

An output of a machine learning survival model is the survival function (SF) denoted as  $S(t|\mathbf{x})$  or the cumulative hazard function (CHF) denoted as  $H(t|\mathbf{x})$ . The SF is the probability of surviving of instance  $\mathbf{x}$  up to time  $t$ , i.e.,  $S(t|\mathbf{x}) = \Pr\{T > t|\mathbf{x}\}$ . The

CHF  $H(t|\mathbf{x})$  is the probability of an event at time  $t$  given survival until time  $t$ . The SF can be expressed through the CHF as:

$$S(t|\mathbf{x}) = \exp(-H(t|\mathbf{x})).$$

One of the most popular survival models is the Cox proportional hazards model [1]. The SF  $S(t|\mathbf{x}, \mathbf{b})$  and the CHF  $H(t|\mathbf{x}, \mathbf{b})$  predicted by the Cox model are computed as

$$S(t|\mathbf{x}, \mathbf{b}) = (S_0(t))^{\exp(\mathbf{b}^T \mathbf{x})},$$

$$H(t|\mathbf{x}, \mathbf{b}) = H_0(t) \exp(\mathbf{b}^T \mathbf{x}).$$

where  $H_0(t)$  and  $S_0(t)$  are the baseline CHF and SF, respectively, which do not depend on vector  $\mathbf{x}$  and vector  $\mathbf{b}$  and estimated by using the Nelson–Aalen estimator [1];  $\mathbf{b}^T = (b_1, \dots, b_m)$  is an unknown vector of regression coefficients.

The Cox model can be extended replacing the linear relationship  $\mathbf{b}^T \mathbf{x}$  with the sum of shape functions  $g_k(x_k)$  of features  $x_k$ , i.e., by GAM

$$\psi(\mathbf{x}, \mathbf{g}) = g_1(x_1) + \dots + g_m(x_m).$$

Here  $\mathbf{g} = (g_1, \dots, g_m)$  is the vector of the GAM shape functions. We will use this representation for constructing the explanation model based on NAM.

It is important to note that if  $t_0 < t_1 < \dots < t_s$  be the distinct times to event of interest, then the predicted SF of a survival model can be often represented as a stepwise function of the form:

$$S(t|\mathbf{x}) = \sum_{j=0}^s S_j(\mathbf{x}) \cdot \chi_j(t),$$

where  $S_j(\mathbf{x})$  is a value of the SF in interval  $[t_j, t_{j+1}]$ ;  $\chi_j(t)$  is the indicator function taking value 1 if  $t \in [t_j, t_{j+1}]$ .

The CHF can be written in the same way:

$$H(t|\mathbf{x}) = \sum_{j=0}^s H_j(\mathbf{x}) \cdot \chi_j(t),$$

where  $H_j(\mathbf{x})$  is a value of CHF  $H(t|\mathbf{x})$  in the interval  $[t_j, t_{j+1}]$ .

### 3 Uncertainty and Certainty Measures of Survival Models

The next question is how to express uncertainty  $u(\mathbf{x})$  of the survival model prediction corresponding to an example  $\mathbf{x}$ . A simple way is to find a distance between the prediction and its true values, for example, between the predicted SF  $S(t|\mathbf{x})$  and some true SF. However, the true SF is unknown, and we can define the most uncertain SF  $S^u(t|\mathbf{x})$  which can be obtained by the given time moments  $t_0 < t_1 < \dots < t_s$  of events. If SFs are represented as stepwise functions, then we can define the most uncertain SF as a function, for which probabilities  $p_j(\mathbf{x}) = S_{j-1}(\mathbf{x}) - S_j(\mathbf{x})$  for every  $j = 1, \dots, s$  are identical. In this case, we have a uniform distribution over the points  $t_0, \dots, t_s$  and get the maximum of the Shannon entropy  $H(\mathbf{p}) = \sum_{j=1}^s p_j \ln p_j$ . Hence, the most uncertain SF is written as

$$S^u(t) = \sum_{j=0}^s \left(1 - \frac{j}{s}\right) \chi_j(t) = \sum_{j=0}^s s_j^u \cdot \chi_j(t).$$

It is interesting to note that  $S^u(t)$  does not depend on  $\mathbf{x}$ . Hence, we can define the certainty measure  $c(\mathbf{x})$  as a distance  $D$  between the predicted SF  $S(t|\mathbf{x})$  and  $S^u(t)$ . Let us suppose that the distance is defined as  $L_2$ -norm. It turns out that the distance between SFs leads to complex expressions. Therefore, we propose to consider the distance between logarithms of the corresponding CHFs  $H(t|\mathbf{x})$  and  $H^u(t)$ , i.e., we represent the certainty measure as

$$c(\mathbf{x}) = \int_0^\infty (\ln H(t|\mathbf{x}) - \ln H^u(t))^2 dt.$$

Substituting the corresponding expressions for  $H(t|\mathbf{x})$  and  $H^u(t)$  into the above expression for the certainty measure and using the fact that the CHFs are stepwise functions, we get after simplification:

$$c(\mathbf{x}) = \sum_{j=0}^s (\ln H_j(\mathbf{x}_k) - \ln(-\ln s_j^u))^2 \tau_j,$$

where  $\tau_j = t_{j+1} - t_j$ .

If we take the Cox model as the black-box one, then the certainty measure can be written in the same way taking into account that the CHF, in this case, is expressed through the baseline CHF  $H_0(t)$ . Hence, we can write the certainty measure as

$$c_{\text{Cox}}(\mathbf{x}, \mathbf{g}) = \sum_{j=0}^s (\psi(\mathbf{x}, \mathbf{g}) + \ln H_{0j} - \ln(-\ln s_j^u))^2 \tau_j.$$



## 4 NAM and Its Use for Uncertainty Explanation

NAM proposed by Agarwal et al. [21] can be represented as a specific neural network that implements the function  $\psi(\mathbf{x}, \mathbf{g}) = g_1(x_1) + \dots + g_m(x_m)$  of GAM such that every function  $g_i(x_i)$  of the  $i$ -th feature is implemented by a separate neural subnetwork  $\psi(\mathbf{x}, \mathbf{g})$ . The whole network consists of  $m$  subnetwork which may be different because they implement different functions of features. The input of the  $i$ -th subnetwork is only the  $i$ -th feature. Outputs of all subnetworks are summed to approximate  $\psi(\mathbf{x}, \mathbf{g})$ . The loss function for the training of the original NAM [21] is based on the  $L_2$ -norm distance between the whole network output  $\psi(\mathbf{x}, \mathbf{g})$  and the class label of the current instance  $\mathbf{x}$  from the training set.

We train NAM with a quite different loss function which takes into account our aim to explain the survival prediction uncertainty. We propose a way how to incorporate the uncertainty into the loss function. The basic idea behind the incorporation is to use the extended Cox model with function  $\psi(\mathbf{x}, \mathbf{g})$ . We are looking for the Cox model which has the same uncertainty (or the same certainty) as the black-box survival model. Since the Cox model is used as an approximation of the black-box model, then actually we are looking for the Cox model whose certainty approximates the certainty of the black-box survival model. This can be formally written as

$$(c(\mathbf{x}) - c_{\text{Cox}}(\mathbf{x}, \mathbf{g}))^2.$$

In order to explain the uncertainty for the instance  $\mathbf{x}$ , we train NAM by generating  $N$  synthetic instances in the local area around the explained instance (see the same approach in LIME [13]). Then the expected loss function for training NAM is

$$L(\mathbf{g}) = \sum_{k=1}^N w_k (c(\mathbf{x}_k) - c_{\text{Cox}}(\mathbf{x}_k, \mathbf{g}))^2.$$

Here  $w_k$  is the weight which is assigned to the synthetic instance  $\mathbf{x}_k$  as a decreasing function  $\psi$  of the distance  $D$  between  $\mathbf{x}$  and  $\mathbf{x}_k$ , i.e.,  $w_k = \psi(D(\mathbf{x}_k, \mathbf{x}))$ . Substituting expressions for  $c(\mathbf{x}_k)$  and  $c_{\text{Cox}}(\mathbf{x}_k, \mathbf{g})$  into the above expression, we obtain

$$\begin{aligned} L(\mathbf{g}) &= \sum_{k=1}^N w_k \left( \sum_{j=0}^s \tau_j \left( \ln H_j(\mathbf{x}_k) - \ln(-\ln s_j^u) \right) \right)^2 \\ &\quad - \sum_{j=0}^s \tau_j \left( \sum_{i=0}^m g_i(x_i) + \ln H_{0j} - \ln(-\ln s_j^u) \right)^2. \end{aligned}$$

Finally, we write the following algorithm for the uncertainty explanation estimate by means of the shape functions, i.e., the algorithm for computing  $\mathbf{g}$  on the basis of the dataset  $Z$  with  $\mathbf{x}_i \in \mathbb{R}^m$ . Let  $\mathbf{x}$  be the explainable instance. Suppose that there is a trained black-box model  $M$ .

1. Determine a neural network architecture, i.e., the number of layers in every subnetwork, the number of units in every layer, activation functions.
2. Define the function  $\psi$  and its parameters for computing weights  $w_k$ .
3. Initialize  $N$ , parameters of distance  $D(\mathbf{x}_k, \mathbf{x})$  between  $\mathbf{x}$  and generated instances  $\mathbf{x}_k$ .
4. Estimate the baseline CHF  $H_0(t)$  by using the Nelson–Aalen estimator.
5. Generate  $N$  instances  $\mathbf{x}_k$  around  $\mathbf{x}$ .
6. Compute weights  $w_k = \psi(D(\mathbf{x}_k, \mathbf{x}))$  of instances  $\mathbf{x}_k, k = 1, \dots, N$ .
7. Compute  $s_j^u, \tau_j$  for all  $j = 0, \dots, s$ .
8. Predict  $H(\mathbf{x}_k)$  for every  $\mathbf{x}_k$  by means of the black-box model,  $k = 1, \dots, N$ .
9. Train NAM on the set of  $N$  generated instances  $\mathbf{x}_k$  using the loss function  $L(\mathbf{g})$ .
10. Visualize obtained shape functions  $g_1, \dots, g_m$ , which explains the prediction  $H(t|\mathbf{x})$ .

## 5 Numerical Experiments with Real Data

In order to illustrate the proposed explanation method, it is studied on real benchmark datasets whose description is given below.

The German Breast Cancer Study Group 2 (GBSG2) Dataset consists of 686 observations such that every instance is characterized by 8 features and 2 target values, including the age of the patients in years, menopausal status, tumor size, tumor grade, number of positive nodes, hormonal therapy, progesterone receptor, estrogen receptor, recurrence-free survival time, censoring indicator (0—censored, 1—event). The dataset is available in the “TH.data” R package.

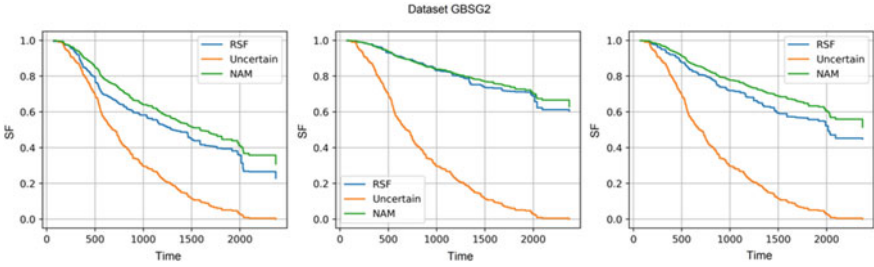
The Monoclonal Gammopathy (MGUS2) Dataset is the natural history of 1384 subjects with monoclonal gammopathy of undetermined significance characterized by 5 features. The dataset is available in the “survival” R package.

The perturbation technique for implementing the local explanation is realized by generating  $N$  nearest points  $\mathbf{x}_k$  around the explained instance  $\mathbf{x}$  in accordance with the normal distribution having the expectation at point  $\mathbf{x}$  and the standard deviation  $r$  which is equal to 5% of the largest distance between points of the corresponding studied dataset. The number of generated points  $N$  is equal to 1000. Weights  $w_k$  are computed by means of the following expression:

$$w_k = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}_k\|_2^2}{r}\right), k = 1, \dots, N.$$

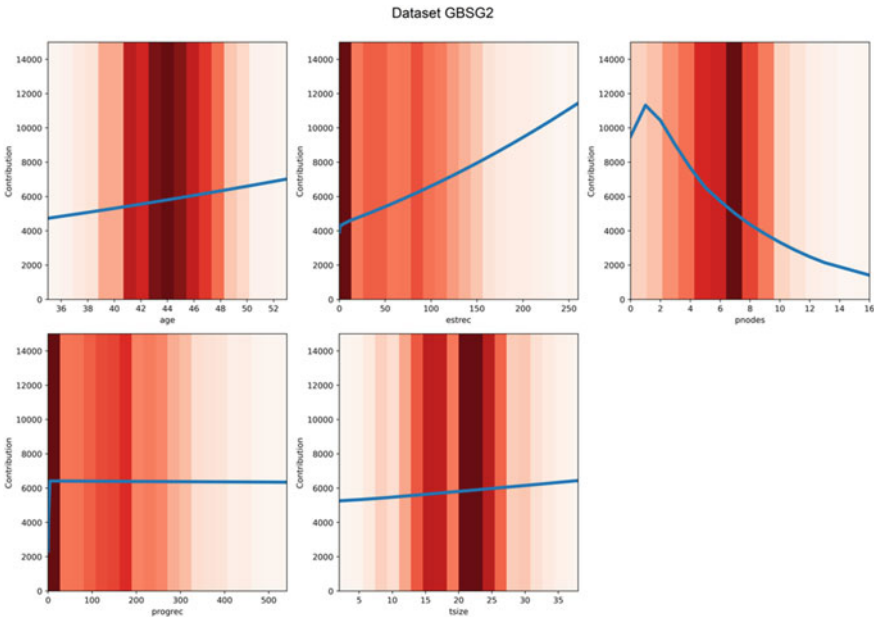
The random survival forest [3] consisting of 1000 decision survival trees is used as a black-box model.

The Python code of NAM for its modification to implement the proposed method is taken from the following site: [https://github.com/google-research/google-research/tree/master/neural\\_additive\\_models](https://github.com/google-research/google-research/tree/master/neural_additive_models). The first dataset for training NAM is the GBSG2 dataset. Figure 1 illustrates three cases of SFs for 3 randomly selected

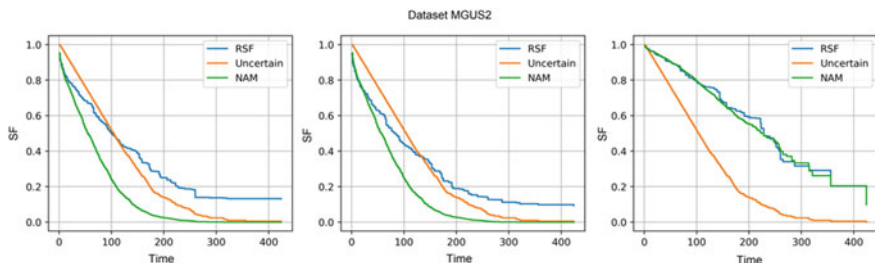


**Fig. 1** SFs obtained by means of the black-box RSF model, the extended Cox model or NAM, and the most uncertain SFs for the dataset GBSG2 and three randomly selected points

instances from the training set. Every picture in Fig. 1 shows three SFs obtained by means of the RSF, by means of the extended Cox model with computed shape functions  $g_i(x_i)$  (or by the modified NAM) and by computing the most uncertain function  $S''(t)$  for given training data. It can be seen from Fig. 1 that SFs predicted by RSF and the proposed method are located approximately at the same distance from the most uncertain SF. This implies that the neural network is acceptably trained and can be used for the uncertainty or certainty explanation. Figure 2 illustrates shape functions for all numerical features. They are computed for an instance with features “age” = 44, “estrec” = 67, “horTh” = yes, “menostat” = Pre, “pnodes” = 6, “progrec”



**Fig. 2** Shape functions explaining the certainty impact of every feature obtained by means of the proposed method for the dataset GBSG2 and the local explanation



**Fig. 3** SFs obtained by means of the black-box RSF model, the extended Cox model or NAM, and the most uncertain SFs for the dataset MGUS2 and three randomly selected points

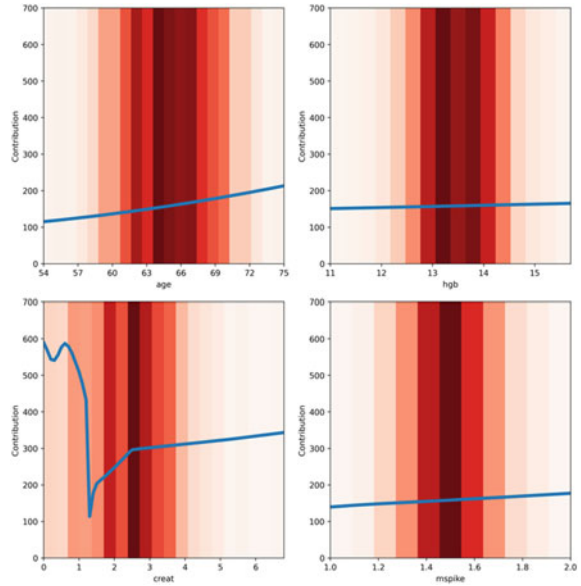
= 150, “tsize” = 20, “tgrade” = 1. Shape functions with large changes (increasing or decreasing) indicate the important features. It can be seen from 2 that features “estrec” (estrogen receptor) and “pnodes” (the number of positive nodes) maximally impact the prediction certainty.

Another dataset illustrating the proposed method is the MGUS2 dataset. Figure 3 illustrates three cases of SFs for 3 randomly selected instances from the training set similarly to the case considered for the dataset GBSG2. Every picture in Fig. 3 shows three SFs obtained by means of the RSF, by means of the extended Cox model with computed shape functions  $g_i(x_i)$  (or by the modified NAM) and by computing the most uncertain function  $S''(t)$  for given training data. It can be seen from Fig. 3 that SFs predicted by RSF and the proposed method are located at different sides of the most uncertain SF (see the first and the second pictures). However, distances from the curves are not significantly different. This implies that the neural network is also acceptably trained. Figure 4 illustrates shape functions for all numerical features. They are computed for an instance with features “age” = 65, “sex” = F, “hgb” = 13.5, “creat” = 2.4, “mspike” = 1.5. It can be seen from 4 that feature “creat” significantly contributes into the prediction certainty. Feature “age” is also important but to a lesser extent.

## 6 Conclusion

A method for uncertainty explanation of machine learning survival models has been proposed and studied by means of numerical experiments. Only the  $L_2$ -norm was considered for representing distances between SFs or CHFs. Moreover, only the  $L_2$ -norm was used to compute the difference between certainty measures of predictions provided by black-box models and the extended Cox model. However, it is also interesting to consider other metrics, including the Manhattan distance (the  $L_1$ -norm) and the Chebyshev distance (the  $L_\infty$ -norm). Their use can be regarded as a direction for further research.

**Fig. 4** Shape functions explaining the certainty impact of every feature obtained by means of the proposed method for the dataset MGUS2 and the local explanation



Only aleatoric uncertainty has been studied in the chapter when the natural randomness takes place, for example, class overlap, data noise, etc. It is interesting to consider also how epistemic uncertainty, which represents our ignorance about the model caused by the lack of observation data, can be explained. This is another direction for further research.

In sum, we have to point out that NAM is a very promising explanation method that can be applied to various objects by changing the loss function in accordance with the properties of the considered object. We have studied only one of the possible applications, but other applications, for example [25, 26], may be also directions for further research

**Acknowledgements** This work is supported by the Russian Science Foundation under grant 21-11-00116.

## References

1. Hosmer, D., Lemeshow, S., May, S.: Applied Survival Analysis: Regression Modeling of Time to Event Data. Wiley, New Jersey (2008)
2. Wang, P., Li, Y., Reddy, C.: Machine learning for survival analysis: a survey. *ACM Comput. Surv. (CSUR)* **51**, 1–36 (2019)
3. Ibrahim, N., Kudus, A., Daud, I., Bakar, M.A.: Decision tree for competing risks survival probability in breast cancer study. *Int. J. Biol. Med. Res.* **3**, 25–29 (2008)

4. Nasejje, J., Mwambi, H., Dheda, K., Lesosky, M.: A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data. *BMC Med. Res. Methodol.* **17**, 1–17 (2017)
5. Wright, M., Dankowski, T., Ziegler, A.: Unbiased split variable selection for random survival forests using maximally selected rank statistics. *Stat. Med.* **36**, 1272–1284 (2017)
6. Belle, V.V., Pelckmans, K., Huffel, S.V., Suykens, J.: Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artif. Intell. Med.* **53**, 107–118 (2011)
7. Faraggi, D., Simon, R.: A neural network model for survival data. *Stat. Med.* **14**, 73–82 (1995)
8. Katzman, J., Shaham, U., Cloninger, A., Bates, J., Jiang, T., Kluger, Y.: Deepsurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 1–12 (2018)
9. Arya, V., Bellamy, R., Chen, P.Y., Dhurandhar, A., Hind, M., Hoffman, S., Houde, S., Liao, Q., Luss, R., Mojsilovic, A., Mourad, S., Pedemonte, P., Raghavendra, R., Richards, J., Sattigeri, P., Shanmugam, K., Singh, M., Varshney, K., Wei, D., Zhang, Y.: One explanation does not fit all: a toolkit and taxonomy of AI explainability techniques. [arXiv:1909.03012](https://arxiv.org/abs/1909.03012) (2019)
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. *ACM Comput. Surv.* **51**, 93 (2019)
11. Molnar, C.: *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.* (2019). Published online, <https://christophm.github.io/interpretable-ml-book/>
12. Murdoch, W., Singh, C., Kumbier, K., Abbasi-Asl, R., Yua, B.: Interpretable machine learning: definitions, methods, and applications. [arXiv:1901.04592](https://arxiv.org/abs/1901.04592) (2019)
13. Ribeiro, M., Singh, S., Guestrin, C.: “Why should I trust You?” Explaining the predictions of any classifier. [arXiv:1602.04938v3](https://arxiv.org/abs/1602.04938v3) (2016)
14. Garreau, D., von Luxburg, U.: Explaining the explainer: a first theoretical analysis of LIME. [arXiv:2001.03447](https://arxiv.org/abs/2001.03447) (2020)
15. Garreau, D., von Luxburg, U.: Looking deeper into tabular LIME. [arXiv:2008.11092](https://arxiv.org/abs/2008.11092) (2020)
16. Kovalev, M., Utkin, L., Kasimov, E.: SurvLIME: A method for explaining machine learning survival models. *Knowl. Based Syst.* **203**, 106164 (2020)
17. Kovalev, M., Utkin, L.: A robust algorithm for explaining unreliable machine learning survival models using the Kolmogorov-Smirnov bounds. *Neural Netw.* **132**, 1–18 (2020)
18. Kovalev, M., Utkin, L.: Counterfactual explanation of machine learning survival models. [arXiv:2006.16793](https://arxiv.org/abs/2006.16793) (2020)
19. Antoran, J., Bhatt, U., Adel, T., Weller, A., Hernandez-Lobato, J.: Getting a CLUE: A method for explaining uncertainty estimates. [arXiv:2006.06848](https://arxiv.org/abs/2006.06848) (2020)
20. Depeweg, S., Hernandez-Lobato, J., Udluft, S., Runkler, T.: Sensitivity analysis for predictive uncertainty in Bayesian neural networks. [arXiv:1712.03605](https://arxiv.org/abs/1712.03605) (2017)
21. Agarwal, R., Frosst, N., Zhang, X., Caruana, R., Hinton, G.: Neural additive models: Interpretable machine learning with neural nets. [arXiv:2004.13912](https://arxiv.org/abs/2004.13912) (2020)
22. Hastie, T., Tibshirani, R.: *Generalized additive models.* Vol. 43. CRC Press (1990)
23. Cox, D.: Regression models and life-tables. *J. Royal Statist. Soc. Series B (Methodological)* **34**, 187–220 (1972)
24. Utkin, L., Satyukov, E., Konstantinov, A.: SurvNAM: The machine learning survival model explanation. [arXiv:2104.08903](https://arxiv.org/abs/2104.08903) (2021)
25. Konstantinov, A.V., Utkin, L.V.: A generalized stacking for implementing ensembles of gradient boosting machines. In: *Cyber-Physical Systems: Digital Technologies and Applications.* Springer International Publishing Series Title Studies in Systems, Decision and Control. Series Vol. 350. pp. 3–16. (2020). [https://doi.org/10.1007/978-3-030-67892-0\\_1](https://doi.org/10.1007/978-3-030-67892-0_1)
26. Bolshakov, A.A., Veshneva, I.V., Lushin, D.: Mathematical model of integration of cyber-physical systems for solving problems of increasing the competitiveness of the regions In: *Studies in Systems, Decision and Control. Society 5.0: Cyberspace for Advanced Human-Centered Society.* Springer Nature Switzerland. Vol. 333. pp. 129–139 (2021). *Decision and Control* 333, [https://doi.org/10.1007/978-3-030-63563-3\\_11](https://doi.org/10.1007/978-3-030-63563-3_11). ISSN 2198-4182

# Biocomputation Using Molecular Agents Moving in Microfluidic Channel Networks: An Alternative Platform for Information Technology



Thomas Blaudeck , Christoph R. Meinecke, Danny Reuter ,  
Sönke Steenhusen, Archa Jain, Sascha Hermann, Stefan E. Schulz,  
Eduard I. Zenkevich , Till Korten , and Heiner Linke 

**Abstract** Deficiencies in software or computer chips cause computers or smart-phones to crash and allow hackers to steal passwords. Automated test procedures could avoid these problems. However, the computing power and cooling requirements of conventional computers increase exponentially with the size of the problem, so that the technological limits for solving these problems will soon be reached. The EU project Bio4Comp aims to develop concepts for a bio-computer to help overcome these two main problems. Compared to conventional computers, computers based on biological molecular motors only consume a fraction of the energy per arithmetic

---

T. Blaudeck (✉) · S. E. Schulz · E. I. Zenkevich  
Research Center for Materials, Architectures and Integration of Nanomembranes (MAIN),  
Chemnitz University of Technology, Rosenbergstr. 6, 09126 Chemnitz, Germany  
e-mail: [thomas.blaudeck@main.tu-chemnitz.de](mailto:thomas.blaudeck@main.tu-chemnitz.de)

T. Blaudeck · C. R. Meinecke · D. Reuter · A. Jain · S. Hermann · S. E. Schulz  
Center for Microtechnologies (ZfM), Chemnitz University of Technology, Reichenhainer Str. 70,  
09126 Chemnitz, Germany

T. Blaudeck · D. Reuter · A. Jain · S. Hermann · S. E. Schulz · E. I. Zenkevich  
Fraunhofer Institute for Electronic Nano Systems (ENAS), Technologie-Campus 3, 09126  
Chemnitz, Germany

S. Steenhusen  
Fraunhofer Institute for Silicate Research (ISC), Neunerplatz 2, 97082 Würzburg, Germany

S. Hermann · S. E. Schulz  
Center for Advancing Electronics Dresden (Cfaed), Chemnitz University of Technology,  
Reichenhainer Str. 70, 09126 Chemnitz, Germany

E. I. Zenkevich  
Department of Information Technologies and Robotics, National Technical University of Belarus,  
Nezavisimosti Ave. 65, 220013 Minsk, Belarus

T. Korten  
Center for Molecular Bioengineering (B CUBE), Dresden University of Technology, Tatzberg 41,  
01307 Dresden, Germany

H. Linke  
Solid State Physics and NanoLund, Lund University, Professorsgatan 1, 22100 Lund, Sweden

operation and scale very well for problems that can be parallelized (“multitasking”). In this article, the topic network-based biocomputation (NBC) i.e. computing with biological molecules as agents that are driven by molecular motors in microfluidic networks, is presented as an alternative approach to computing, data processing, and information technology.

**Keywords** Information technologies · Biological molecular motors · Microfluidic networks · Nanooptical methods of fabrication and analytics

## 1 Introduction

### 1.1 *Alternative Computing Approaches*

The global trend of digitalization is pushing classical approaches of computation and information technologies to the limits of their efficiency. This has led to alternative, often hybrid computing approaches in research and technology that fulfil the new requirements regarding computational power, trustability and energy efficiency. Particular research fields worldwide are Quantum Computation (QC), but also other approaches such as DNA Computation and Network-Based Biocomputation (NBC) are to be mentioned.

**Quantum computation (QC).** QC follows the idea of the superposition of quantum states: OFF “0” or ON “1” are not any more, as in classical theory, distinct, but may exist simultaneously.  $N$  qubits would decode  $2^N$  classical bits, allowing a massive parallel computing [1, 2]. Tractable problems are integer factorization e.g. for applications in cryptography,  $10^{10}$  steps in quantum cryptography would replace  $10^{24}$  steps in classical cryptography. Further applications regard optimization tasks for database searches in economy or logistics, simulations e.g. towards novel chemicals for application in bioelectronics and medicine or new materials for advanced electronic devices, e.g. energy storage and transmission [3]. Quantum logic gates have currently realized up to ca 70 qubits. Using a quantum approximate optimization algorithm, Bengtsson et al. solved small instances of the NP (nondeterministic polynomial time)-complete exact-cover problem, with 96.6% success probability [4]. However, the infrastructure for QC requires huge investments and efficient measures to overcome the decoherence of the quantum states of the qubits that leads to errors.

**DNA computation.** DNA computation [5, 6] belongs to the bio-inspired computation approaches. It uses programmable modules to solve mathematical problems. The idea is that, in theory, six grams [=0,006 kg] of DNA/RNA ([desoxy-] ribonucleic acid) do theoretically have a storage capacity of  $3072 \text{ EiB} = 3 \text{ ZiB}$  [1 Zettabyte  $\sim 3,072 \cdot 10^{21}$  byte]. Logic gates can be made from molecules of DNA and RNA and their base pairs and be trained to perform massively parallel calculations of about  $10^{18}$  operations/second (for comparison, commercial computer achieve  $10^{16}$  operations/second). For instance, with an early DNA computer, the solution of a



20-variable instance of a 3-SAT problem was solved after exhaustive search over  $2^{20}$  (1,048,576) possible solutions [7]. A fabrication technology can be based on so-called DNA origami. According to this concept, self-assembled molecular DNA scaffolds are placed on surfaces and in microscopic reactive environments to form nano-scaled circuits and devices. Problems tractable by DNA computing are all those which require a massive parallelization, e.g. simple variants of the travelling-salesman problem (TSP) based on a Hamiltonian cycle, however, as DNA is subject to base pair errors during reproduction, accuracy and liability have their limits.

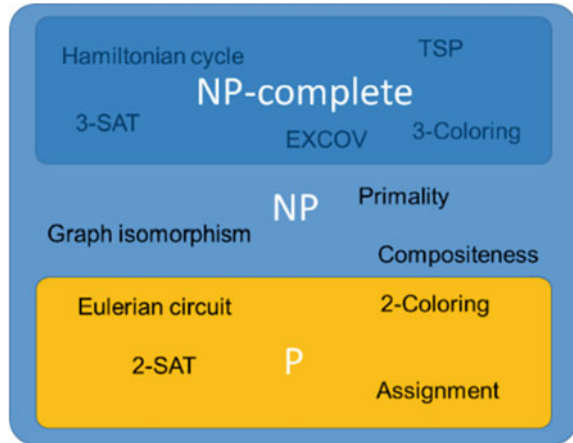
**Network-based biocomputation (NBC).** NBC uses biomolecular agents driven by biomolecular motors inside networks of micro- and nanofluidic channels, acting as programmable modules for mathematical problem solving [8, 9]. The storage, manipulation (i.e. calculation) and readout of data are inspired by and realized with respect to biochemical processes in nature. The instance of the mathematical problem is coded into a micro fluidic channel network that can be fabricated by conventional semiconductor nanotechnology, comprising initial landing zones and confined channels, junctions (split junctions, pass junctions), detectors and further architectural elements for the computational processing [10]. The biological molecules or bacteria act as filament agents and could in principle be labelled or tagged with molecular flags e.g. from DNA, nanoparticles, quantum dots, fluorescent molecules, and other building blocks. Split junctions are designed in a way so that the agents face an exact 50% probability to turn left or right, pass junctions for an error-free passage of the agents without making a turn (cf. Sect. 2.2). Hence, physical guiding and propelling of regenerative agents induced by biochemical energy supply moves the agents through network [11].

## 1.2 Computational Complexity and Energy Considerations

The theory of computational complexity helps to answer which class of problems are tractable by NBC as an alternative computation. It relates to the assumption that the tractability of an optimization problem is not much larger than the tractability of the associated decision problem [12]. The associated algorithm can be translated into a binary search over the possible objective function values to solve the optimization problem. Figure 1 contains relevant mathematical classes.

The class P (“polynomial time”) is defined as the set of decision problems, for which an efficient (i. e. running in polynomial time) solution algorithm exists. The class NP (“non-deterministic polynomial time”) contains all decision problems with this property: for a specific instance of the problem (“yes”, “no”) there exists an efficient solution algorithm, i. e. a certain asymmetry between the instances is possible. If an efficient solution algorithm for each single instance of the associated decision problem exists, then there exist efficient solution algorithms for all problems in the class (NP-complete problems). NP-complete problems are the main application area for NBC. Tractable mathematical problems by NBC are, aside of the

**Fig. 1** The class of NP-complete problems with respect to class NP (“non-deterministic polynomial time”) and class P (“polynomial time”) and examples for the different classes. Scheme adapted from S. Mertens, Computational Complexity for Physicists, Ref. [15]



travelling-salesman problem TSP, satisfiability problems (SAT), exact-cover problems (EXCOV), and subset-sum problem (SSP), as representatives of the class of NP-complete problems.

A tremendous potential benefit of NBC with respect to classical computing is the energy efficiency during operation. Both DNA computing and NBC have been estimated to be able to work very close to the Landauer limit (energy per operation). Given, for example, an integer SSP, estimates show that NBC consumes  $10^{-14}$  J/operation compared to  $10^{-10}$  J/operation in classical computing [10, 13]. As outlined recently by Konopik et al., this finite-time energy cost is independent of problem size [14]. From the perspective of finite-time energy cost, biological computers thus offer a potentially large, fundamental advantage over electronic computers.

## 2 Fabrication of Network-Based Biocomputation

### 2.1 Manufacturing Paradigma

As discussed above, the NBC concept is based on biological agents that travel through micro and nanofluidic network systems designed in a way to decode the particular instance of a mathematical problem. To date, agents are represented by cytoskeletal filaments [9], fungi [10], or biomolecular microtubuli [16, 17]. The hardware fabrication is related to scale-bridging manufacturing schemes on the basis of established manufacturing methodology in microelectronic device fabrication, including aspects of nanotechnology and schemes for heterogeneous systems integration.

**NBC networks based on microfluidic systems.** To allow scaling of the problems to be solved, one clear aim of manufacturing science and engineering is to down-scale the NBC feature size towards nanostructured fluidic channel systems and—at the same time—upscale of NBC network in area (footprint) and volume allowing large-area networks e.g. on silicon or quartz wafers. We use the term *space-encoded NBC networks* to highlight that the agent position in the channels contain information, opposed to *agent-encoded NBC networks* in which the agents carry tags. The manufacturing steps comprise particular biofunctionalization with molecular motors that are statically bound to the channel to drive the mobile agents. It further includes integration concepts for architectural elements such as error-free and, in the future, switchable split junctions and pass junctions, biochemical energy supply, encapsulation, and nonetheless the integration of data handling.

**Agent supply, identification and multiplication.** For sufficiently small problems, the NBC networks are fed with motile agents through the entrances. Motility can be achieved via bio-inspired methods, e.g. by coating the channels with kinesin molecules that induce a translational motility to the microtubules [16, 17]. An alternative pair is myosin as the motility enabler and actin filaments as the agent [18, 19]. The agents pass the NBC network through junctions and leave it through defined exits. To allow a scaling of the tractability of the mathematical problem, agent multiplication can be considered.

**Agents, labels, and tags.** Monitoring the agents' exit from the NBC chip or even observing their entire trajectory through the network is a critical paradigm. At the first glance, in purely space-encoded NBC networks, agents carry information on their presence by their dielectric properties that can be employed e.g. for electrical means of detection. At the second glance, agents can be dynamically supplied with flags during the computational process that carry information about the read-out. Aside of space encoding of the problem or its instances, agent encoding of the problem or its instances can be pursued. The manufacturing paradigms thus have to include integration aspects for physical, chemical and biological marking (barcoding, tagging, labelling) of the agents.

## 2.2 Architectural Elements and Fabrication Tools

The manufacturing paradigms invoke design rules for NBC networks using certain architectural elements that fit into a seamless manufacturing chain.

**Microfluidic channel network.** The overall structuring of the NBC network system has to fulfil the aspects both of downscaling in feature size and the upscaling in principal area. The patterning of the microfluidic channel system has to provide compatibility of large-area (e.g. wafer-scale) fabrication with highest precision, reproducibility and reliability. A premier tool for the fabrication of the NBC computers is electron beam lithography (EBL) addressing the necessary channel structures. EBL

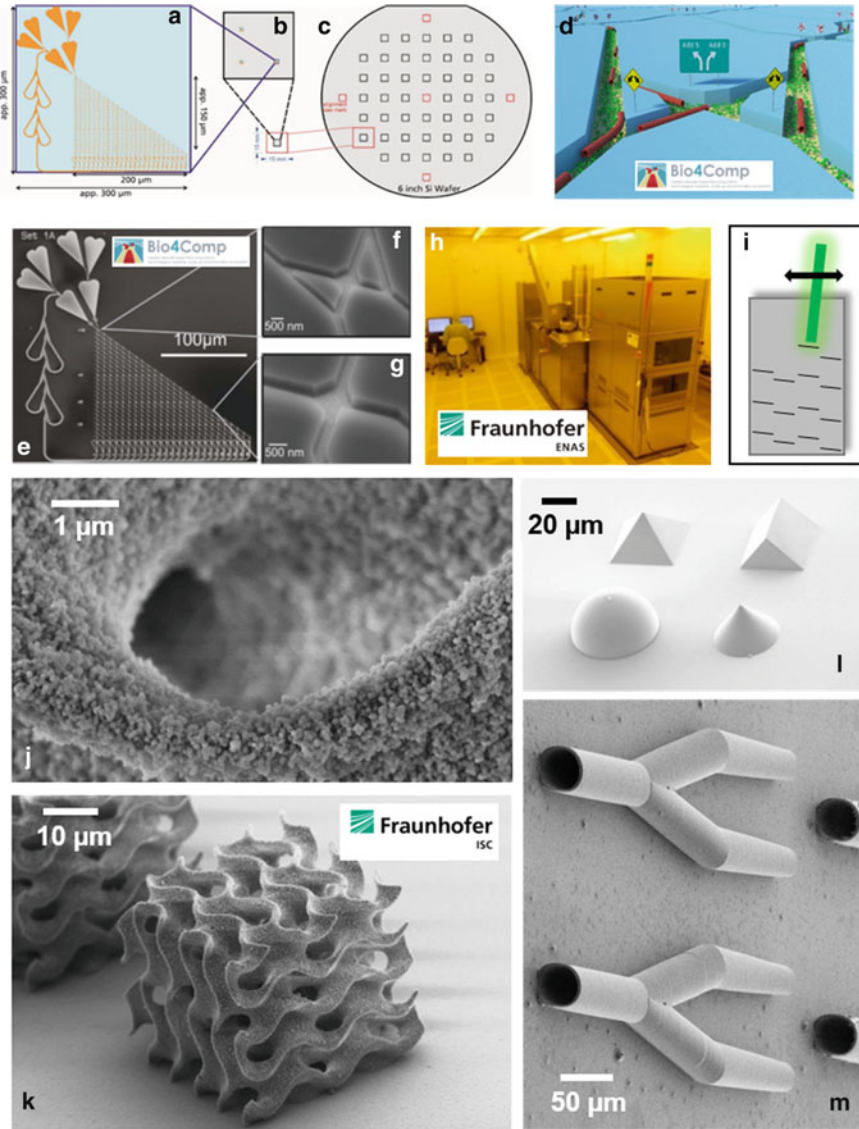
allows for high throughput, high accuracy patterning below the 32 nm node regime. In the current research work, a Vistec SB 254 EBL setup in variable-shape beam technology, equipped with a 50 keV electron accelerator and a scanning-electron inspection system allows high-resolution and fast patterning along vector scans in a write-on-the-fly mode.

**Junctions between microfluidic channels and periphery.** For the space-encoding NBC networks, junctions between the channels in the microfluidic system decode the instance of the given mathematical problem. In contemporary NBC computing, they comprise pass junctions and split junctions operating at highest precision, that is, lowest possible error rates. For that end, two-photon polymerization (TPP) can be seen as a viable direct-write approach for the manufacturing of split junctions and pass junctions as it allows writing of three-dimensional over-/underpasses e.g. in a lab-on-chip matrix material. Steenhusen et al. employed TPP on proprietary hybrid materials (Ormocers®) as negative-tone resists and induce photopolymerization with femtosecond laser pulses through non-linear optical effects as a means for closed-loop underpasses over-/underpasses [20, 21]. Perspectively, junctions that are switchable between passing and turning, literally ‘kicking’ the motile agents in the right channel dynamically, can help to increase the addressable tractability of the mathematical problem. The periphery of the NBC networks include landing zones and other parts to feed agents into the fluidic system.

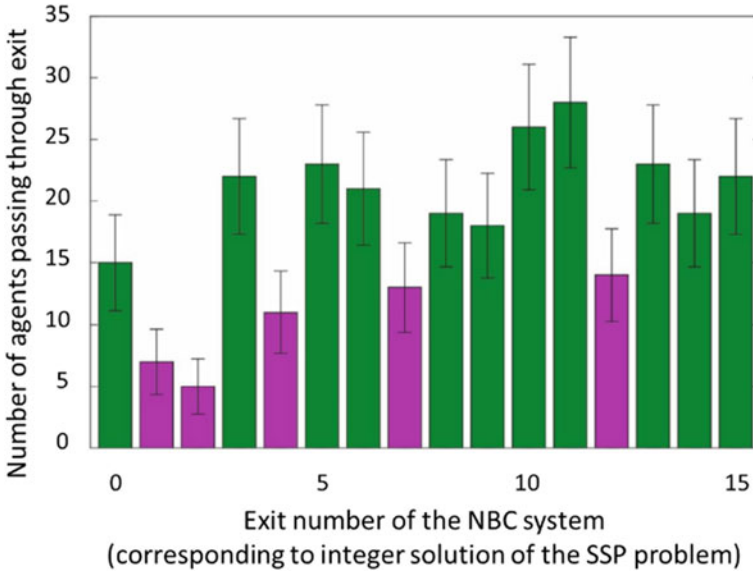
**Detectors for identification and data integration.** Candidates for identification of the motile agents include (but are not limited to) fluorescence labels (e.g. fluorescent dye molecules bound by oligonucleotide linkers [22], perspectively quantum dots [23, 24], at best in supramolecular complexes that are stable in aqueous environment over a considerable pH range [25]), oligonucleotides (e.g. DNA), magnetic information or spin (nanoparticles, complex molecules), etc. The handling and integration of data and other information from the NBC chip to the adjacent peripheral electronic interface includes the readout of signals, local storage and transfer of results to classical electronic systems in the periphery.

### ***2.3 Integrated Manufacturing Scheme for a Functional NBC Network Including Microfluidic Channels, Junctions and Detection of Agents***

The subset-sum problem (SSP) is a well-known problem in computer science especially in cryptography and complexity theory [26]. SSP asks whether a target sum  $T$  can be reached by summing up any combination of integers from set  $S$ . For example, given the set  $S = \{3, 5, 6, 10\}$  and the target  $T = 11$  we would sum up  $5 + 6$ , while for  $T = 12$  there is no solution. Design, manufacturing scheme and evaluation of the results of a successfully functional NBC network for an instance of a mathematical problem are presented in Figs. 2 and 3, respectively.



**Fig. 2** Panel figure: **a** Design of a typical NBC fluidic network with landing zones and reflux system. **b** Several NBC networks are associated on one silicon chip. **c** Several chips are fabricated on a wafer. **d** Molecular agents travel through the fluidic channel system driven by molecular motors, as shown in the graphic-artist's view. **e** SEM image of the designed NBC network including split junctions **f** and pass junctions **g** governing the movement of the molecular agents. **h** Electron beam lithography with a 50 keV electron accelerator **i** and vector scan function is used for fabrication of the NBC fluidic network. Alternatively, two-photon polymerization (TPP) is employed to design voids **j** and **k**, microoptical elements **i** and channel structures **m** on the micrometer scale. Copyrights: **a** to **d**: by the Authors; **e** to **g**: by Cornelia Kowol (TU Chemnitz); **h**, **i**: by the Authors; **j**, **k**: by the Authors, featured in Ref. [33]; **l**: by the Authors, adapted from Ref. [20]; **m**: by the Authors, adapted from Ref. [34]



**Fig. 3** Aggregated results of agent exit passages after their exploration of a space-encoded NBC network decoding a four-instance  $\{3, 5, 6, 10\}$  SSP problem. Exit passages were optically detected and aggregated by image correlation; due to limited field-of-view, only 16 of the 25 exits are shown. On the abscissa, exits corresponding to correct solutions of the SSP problem are labelled in green; exits corresponding to incorrect solutions of the SSP problem are indicated in magenta. Copyright by the Authors, adapted from C. R. Meinecke et al. Ref. [ 27]

Figure 2 illustrates the design and manufacturing aspects. The panel parts (A), (B) and (C) contain the implementation of this problem into a fluidic network design on the levels of the NBC device, chip, and wafer, respectively. (D) shows a graphic-artist’s view on the principle of the travel of the agents through channels and pass junctions. The parts (E), (F) and (G) present the scanning electron microscopy images of an entire NBC network and individual examples for a pass junction and a split junction, respectively. One recently investigated manufacturing scheme is comprised by the parts (H) and (I) showing electron beam lithography (EBL) as a method for structuring the NBC network and (J), (K), (L), (M) show examples for two-photon polymerization (TPP) evaluated for writing error-free pass junctions in three dimensions. Figure 3 presents a histogram of agent exit passages after their exploration of a space-encoded NBC network decoding a four-instance  $\{3, 5, 6, 10\}$  SSP problem. It becomes clear that correct and incorrect solutions can be discriminated by a clear threshold (cf. Ref. [27] for details).

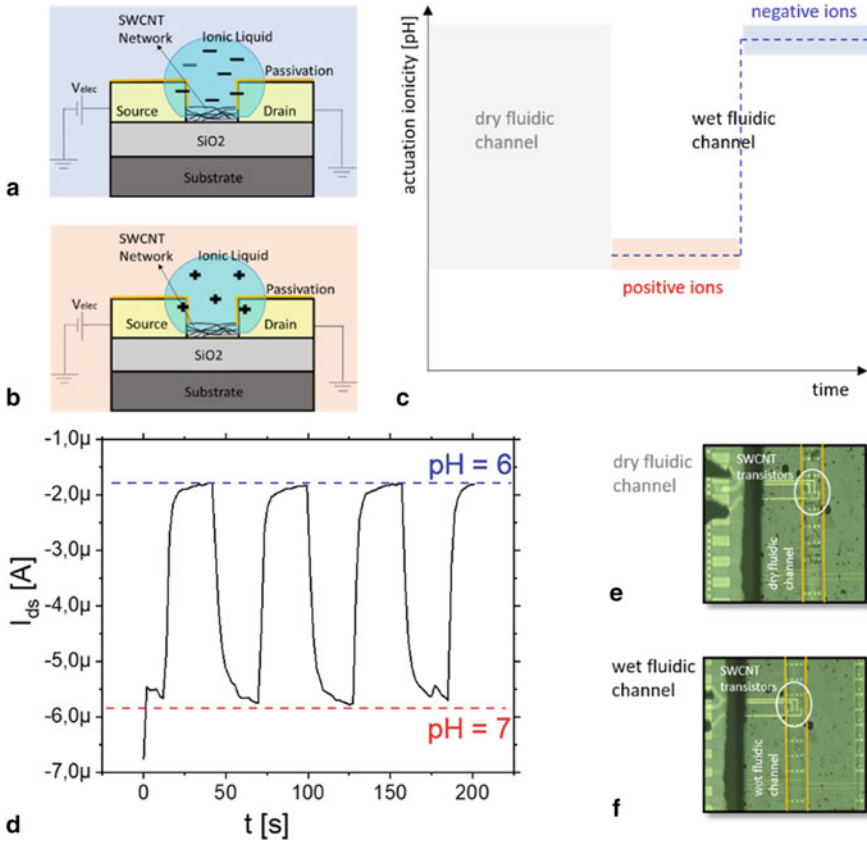
Contemporary NBC uses fluorescence interference contrast microscopy based on wide-field optical microscopy to detect the passage of the agents from the NBC network to the exits [16]. Alternative detection techniques that pass the information from the agents to the peripheral computing system. This paradigm includes to conceptualize integration of automatic path-specific read-out of the agents on

chip, using local optical or plasmonic techniques or alternative detection methods using electrical, biochemical, or magnetic information. The latter approaches have to operate in biological environments e.g. buffer solutions at premier stability and reliability.

## 2.4 *Integrated Electronic Detectors in Microfluidic Channels Actuated by Varying Ionicity*

Architectural elements for NBC require a resilient (with respect to long-term, repetitive multi-time operability) and reliable (with respect to a low error rate) function in biological and bio-inspired, yet technically fabricated, environments, whence the technical definitions for *resilience* and *reliability* still require clear definitions for NBC. Nanoelectronic electrical detectors are one candidate to monitor the trajectory of agents through the NBC system e.g. by dielectric triggering. The online monitoring of the pH value in microfluidic environments can be considered an intermediate step on the way to passage charge detectors, with the time-dependent background charge originating from the carrier liquid anyhow considered subject to monitoring. In the light of this aim, we performed a case study using existing technology oriented to check the resilience and operational stability of nanoelectronic detectors in fluidic environments at variable ionicity. Figure 4 shows a case study on the electrical characteristics of such detectors immersed in a microfluidic system actuating in a cycling manner with variable pH values. Carbon-nanotube field-effect transistors (CNT-FETs) with single-walled carbon nanotubes as transistor material were fabricated on silicon wafers according to protocols established by Hermann et al. [28–30]. Experimental details are found elsewhere [31]. The design was chosen in a way that, on roof-top of the transistors, microfluidic actuation of an individual (or a group of) CNT-FETs and the real-time electrical analysis of the switching properties was possible.

A key and crucial step was the implementation of a patterned dielectric material (metal oxide) on top of the source and drain electrodes in order to prevent leakage currents from the gate (as a reference electrode in contact with the fluid) to source and/or drain. In Fig. 4, parts **a** to **c** indicate the actuation principle for the nanoelectronic detectors in the microfluidic environment. Part **d** shows that the CNT-FETs addressed in an alternating flow of aqueous buffer of variable pH value shifts its drain current accordingly. Note that, for this case study, the geometrical constraints **e**, **f** were kept factors 3 to 10 above the target constraints already implemented in the NBC platform introduced above to enable standard wiring procedures for the electrical mapping, hence the involved absolute values are not yet be comparable to the NBC systems. Further experiments are ongoing to address accurate buffers and downscaling effects.



**Fig. 4** Concept for electronic detection in pH environments. Field-effect transistors (FETs) use integrated single-walled carbon nanotubes (SWCNTs) as transistor channel material. The source and drain electrodes are passivated with metal oxides towards the microfluidic channel so that the SWCNT-FETs can act as charge detectors in microfluidic environments. **a** Sketch of actuation with a basic liquid. **b** Sketch of actuation with an acidic liquid. **c** Principle of an actuation timeline during a microfluidic experiment. **d** Time series of an aqueous solution actuated through the microfluidic channel and the responding change of the drain current  $I_{ds}$  with time of the SWCNT-FETs (Gate voltage: fixed to the reference electrode in the liquid reservoir). A clear correlation between the drain current and the pH value of the actuating fluid can be seen. Insets **e** and **f**: light microscope images of electrode structures (yellow) to SWCNT-FETs (not visible) in a dry (left) and a wet (right) microfluidic channel (highlighted orange) (base length: ca. 500  $\mu\text{m}$ , width of microfluidic channel: ca. 200  $\mu\text{m}$ ). Copyrights: **a** to **c**, **e** and **f**: by the Authors, featured in A. Jain et al., Ref. [31], **d**: S. Hermann (TU Chemnitz)

### 3 Summary and Outlook

Results with the NBC devices designed and manufactured in the project show a successful implementation of formal semantics and verification of NBC circuits [32] for the solution of integer SAT problems. To date, for evaluation of the calculations,



passages of the labelled agents are detected using fluorescence interference contrast microscopy [16]. Alternative designs for space-encoded and agent-encoded NBC, new detection methods and advanced concepts for systems integration on the micro- and nanoscopic scales are challenging tasks for research and technology to push NBC as an alternative, energy-efficient computing platform.

**Acknowledgements** This work received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement 732842 (project Bio4Comp), the German Science Foundation (Deutsche Forschungsgemeinschaft, DFG) within the Research Area FOR 1713 "Sensoric Micro and Nano Systems" and the Lusitan Coal Foundation (Stiftung Lausitzer Braunkohle) by an academic school-scholarship (A.J.).

## References

1. Vandersypen, L.M.K., Steffen, M., Breyta, G., Yannoni, C.S., Sherwood, M.H., Chuang, I.L.: Experimental realization of Shor's quantum factoring algorithm using nuclear magnetic resonance. *Nature* **414**(6866), 883–887 (2001)
2. Xu, N., Zhu, J., Lu, D., Zhou, X., Peng, X., Du, J.: Quantum factorization of 143 on a dipolar-coupling nuclear magnetic resonance system. *Phys. Rev. Lett.* **108**(13), 130501 (2012)
3. Shor, P.W.: Polynomial-time algorithms for prime factorization and discrete logarithms in a quantum computer. *SIAM J. Comput.* **26**(5), 1484–1509 (1997)
4. Bengtsson, A., Vikstål, P., Warren, C., Svensson, M., Gu, X., Frisk Kockum, A., Krantz, P., Križan, C., Shiri, d., Svensson, I.-M., Tancredi, G., Johansson, G., Delsing, P., Ferrini, G., Bylander, J.: improved success probability with greater circuit depth for the quantum approximate optimization algorithm. *Phys. Rev. Appl.* **14**, 034010 (2020)
5. Adleman, L.M.: Molecular computation of solutions to combinatorial problems. *Science* **266**(5187), 1021–1024 (1994)
6. Boneh, D., Dunworth, C., Lipton, R.J., Sgall, J.Í.: On the computational power of DNA. *Discret. Appl. Math.* **71**(1–3), 79–94 (1996)
7. Braich, R., Chelyapov, N., Johnson, C., Rothmund, P., Adleman, L.: Solution of a 20 variable 3-SAT problem on a molecular computer. *Science* **296**(5567), 499–502 (2002)
8. Hanson, K.L., Nicolau, D.V., Jr., Filipponi, L., Wang, L.-S., Lee, A.P., Nicolau, D.V., Sr.: Fungi use efficient algorithms for the exploration of microfluidic networks. *Small* **2**(10), 1212–1220 (2006)
9. Sundberg, M., Bunk, R., Albet-Torres, N., Kvennefors, A., Persson, F., Montelius, L., Nicholls, I.A., Ghatnekar-Nilsson, S., Omling, P., Tågerud, S., Månsson, A.: Actin filament guidance on a chip: toward high-throughput assays and lab-on-a-chip applications. *Langmuir* **22**, 7286–7295 (2006)
10. Nicolau, Jr., D.V., Lard, M., Korten, T., van Delft, F.C.M.J.M., Persson, M., Bengtsson, E., Månsson, A., Diez, S., Linke, H., Nicolau, Sr., D.V.: Parallel computation with molecular-motor-propelled agents in nanofabricated networks. *Proc. Natl. Ac. Sci.* **113**(10), 2591–2596 (2016)
11. Rahman, M.A., Reuther, C., Lindberg, F.W., Mengoni, M., Salhotra, A., Heldt, G., Linke, H., Diez, S., Månsson, A.: Regeneration of assembled, molecular-motor-based Bionanodevices. *Nano Lett.* **19**(10), 7155–7163 (2019)
12. Cai, L.-M., Chen, J.-N.: on fixed-parameter tractability and approximability of NP optimization problems. *J. Comput. Syst. Sci.* **54**, 465–474 (1997)
13. van Delft, F.C.M.J.M., Ipolitti, G., Nicolau Jr., D.V., Sudalaiyadum Perumal, A., Kašpar, O., Kheireddine, S., Wachsmann-Hogiu, S., Nicolau Sr., D.V.: Something has to give:

- scaling combinatorial computing by biological agents exploring physical networks encoding NP-complete problems. *Interface Focus* **8**(6), 20180034 (2018)
14. Konopik, M., Korten, T., Lutz, E., Linke, H.: Fundamental energy cost of finite-time computing, [arXiv:2101.07075](https://arxiv.org/abs/2101.07075) [cond-mat.stat-mech] (2021)
  15. Mertens, S.: Computational complexity for physicists. *Comput. Sci. Eng.* **4**(3), 31–47 (2002)
  16. Nitzsche, B., Bormuth, V., Bräuer, C., Howard, J., Ionov, L., Kerssemakers, J., Korten, T., Leduc, C., Ruhnow, F., Diez, S.: Studying kinesin motors by optical 3D-nanometry in gliding motility assays. *Methods Cell Biol.* **95**, 247–271 (2010)
  17. Korten, T., Chaudhuri, S., Tavkin, E., Braun, M., Diez, S.: Kinesin-1 expressed in insect cells improves microtubule in vitro gliding performance, long-term stability and guiding efficiency in nanostructures. *IEEE Trans. Nanobiosci.* **15**(1), 62–69 (2016)
  18. Kumar, S., ten Siethoff, L., Persson, M., Lard, M., te Kronnie, Geertruy, Linke, H., Månsson, A.: Antibodies covalently immobilized on actin filaments for fast myosin driven Analyte transport. *PLoS ONE* **7**(10), e46298 (2012)
  19. Lindberg, F.W., Norrby, M., Rahman, M.A., Salhotra, A., Takatsuki, H., Jeppesen, S., Linke, H., Månsson, A.: Controlled surface Silanization for actin-myosin and biocompatibility of new polymer resists. *Langmuir* **34**, 8777–8784 (2018)
  20. Steenhusen, S., Burmeister, F., Groß, M., Domann, G., Houbertz, R., Nolte, S.: Heterogeneous microoptical structures with sub-micrometer precision. *Thin Solid Films* **668**, 74–80 (2018)
  21. Heldt, G., Meinecke, C.R., Steenhusen, S., Korten, T., Groß, M., Domann, G., Lindberg, F.W., Reuter, D., Diez, S., Linke, H., Schulz, S.E.: Approach to combine electron-beam lithography and two-photon polymerization for enhanced nanochannels in network-based biocomputation devices. *Proceedings SPIE 34th European Mask and Lithography Conference 2019*, pp. 175–182 (2019)
  22. Micolich, A. P.: Colorimetric path tagging of filaments using DNA-based metafluorophores, [arXiv:1909.06024v1](https://arxiv.org/abs/1909.06024v1) [cond-mat.soft] (2018)
  23. Zen'kevich, É.I., Blaudeck, T., Heidernätsch, M., Cichos, F., von Borczyskowski, C.: Effects of electron tunneling and nonresonance quenching of photoluminescence in semiconducting CdSe/ZnS and CdSe nanocrystals by porphyrin molecules in joint complexes. *Theor. Exp. Chem.* **45**, 23–34 (2009)
  24. Zen'kevich, É.I., Stupak, A.P., Kowanko, D., von Borczyskowski, C.: Temperature dependence and change over time of the spectral properties of CdSe/ZnS–organic dye nanocomposite quantum dots in solution and single nanostructures. *Theor. Exp. Chem.* **48**, 21–32 (2012)
  25. Motevich, I.G., Zen'kevich, É.I., Stroyuk, A.L., Raevskaya, A.E., Kulikova, O.M., Sheinin, V.B., Koifman, O.I., Zahn, D.R.T., Strekal, N.D.: Influence of the pH and polyelectrolytes on the spectral-kinetic properties of the AIS/ZnS semiconductor quantum dots in water solutions. *J. Appl. Spectr.* **87**(6), 926–935 (2020)
  26. Martello, S., Toth, P., 4-subset-sum problem. In: Martello, S., Toth, P. (Eds.), *Knapsack problems: Algorithms and Computer Interpretations*, pp. 105–136. Wiley, New York (1990)
  27. Meinecke, C.R., Korten, T., Heldt, G., Reuter, D., Diez, S., Schulz, S.E.: Fabrication and operation of protein-powered Biocomputation using nanostructured networks. *Smart Syst. Integr.* **2018**, 102–109 (2018)
  28. Blaudeck, T., Adner, D., Hermann, S., Lang, H., Gessner, T., Schulz, S.E.: Wafer-level decoration of carbon nanotubes in field-effect transistor geometry with preformed gold nanoparticles using a microfluidic approach. *Microelectron. Eng.* **137**, 135 (2015)
  29. Rodriguez, R.D., Blaudeck, T., Kalbacova, J., Sheremet, E., Schulze, S., Adner, D., Hermann, S., Hietschold, M., Lang, H., Schulz, S.E., Zahn, D.R.T.: Metal nanoparticles reveal the organization of single-walled carbon nanotubes in bundles. *RSC Adv.* **6**, 15753 (2016)
  30. Blaudeck, T., Preuß, A., Scharf, S., Notz, S., Kossmann, A., Hartmann, S., Kasper, L., Mendes, R.G., Gemming, T., Hermann, S., Lang, H., Schulz, S.E.: Photosensitive field-effect transistors made from semiconducting carbon nanotubes and non-covalently attached gold nanoparticles. *Phys. Status Solidi A* **216**(19), 1900030 (2019)
  31. Jain, A., Blaudeck, T., Hermann, S., Schulz, S. E.: Carbon-Nanotube Field-Effect Transistors as Charge Detectors in Liquid Environments: Design and Proof-of-Principle, Technical Report, Chemnitz University of Technology, Center for Microtechnologies (2018)

32. Aluf-Medina, M., Korten, T., Raviv, A., Nicolau, D.V., Jr., Kugler, H.: Formal Semantics and Verification of Network-Based Biocomputation Circuits. In: Henglein, F., Shoham, S., Vize!, Y. (eds.) Proc. Verification, Model Checking, and Abstract Interpretation 2021, Lecture Notes in Computer Science, vol. 12597, pp. 464–485. Springer, Cham (2021)
33. Steenhusen, S.: 3D printing goes micro, *Physics World*, Optics and Photonics Focus (2017). <https://physicsworld.com/a/3d-printing-goes-micro/>
34. Burmeister, F., Steenhusen, S., Houbertz, R., Asche, T. S., Nickel, J., Nolte, S., Tucher, N., Josten, P., Obel, K., Wolter, H., Fessel, S., Schneider, A. M., Gärtner, K.-H., Beck, C., Behrens, P., Tünnermann, A., Walles, H.: Two-photon polymerization of inorganic-organic polymers for biomedical and microoptical applications. In: König, K., Ostendorf, A. (eds.), *Optically Induced Nanostructures Biomedical and Technical Applications*, chapter 5. Walter de Gruyter, Berlin (2015)

# Artificial Immune Systems—Models and Applications



Yuriy Skobtsov 

**Abstract** The main models of artificial immune systems are considered: algorithms for clonal and negative selection, network models. Their applications are also presented: computer security, anomaly and malfunction detection, numerical and combinatorial optimization, pattern recognition. An algorithm for clonal selection is presented, a conclusion is made about its closeness to the method of evolutionary strategies of evolutionary computations with a more developed mutation operator, adaptive control of mutation. Two properties of the negative selection algorithm are highlighted: (1) the basic concept is to use a complement to the set of own cells; (2) the goal is to learn how to differentiate cells: “ours” or “aliens” with access only to samples of “own” cells. A comparative table of the temporal and spatial complexity of the generation of detectors is presented. It is shown that the positive selection method provides more accurate results, but is more time and memory-consuming than the negative selection method. Both approaches can be used in different settings. Various applications with a large amount of “friendly” data are more preferable for the negative selection method. For the network model, three types of mutation operators have been introduced: multipoint mutation, substring regeneration, and simple replacement. The chapter concludes with a comparative description of immunological computational models and concepts.

**Keywords** Immune computation · Clonal selection · Negative selection · Idiopathic network

## 1 Introduction

Artificial immune systems [1] (also sometimes called immunological computing [2]) is a field of artificial intelligence dedicated to the development of computational models based on the principles of the biological immune system. Obviously,

---

Y. Skobtsov (✉)

Saint Petersburg State University of Aerospace Instrumentation, 67 Bolshaya Morskaya str, St. Petersburg, Russia

e-mail: [ya\\_skobtsov@list.ru](mailto:ya_skobtsov@list.ru)

in the process of evolution, nature has created a very effective system of various mechanisms, that are capable to protect itself from a wide range of pathogens, such as bacteria, fungi, parasites, etc. The powerful capabilities of the immune system to process information, such as function mining, pattern recognition, learning, memory, and its distributed nature promise great opportunities for its artificial analog. In general, the immune system is able to recognize dangerous elements and take the appropriate response, at the same time it is “tolerant” of its own molecules and ignores many harmless substances.

## 2 Basic Models

Artificial immune systems (AIS) or immunological computing are developing rapidly and there is growing interest in the development of computational models based on immunological principles. Some models imitate abstract mechanisms in the biological immune system in order to better understand its natural processes and simulate its dynamic behavior in the presence of antigens or pathogens. Others focus on computational algorithm design, methods that use simplified concepts of various immunological processes and functions. Note that AIS (at least some models—for example, the clonal selection algorithm) in their ideology are quite close to evolutionary computations (EC) [3]. In the process of solution searching, they use a population of potential solutions, each of which is represented by an artificial lymphocyte (AL)—an analog of an individual in EC. At the same time, just like an individual in EC, an artificial lymphocyte is represented (encoded) most often by a binary string or a vector of real numbers. When forming the next population, more advanced mutation operators are applied in AIS. The quality of AL is evaluated by comparison using the affinity function (analog of the fitness function in EC) [3]. To build an AIS model, there are several main aspects that must be taken into account [2]:

- There should be trained detectors (artificial lymphocytes) that detect “non-self” patterns with a certain affinity (similarity).
- AIS may require a good repository of one’s own or someone else’s patterns for training artificial lymphocytes (IL).
- A tool for measuring the similarity (affinity) between AL and the pattern is required. The measured affinity should show a pattern detection rate.
- To be able to measure the similarity (affinity), the representation (coding) of patterns (potential solutions) and AL must have the same structure.
- It is necessary to measure the affinity between two ALs. The measured affinity indicates the extent to which ALs can communicate with each other to form a network.
- AIS must have memory, which is created by artificial lymphocytes, which often show patterns that are not associated with themselves.
- When AL detects “non-self” patterns, they can be cloned, and clones can be mutated to have more variety in the search space.

### 3 Clonal Selection Algorithm

This algorithm is based on the theory of clonal selection [4], which suggests that lymphocytes (B cells and T cells) are used to destroy or neutralize an antigen (pathogen). When a lymphocyte is selected and associated with an antigen, it multiplies and differentiates into plasma and memory cells. Plasma cells have a short lifespan and produce large numbers of antibody molecules, while memory cells live for a long period, expecting the same antigen in the future. Based on the principles of Darwin's evolutionary theory of natural selection, this theory uses the cloning of selected lymphocytes. Further, clones undergo mutation, which increases the efficiency of antigen fighting. Below is the clonal selection algorithm.

Begin

Initialization: determination of parameters such as repertoire size  $M$ , stopping criterion (for example, the maximum number of generations  $\text{maxgen}$ ), cloning factor  $\beta$ , etc.

Generation of a random initial repertoire of antibodies, evaluation of the affinity of antibodies,  $\text{gen} = 0$ .

While (stop criterion not met—for example  $\text{gen} > \text{maxgen}$ ).

Generation of round  $(\beta \cdot M)$  identical clones by each antibody.

This increases the size of the repertoire after cloning  $N_c$  as follows

$$N_c = \sum_{i=1}^M \text{round}(\beta \cdot M) + M$$

Somatic hypermutation execution for clones

(that is, the degree of the mutation of a similar clone (having a high affinity is less, while the amount of the original antibodies remains unchanged).

This stage is also called the process of affinity maturation).

Evaluation of the affinity measure for mutants.

Selection of the antibodies with the highest affinity among the  $N_c$  antibodies and discarding the rest antibodies.

$\text{gen} = \text{gen} + 1$ .

End.

Output: generated antibody repertoire.

Note that this algorithm is very similar to the method of evolutionary strategies (ES) of EC with a more developed mutation operator and it can be roughly considered as a parallel version  $(1 + \text{round}(\beta \cdot M))$  [3]—ES with adaptive control of mutations. Note that versions of clonal selection have been developed with both discrete and real coding [4, 5]. This algorithm is most often used to solve optimization problems.

## 4 Negative Selection Algorithm

This algorithm [6, 7] is based on the principle of recognizing «self- nonself” in the immune system and is the most popular in AIS applications. In the natural immune system, such recognition is provided by T-lymphocytes and other cells that have receptors on their surface that can detect foreign proteins (antigens). How to distinguish self-proteins from pathogenic foreign proteins using T cells? Once in the thymus (the thymus gland is the central organ of the immune system), T cells undergo negative selection, which consists in the fact that the cells that have reacted with their own proteins are destroyed. While the rest cells (which did not reacted) are given the opportunity to leave the thymus. These T cells then circulate throughout the body and act as a defense against antigens.

The negative selection algorithm (NSA) works in a similar way, randomly creating detectors and removing those, that recognize "self" cells so that the remaining detectors can detect any "non-self" cells. The main goal of the NSA is to cover the "non-self" space (antigens) with a set of detectors.

The two most important aspects of NSA are:

1. The basic concept of the algorithm is the use of a complement to the set of “self” (cells).
2. The goal is to learn to distinguish between “self” and “non-self” cells, while only samples of “self” cells are available.

NSA has two stages [1, 2]: (1) “generation of detectors” and (2) detection of “non-self”. At the first stage, a detectors’ set is generated by some randomized process, which uses a set of "self" -Sself as input. Candidates for detectors that detect any "self" elements are removed, and detectors that do not recognize "self" are retained. The pseudocode of the simplest detector selection algorithm is presented below.

```

input: Sself - a set of "self" elements
Begin
  population = {}
  While (stop criterion not met)
    New_antibody = random_antibody_generation
    match = False
    For each {s | s ∈ Sself}
      If there is a match (s, new_antibody)
        match = True
    End
    If there is no match
      population = population ∪ new_antibody
    End
  End
End
Output: detectors set.

```

**Table 1** Time and space complexity of detector generation

Algorithm	Time	Space
Exhaustive search	$O(m^l \cdot N_S)$	$O(l \cdot N_S)$
Linear	$O((l-r+1) \cdot N_S \cdot m^r) + O(l-r+1) \cdot m^r + O(l \cdot N_R)$	$O((l-r+1)^2 \cdot m^r)$
Greedy	$O((l-r+1) \cdot N_S \cdot m^r) + O((l-r+1) \cdot m^r) \cdot N_R$	$O((l-r+1)^2 \cdot m^r)$
$N_S$ —mutation	$O(m^l \cdot N_S) + O(N_R \cdot m^r) + (N_R)$	$O(l \cdot (N_S + N_R))$

In fact, to generate detectors when solving real problems, various algorithms are used: random, greedy, evolutionary, dynamic programming, etc., which have different complexity and adequacy [2, 8]. The time and space complexity of the main algorithms for generating detectors are shown in Table 1, where  $m$  is the alphabet cardinality,  $l$  is the string,  $r$  is the matching threshold,  $N_S$  is the number of elements in the set of "self",  $N_R$  is the number of detectors.

In the detection phase, the detectors set (generated in the first phase and saved for later use) are used to check if new incoming patterns match "self" or "non-self" patterns. If the input sample matches the detector, it is identified as part of the "non-self", which in most applications means that there is an anomaly/change. A particular NSA is characterized by how the detectors are presented, what rules are used to determine the match between the sample and the detectors, and the mechanisms for creating and removing "self—non-self" detectors. In most works on NSA, potential solutions are encoded either by binary strings or real vectors, although more complex constructions are also used.

## 5 Positive Selection

Positive selection is also used in the natural immune system and, although not as widely as negative selection. A positive selection is made based on filter receptors found on the surface of immature T cells. A receptor is an MHC molecule that allows the body to filter out immature T cells. Positive selection works similarly to the negative selection algorithm, but instead of removing antibodies to the population, if they match, on the contrary, they are added to the repertoire. In contrast to negative selection, "positive detection methods" are widely used in pattern recognition, clustering, and other fields, where they generate the detectors set that match (recognize) "self" patterns (instead of "non-self"). In this case, the model of the "self" set (training data) is used to classify the sample as part of "self" or "non-self". A simple positive detection model can be built using the nearest neighbor approach. If the pattern is in the vicinity of the standard, then it is marked as belonging to this set. As a rule, a positive detector determines the surroundings, assuming a hypersphere with a certain radius centered in each of its "self" patterns. Moreover, detectors can be defined in a more complex way, using some clustering algorithm on the elements of the set of "self". Thus, a recognized pattern can be classified as belonging to a cluster by measuring



its distance, for example, in the Minkowski metric. Dasgupta and Gonzalez [2, 9] compared negative and positive selection methods. Although the positive selection was more accurate, it was more time and memory-consuming than the negative selection. Both approaches can be used in different settings. Numerous applications with a lot of data “self” seem to be more suitable for the negative selection method.

## 6 Immune Network Models

These models suggest a hypothesis that the immune system is a regulated network of molecules and cells that recognize each other even in the absence of antigen [12]. Such structures are often referred to as idio-cy nets, which serve as the basis for studying the behavior of the immune system. Network AIS models can be divided into two categories: continuous and discrete models [1, 2]. In continuous models, the immune response is assumed to be continuous, in contrast to discrete models, where it occurs in discrete time steps. Continuous models based on differential equations do not focus on the structure of IS but consider the concentrations of antibodies and foreign antigens, although they assume that all antibodies interact with each other and antigens also interact with all antibodies. The change in the concentration of a specific antibody  $x_i$  is presented as the sum:  $dx_i / dt = \text{internal network dynamics} + \text{dynamics driven by the antigen}$ . The first term models the interaction between antibodies, on the contrary, the second term is the stimulation of antibodies by antigens.

In Erne’s model [10], the following equation is used to describe the change in the concentration of lymphocytes of a certain type:

$$\frac{dx_i}{dt} = x_i \sum_{j=1}^N f(E_j, K_j, t) - x_i \sum_{j=1}^N g(I_j, K_j, t) + k_1 - k_2 x_i,$$

where the first term represents the general stimulation of type  $i$  lymphocytes by excitation of signals in the form of the sum of excitatory signals received from stimulating lymphocytes. Accordingly,  $f(E_j, K_j, t)$  is a measure of the signal excitation from idiotypes in  $E_j$  on type  $i$  lymphocyte at time  $t$ . Here  $K_i$  is a constant associated with the strength of affinity between type  $i$  lymphocytes and idiotypes in  $E_j$ . Similarly, the second term expresses the overall effect of inhibitory signals from other lymphocytes on type  $i$  lymphocytes. Thus,  $I_j$  expresses a lymphocyte that (by site-bonding) recognizes idiotypes on type  $i$  cells. In addition,  $k_1$  is the rate at which type  $i$  lymphocytes enter the network, and  $k_2$  is the ratio that determines the natural mortality of lymphocytes Type  $i$  in the absence of antigen. In this model, the differential equation describes the change in the concentration of lymphocytes of each type. Thus, the network demonstrates dynamic behavior even in the absence of stimulating antigens. To describe the dynamic behavior of a foreign antigen, it is necessary to

include an additional term representing the interaction of the corresponding type I lymphocyte with external antigens.

However, discrete models are usually abstract functional models that are used by AIS, and their purpose is to solve real computational problems. Discrete models consider the AIS as a set of B-cells that interact with each other in accordance with their affinity. The first discrete model of IIS [11] considers IIS as a set of B-cells that interact with each other in accordance with their affinity. B-cells are represented as binary strings, following some earlier work [12]. Therefore, the affinity between B cells is determined based on the Hamming distance. If the stimulation of B cells with foreign antigens is above a certain threshold, then they clone and mutate. Cloning produces a certain number of exact copies of a B cell. However, the number of copies depends on the level of B cell stimulation. In addition, in the simple substitution operator, a small (less than half) part of the substring representing the B-cell is replaced by the corresponding elements of another randomly selected B-cell. Also, three types of mutation operators have been introduced: multipoint mutation, substring regeneration, and simple substitution.

## 7 The Danger Theory

The theory of danger (of the natural immune system) [13] suggests that the immune system is activated when it receives molecular signals that indicate damage (or stress) of the body, and not on the comparison of “self”—“nonself” patterns. Accordingly, damaged cells and tissues transmit danger signals, which leads to the capture of antigens by antigen-presenting cells (APCs), such as macrophages. These cells then travel to the local lymph node and present the antigen to the lymphocytes. Essentially, a danger zone exists around every danger signal. Consequently, only those B cells whose antibodies correspond to antigens in the danger zone will be stimulated, and then will undergo clonal expansion. Eikelin and Kaiser [14] include the following aspects of danger theory in their principles of creating an artificial immune system: it is necessary to model the appropriate APCs to signal danger; a danger signal can be either positive or negative, which means the presence or absence of a signal; Although in biology the danger zone is spatial, the computational model can use other concepts of proximity, such as, for example, temporal proximity.

## 8 Comparison of the Main Models

Table 2 presents the most common computational models of IMS, which shows the use of specific immunological concepts in various models and their main applications.

**Table 2** Immunological computational models and concepts

Criteria	Clonal selection	Negative selection	Network models
Immunological concepts and essences	Clonal expansion, maturation of affinity, B cell	Recognition of "self -nonself " based on T cells	Idiotypic network, immune memory, B cell
Data types	Binary strings, Real vectors	Binary strings, Real vectors	Binary strings, Real vectors
Elements	Epitopes, plasma cells, memory cells	Detectors, receptors, anti-genes	B lymphocytes, Y antibodies
Operations	Clonal selection, somatic mutation, calculation of affinity, proliferation, differentiation	Negative selection, match, binding, generation of detectors, affinity calculation	Cloning, mutation, rejection, and growing of arcs based on affinity, stimulation, and suppression to stabilize the network
Benefits	Increasing the repertoire diversity protects against previously not encountered antigens and forms a more effective secondary response. Good parallelization. Few user parameters, low complexity	Ability to distinguish between "self-non-self". No prior knowledge of the multitude of "non-self" elements is needed. The existence of the algorithms for the detectors generation with varying complexity. Ability to detect anomalies. Good parallelization	Continuous adaptation of the network to maintain a steady state. Flexible selection mechanism, autonomous and fully decentralized
Disadvantages	There is an early convergence to local extrema. An accurate evaluation of the repertoire power is desirable. Weak feedback	It is inappropriate if the set of "self" is small or the search space is infinite. The standard algorithm for detectors generating has exponential complexity. The usual software is required for the iterative increase of the "self" samples	Limited applicability—high overhead costs, computational complexity and misunderstanding of their dynamics. Application of the properties of excitation and suppression of the network model is supported by the usual software

(continued)

**Table 2** (continued)

Criteria	Clonal selection	Negative selection	Network models
Applications	Numerical and combinatorial optimization, pattern recognition, classification, pattern recognition, machine learning, numerical data classification. Flexible alternative to the genetic algorithm	Detection of anomalies, malfunctions, changes, computer security, protection against viruses and network fraud, intrusions, binary classification	Machine learning (supervised, unsupervised), management, clustering, data extraction, text classification, pattern recognition

## 9 Applications

As a tool, AIS are used in real applications in the following areas: computer security, fraud detection, robotics, fault detection, data mining, text analysis, pattern and image recognition, bioinformatics, games, planning, optimization, classification, clustering, anomaly detection, machine learning, adaptive control, and associative memory, etc., which are shown in Table 2. They are also used in combination with other techniques such as genetic algorithms, neural networks, fuzzy logic, and swarm intelligence. A general description of the solution process using models of the immune system is provided by some general applications, such as, for example, ARTIS [15] and LISYS [16]. In addition, some AIS applications are briefly described below to demonstrate how these methods can be used to solve real-world problems.

ARTIS (Artificial Immune System) was developed by Hofmeyr and Forrest [15] in 2000. It is an AIS that simulates many processes and properties of the immune system of vertebrates, including some of the concepts and algorithms described in the previous sections. The purpose of ARTIS is to point out the elements of a general adaptive distributed system without reference to any particular application. These common elements must then be specified according to the characteristics of the application. LISYS [16] (Light Intrusion Detection System), a network intrusion detection system based on ARTIS, is considered as an example of this specification. Each node in the protected system is a networked computer that has a local collection of receptors and a local sensitivity level. Antigens to be monitored by detectors are strings containing information about network traffic that affects protected nodes. The detection of abnormal lines results in the generation of an alarm for the human operator.

Baltrop et al. [17] used a version of LISYS to monitor network traffic. The system used a negative selection algorithm (to create mature 49-bit binary detectors, that is, triplets representing Transmission Control Protocol [TCP] connections that were checked for connections collected during the training period. The matured detectors were then distributed to each host in the existing one. At the same time, diversity is

created through each host, independently reacting to "own" and "alien" (normal and abnormal).

Computer security is in many respects similar to biological security [6, 18, 19]. In this way, we can learn lessons and useful knowledge from the natural immune system to boost digital immunity. Most of the first works of AIS were devoted to the use of some immunological models for the development of the protection of digital systems [6, 7, 19]. AIS used various aspects of data protection and search for anomalies while providing a universal protection system and improving existing computer security systems. The security of computer systems depends on actions such as detecting unauthorized use of computing equipment, maintaining the integrity of data, files, and preventing the spread of computer viruses. Forrest et al. [6] first proposed the use of negative selection in computer security. They proposed to consider the problem of protecting computer systems from harmful viruses as a special case of the general problem of recognizing "ours" (legitimate users, integral data, etc.)—(dangerous) "aliens" (unauthorized users, viruses, and other malicious agents). This method was intended to complement the more traditional methods of cryptographic and deterministic file authentication for the problem of detecting computer viruses.

Kephart [20] proposed another immunologically inspired approach (based on the invasion hypothesis) for virus detection. With this approach, known viruses are detected by their sequences of computer codes (signatures), and unknown viruses by their unusual behavior in the computer system. This virus detection system constantly scans your computer software for typical signs of a virus infection. These signatures trigger the launch of decoys whose sole purpose is to infect a virus.

An automatic detection and response system for detecting self-propagating malicious code and preventing its spread, called the Cooperative Automated worm (CARDINAL), was proposed in [21]. This method is based on the concept of differentiation of T-cell states. Specifically, three key properties of T cells have been identified: T cell proliferation to optimize the number of hosts interviewed, T cell differentiation to assess the severity and reliability of an attack, and T cell modulation and interaction to balance local and peer-to-peer information. The aim of this work was to use a variety of T cell types to act as a collaborative automated system for detecting and responding to worm-type viruses.

Bradley and Tyrrell in 2002. [22] applied AIS concepts to solving the problem of detection and elimination of malfunctions in digital electronic systems. The classical approaches to the detection and elimination of malfunctions in artificial systems are redundancy and the addition of protection systems that check and, possibly, correct the technical state of the system. The immunotronic approach defines a system of the same type but uses the AIS concept to recognize "self -nonself" and automate the generation of verification criteria used to protect the system. This approach is applicable to finite state machine models (FSM), a class of systems where operation is modeled in terms of states and transitions between them and encompassing most of the existing electronic systems. In this case, the set of "self" can be defined as a set of rows that represent legal transitions between states of the machine. For example, strings can be formed by concatenating a string containing the values of the current

input, a string containing the current state, and a string containing the next state generated by the automaton (Fig. 7.8).

Forrest et al. [19] applied a negative selection algorithm to monitor UNIX processes in a dynamic computing environment in such a way that self-determination is susceptible to malicious attacks. This work is based on the assumption that the system calls root processes, which are inherently more dangerous than user processes. In addition, root processes have a limited range of behavior, and their behavior is relatively stable over time; accordingly, "normal (or" friendly ") is determined by the correlation of "short-range" in the system calls of the process (called tracing). They have experimented with several common send mail-related intrusions, such as: send mail traces of successful (trails), traces of send mail failed intrusions and traces of conditional errors.

Gonzalez [23] proposed using the negative selection method to form attack detection rules (NSDR) when monitoring network traffic. The real encoding was used for developing hyper-rectangular detectors, interpreted in the form of "if-then-rules", for high-level characterization of the space of oneself "friend or foe" (that is, normal and abnormal traffic). The experiments were conducted using the 1999 Defense Advanced Research Projects Agency (DARPA) Intrusion Detection Assessment dataset. This data represents normal and abnormal information collected on the test net on which the attack was simulated. The immune genetic approach yielded detectors that gave a good estimate of the degree of deviation from the norm. In further works, this algorithm (NSDR) is extended by using fuzzy detection rules and is called NSFDR. This increases the accuracy of the method and gives a measure of deviation from the normal, so there is no need for a discrete division of the "friend or foe" space. This modification provides a better definition of the boundary between normal and abnormal. In another work, Kim and Bentley [24] used three evolutionary approaches: gene library evolution, negative selection, and clonal selection to design an effective network intrusion detection system. Here, detectors (in the form of classifiers) are developed using a clonal selection algorithm, in which the developing population of detectors is grouped into "niches" that help distinguish "friend or foe" in network traffic data.

Aicklen et al. [25] first proposed using the concept of Danger Theory (DT) for intrusion detection. Their system behaves like dendritic cells (DCs) looking for danger, observing signals such as a sudden increase in network traffic or an unusually high number of message errors. If these signals exceed a preset threshold, this will trigger a warning. An artificial immune system called JISYS has been developed to detect fraud [11, 26]. This system forms a network of artificial B-cell objects, where each B-cell represents an application for a loan. The system was developed in a subsequent work by Hunt [27, 28], which presents the results of fraud detection. The work presented by Neal [29] examines an immune-inspired (supervised) learning system called Immunos-81. Two standard machine learning datasets were used to test the recognition capabilities of the system. They use software abstractions (models) of T cells, B cells, antibodies, and their interactions. At the same time, artificial T-cells control the generation of populations of B-cells (clones), which fight for the recognition of "unknown" (strangers).

## 10 Conclusion

Models of artificial immune systems are being actively developed and applied in real applications in many areas. First of all, it is computer security, fraud detection, robotics, fault detection, data mining, etc. They are also used in combination with other computational intelligence techniques such as evolutionary algorithms, neural networks, fuzzy logic, and swarm intelligence. Such hybrid systems [30] are able to take advantage of individual strengths and overcome the shortcomings of different paradigms, thus offering powerful algorithms for solving complex problems. It should be noted that the application of artificial immune systems is promising for solving problems of the security for cyber-physical systems, such as, for example, considered in [31–33]. The use of immune algorithms for the detection of various viruses seems to be quite natural, similar to the work [34].

## References

1. Dasgupta, D.: *Artificial Immune Systems and their Applications*. Springer (1999)
2. Dasgupta, D., Luis, F.N.: *Immunological Computation-Theory and Applications*. CRC Press, Boca Raton, FL (2009)
3. Skobtsov Y.A, Speransky D.V.: *Evolutsionnye vychislenia: uchebnoe posobie [Evolutionary computation: hand book]*. Moscow: The National Open University “INTUIT”, 2nd edn (2016)
4. Castro, L.N.: von Zuben: learning and optimization using clonal selection principle. *IEEE Trans. Evoluti. Comput.* **6**, 239–251 (2002)
5. Yu, X., Gen, M.: *Introduction to Evolutionary Algorithms*. Springer (2010)
6. Forrest, S.: Self-nonsel self discrimination in a computer. In: Forrest, S., Perelson, A.S., Allen, L., Cherukuri, R. (eds.), *Proceedings of the 1992 IEEE Symposium on Security and Privacy*, pp. 202–212 (1994)
7. D’haeseleer, P., Forrest, S., Helman, P.: An immunological approach to change detection: Algorithms, analysis, and implications. *Proceedings of the 1996 IEEE Symposium on Computer Security and Privacy*, IEEE Computer Society Press, Los Alamitos, CA, pp. 110–119 (1996)
8. Yang, H., Li, T., Hu, X., Wang, F., Zou, Y.: A survey of artificial immune system based intrusion detection. *Sci. World J*, Article ID 156790 (2014)
9. Dasgupta, D., Gonzalez, F.: An immunity-based technique to characterize intrusion in computer networks. *IEEE Trans. Evol. Comput.* **6**(3), 1081–1088 (2002)
10. Jerne, N.K.: Towards a network theory of the immune system. *Ann. Immunol.* **125**C(1–2), 373–389 (1974)
11. Hunt, J.E., Cooke, D.E.: Learning using an artificial immune system. *J. Netw. Comput. Appl.* (special issue on Intelligent systems: Design and application) **19**, 189–212 (1996)
12. Farmer, J.D., Packard, N.H., Perelson, A.S.: The immune system, adaptation, and machine learning. *Phys. D* **2**(1–3), 187–204 (1986)
13. Matzinger, P.: Tolerance, danger and the extended family. *Annu. Rev. Immunol.* **12**, 991–1045 (1994)
14. Aickelin, U., Cayzer, S.: The danger theory and its application to artificial immune systems. *1st International Conference on Artificial Immune Systems*, Canterbury, U.K. (2002)
15. Hofmeyr, S.A., Forrest, S.: Architecture for an artificial immune system. *Envol. Comput.* **7**(1), 45–68 (2000)

16. Balthrop, J., Forrest, S., Glickman, M.R.: Revisiting LISYS: parameters and Normal Behavior. Published in the proceedings of the special sessions on artificial immune systems in the 2002 Congress on Evolutionary Computation, IEEE World Congress on Computational Intelligence, Honolulu, Hawaii, 20 (2002)
17. Balthrop, J., Esponda, F., Forrest, S., Glickman, M.R.: Coverage and Generalization in an Artificial Immune System (AAAA). Published in the proceedings of the International Conference Genetic and Evolutionary Computation (GECCO), New York, July 9–13 (2002)
18. Forrest, S., Hofmeyr, S., Somayaji, A.: Computer immunology. *Commn. ACM* **40**(10), 88–96 (1997)
19. Forrest, S., Hofmeyr, S.A., Somayaji, A., Longstaff, T.A.: A sense of self for Unix processes. Proceedings of 1996 IEEE Symposium on Computer Security and Privacy, Oakland, CA (1996)
20. Kephart, J.O.: A biologically inspired immune system for computers. In: Brooks, R.A., Maes, P. (Eds.), *Artificial Life IV. Proceedings of the 4th International Workshop on the Synthesis and Simulation of Living Systems*, MIT Press, Cambridge, MA, pp. 130–139 (1994)
21. Kim, J., Wilson, W.O., Aickelin, U., McLeod, J.: Cooperative automated worm response and detection immune algorithm (CARDINAL) inspired by T-cell immunity and tolerance. The Proceedings of the Fourth International Conference on Artificial Immune Systems, Banff, Alberta, Canada, pp. 168–181, August (2005)
22. Bradley, D.W., Tyrrell, A.M.: Immunotronics—novel finite-state-machine architectures with built-in self-test using self-nonsel self differentiation. *IEEE Trans. Evolut. Comput.* **6**(3), 227–238 (2002)
23. González, F., Dasgupta, D., Kozma, R.: Combining negative selection and classification techniques for anomaly detection. Proceedings of IEEE Congress on Evolutionary Computation, Honolulu, Hawaii, vol. 1, pp. 705–710, May (2002)
24. Kim, J., Bentley, P.J.: An evaluation of negative selection in an artificial immune system for network intrusion detection. Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2001), San Francisco, CA (2001)
25. Aickelin, U., Bentley, P., Cayzer, S., Kim, J., McLeod, J.: Danger theory: The link between ais and ids? Proceedings of the Second International Conference on Artificial Immune Systems (ICARIS 2003), vol. 2787 of LNCS, Springer-Verlag, pp. 147–155 (2003)
26. Hunt, J.E., Fellows, A.: Introducing an immune response into a CBR system for data mining. BCS ESG'96 Conference and Published as Research and Development in Expert Systems XIII (1996)
27. Hunt, J., Timmis, J., Cooke, D., Neal, M., King, C.: Jisys: the development of an artificial immune system for real world applications. In: Dasgupta, D. (ed.) *Applications of Artificial Immune Systems*, pp. 157–186. Springer, Berlin (1999)
28. Ishida, Y.: Fully distributed diagnosis by PDP learning algorithm: towards immune network pdp model. Proceedings of International Joint Conference on Neural Networks, San Diego, CA, pp. 777–782 (1990)
29. Neal, M., Hunt, J., Timmis, J.: Augmenting an artificial immune network. Proceedings of International Conference Systems and Man and Cybernetics, San Diego, CA, IEEE, pp. 3821–3826 (1998)
30. Skobtsov, Y.: Prospects of the interdisciplinary course “Computational intelligence” in engineering education. *Studies in Systems, Decision and Control*.-Volume 342. *Cyber-Physical Systems: Design and Application for Industry 4.0*, pp. 431–441. Springer (2021)
31. Rimsha, A.S., Rimsha, K.S.: The problem of selecting APCS' information security tools. *Studies in Systems, Decision and Control*. Volume 342. *Cyber-Physical Systems: Design and Application for Industry 4.0*, pp. 211–223. Springer (2021)
32. Kravets, A.G.: Natalia Salnikova, Kirill Dmitrenko and Mikhail Lempert.: *Industrial Cyber-Physical Systems: Risks Assessment and Attacks Modeling*. *Studies in Systems, Decision and Control*.-Volume 342. *Cyber-Physical Systems: Design and Application for Industry 4.0*, pp. 197–210. Springer (2021)



33. Kravets, A.G., Bui, N.D., Al-Ashval, M.: Mobile security solution for enterprise network. Knowledge-Based Software engineering. Communications in Computer and Information Science,466. 11th Joint Conference, JCKBSE 2014 Volgograd, Russia, September 17–20, 2014, Proceedings, pp. 371–382. Springer (2014)
34. Zharkikh, L.I., Smirnova, Yu.A., Azhmukhamedov, I.M., Golubkina, E.V., Trizno, M.N.: Inhibitors selection to influenza virus a by method of blocking intermolecular interaction. Studies in Systems, Decision and Control.-Volume 333. Cyberspace for Advanced Human-Centered Society, pp. 227–238. Springer (2021)

# The Analysis of Technology Development Trends Based on the Network Semantic Structure “Subject-Action-Object”



Veronika Kolesnikova, Dmitriy Korobkin , Sergey Fomenkov , Eduard Rayushkin , and Vsevolod Glushkin

**Abstract** The chapter presents the approaches and technologies used to solve the problem of analyzing trends in technology development based on the network semantic structure “Subject-Action-Object”. From the point of view of information about the invention itself, the most important is the description of the invention to the patent. In electronic databases of patents, all patents begin precisely with the description of the invention to the patent, which in turn has its title page. This form of the description of the invention to the patent is unified, and all patents are presented in this form, that is, all patents are equally structured. It is this block of the patent—information about the invention must be investigated using the network semantic structure “Subject-Action-Object”. To solve this problem, the structure of the patent was studied; Hadoop technologies, Spark MLlib, clustering methods. Grid computing technologies have been chosen as a successful and efficient means of processing large text data in the form of patents. An algorithm for parsing a patent document has been developed; an algorithm for preprocessing text documents of a patent selection; a Subject-Action-Object (SAO) extraction algorithm; an algorithm for forming a patent landscape for a certain period. The concept and architecture of the automated system have been formed, the proposed algorithms have been implemented in software.

**Keywords** SAO-structures · Technology development trends · Semantics · Automated system

---

V. Kolesnikova (✉) · D. Korobkin · S. Fomenkov · E. Rayushkin  
Volgograd State Technical University, 28 Lenina, Volgograd 400005, Russia  
e-mail: [verona.7@yandex.ru](mailto:verona.7@yandex.ru)

D. Korobkin  
e-mail: [dkorobkin80@mail.ru](mailto:dkorobkin80@mail.ru)

S. Fomenkov  
e-mail: [saf550@yandex.ru](mailto:saf550@yandex.ru)

V. Glushkin  
Volgograd State University, 100 Prospect Universitetsky, Volgograd 400062, Russia

## 1 Introduction

Modern methods of clustering the patent space use key terms in the form of unigrams or n-grams. The representation of the key features of a patent in the form of “Subject-Action-Object” structures will improve the accuracy of clustering the patent space and identifying patent trends [1–3].

Thus, the current task is to develop software for identifying trends in the development of technologies based on the network semantic structure “Subject-Action-Object” [4].

The existing analogs of the development trends analysis were compared (Table 1) [8–13]. Based on the results of the comparison, it can be concluded that at the moment the Questel Orbit system is the best information service in the world, which guarantees the maximum completeness and reliability of patent documentation research. However, it lacks a method for presenting the key features of a patent in the form of “Subject-Action-Object” structures, which will allow for a more accurate comparison of patents [5, 6].

## 2 Algorithm for Parsing English-Language Patents

Patent parsing is an important part of identifying patent trends. The next stages of patent trend analysis directly depend on the correct analysis of the patent at the initial stage. At the input, the algorithm receives USPTO patents in XML format, files are read from the catalog and stored in RDD using the `wholeTextFiles()` function [7].

The first step is to extract from the file the necessary fields containing the patent name, patent number, etc. Records from the RDD schema are preprocessed using the functions shown in Table 2.

After that, the patent text is processed using the “`preparing_text`” function: removing line breaks, removing references and numbering, separating sentences by punctuation marks and marker words.

At the output of this algorithm, we get a record in the database and a `DataFrame`, the fields of which are shown in Fig. 1.

The resulting text, located in the “`patent_description`” column, will pass to the input of the next algorithm for further syntactic processing. The patent parsing algorithm is shown in Fig. 2.

## 3 The Subject-Action-Object Extraction Algorithm (SAO)

The SAO extraction algorithm is a syntactic analysis of the sentences of the text obtained from the previous stage. The main task of the algorithm is to determine the semantic relations between words [14].

**Table 1** Comparison of analogs

Criteria	Questel orbit	Patent scope	Global patent index	Patent family group	Minesoft	STN AnaVist
Content of the patent database	60 million documents from 95 countries and International Patent offices	3.8 million published international applications, 84 million patents	World bibliographic data set from 87 states and 5 international organizations	Patents, R&D, and publications	Patent data from more than 106 countries	patents from databases: CAplus, USPAT2, USPATFULL
Ability to translate patents	Available in 30 languages	Has cross-language search	–	–	–	–
Types of patent searches	Simple, semantic, similarity search	Simple, advanced, cross-language search, chemical structure search	Multidimensional patent search, thematic search, search for patent analogs	Analytical report	Search for patents, IP documents, analytics in patents and competitive intelligence systems	Analysis of the patent environment, visualization of the information contained in the patent
Research of patent activity of competitors	Research by year, strategic markets of firms and joint development partners	–	Regular monitoring of technical areas or companies for new patent documents, visualization of patent activity of companies	The ability to determine the geography of the company, the best developers, and the intentions of competitors	Information about permanent inventors, which companies are applying, identifying gaps in areas of interest	–
Visualization of results	Various charts and graphs	Concept map	Bubble chart	Patent landscape	Topographic map of the landscape	Topographic map of the landscape
Free of charge	–	+	–	–	Trial period	–

**Table 2** Patent parsing functions

Function	Description
get_patent_name	Gets the name of the patent by extracting the text from the invention-title tag
get_country	Extracts the country that issued the patent from the country tag
get_number	Extracts the patent number from the doc-number tag
get_date	Extracts the patent grant date from the date tag
get_description	Extracts the patent text from the description tag
get_uniq_id	Combines information from the doc tags-number, country, king, gets a unique identifier for the patent

```

+-----+-----+-----+-----+-----+-----+
|      uniq|      patent_name|country|  number|   date| patent_description|
+-----+-----+-----+-----+-----+-----+
|US10548545B2|Anatomical imagin...|    US|10548545|20200204|An imaging system...|
|US10548547B2|X-ray computed to...|    US|10548547|20200204|The rigidity of a...|
|US10548537B2|Interface for a h...|    US|10548537|20200204|A testing system ...|
|US10548548B2|System and method...|    US|10548548|20200204|Systems and metho...|
|US10548539B2|X-ray CT apparatus...|    US|10548539|20200204|According to an X...|
|US10548544B2|Rotating-slit gam...|    US|10548544|20200204|A method for form...|
|US10548546B2|Medical examinati...|    US|10548546|20200204|A medical examina...|
|US10548533B2|Method and system...|    US|10548533|20200204|A method and syst...|
|US10548534B2|System and method...|    US|10548534|20200204|This application ...|
|US10548536B2|Device and method...|    US|10548536|20200204|A device and meth...|
+-----+-----+-----+-----+-----+-----+
    
```

**Fig. 1** Parsing patent function output format

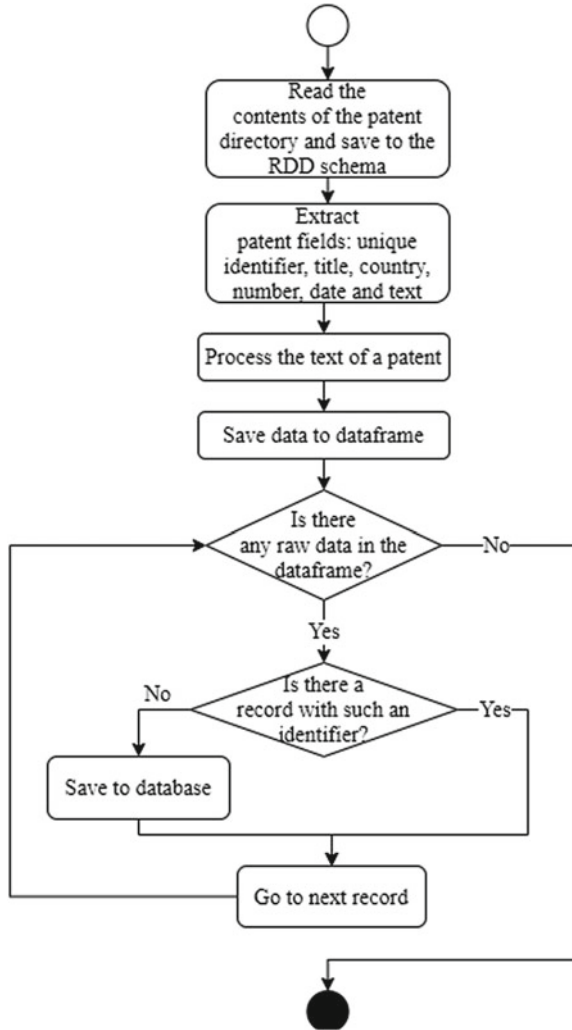
SAO (subject—“subject”, action—“action”, object—“object”) is a triple semantic construction that can be used to represent technical functions derived from the patent text. An object and subject is a word or phrase related to the semantics of a text. An action is a verb that is an operation by which an object and a subject are connected, or a relationship between them. Modern natural language processing technologies enable SAO structures to describe fairly complete semantic information about technical functions [15, 16].

The UDPipe library is used for parsing. The input and output data format for UDPipe is CoNLL [17]. Using the UDPipe library, a dependency tree is built, presented in the form of Table 3.

The fields containing the index, the word, the initial form of the word, the part of speech, the number of the parent vertex, and the relationship to the parent vertex are important for further work.

From the resulting trees, you need to remove vertices that do not carry a semantic load, such as: “aux”—auxiliary verb; “auxpass”—auxiliary verb in the

**Fig. 2** Algorithm for parsing a patent



**Table 3** UDPipe parsing table fields

Column name	Description
Id	Index of a word in a sentence
Form	Word
Lemma	An initial form of the word
CPosTag	Part of speech
Feats	Morphological features
Head	Index of the parent vertex
DepRel	Relationship to the parent vertex
Deps	Secondary dependencies

compassionate voice; “cc”—coordination; “det”—determinant; “predet”—determinant; “prep”—preposition modifier; “prt”—phrasal verb particle; “punct”—punctuation mark; “quantmod”—quantifier phrase modifier; “tmod”—temporary modifier. And the vertex with the relation “conj” gets the semantic role of its parent vertex.

The next stage of transformations is the deep-syntactic structures. We combine the dependencies obtained in the previous step into the following groups:

- actant relation I: “agent”—agent, “csubj”—clausal subject, “csubjpass”—clausal passive subject, “nsubj”—nominal subject, “nsubjpass”—nominal passive subject, “subj”—subject, “xsubj”—controlling subject;
- actant relation II: “dobj”—direct object, “iobj”—indirect object, “obj”—object, “pobj”—preposition object;
- attribute relation (attr): “amod”—adjective modifier, “cop”—copula, “nn”—compound noun modifier, “poss”—possession modifier;
- compositional relation (coord): “acomp”—adjective complement, “advmod”—adverb modifier, “ccomp”—clausal complement, “pcomp”—propositional complement.

Next, you need to extract the “Action” from the resulting transformed tree, it will be designated in the tree by the vertex “root”, and its child elements with the connection I and II [18].

As a result of this algorithm, the resulting SAO structure is stored in a data frame (Fig. 3).

To form the patent landscape, it is necessary to compare the SAO structures of each patent with each other and divide them into clusters.

The first step is the comparison by Action, that is, by the ROOT vertex. If the vertices of the patents do not match, the subsequent comparison of the trees of this offer is not performed, therefore, the similarity coefficient for Action is 0 and the comparison for the next offer of the patent begins. When the vertices match, the comparison of the attribute (ATTR) structures connected to the ROOT vertex begins.

The similarity coefficient of two SAO structures is calculated using the formula (1):

$$K = (K_{ATTR}^A + K^I + K^{II})/5, \quad (1)$$

where  $K_{ATTR}^A$  is the similarity coefficient of the attribute (ATTR) structures associated with the ROOT vertex (Action);

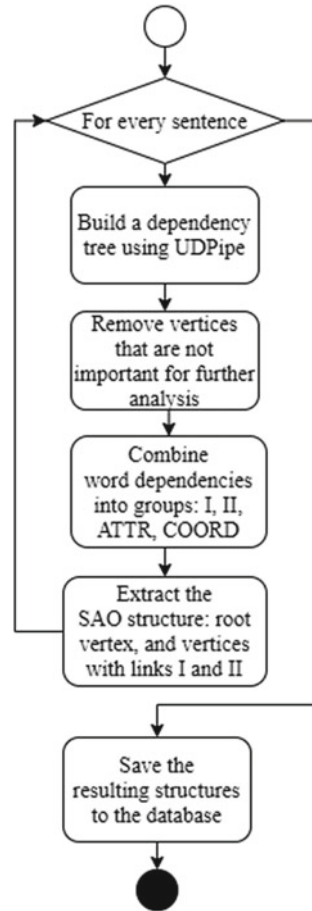
$K^I$ —similarity coefficient of subjects (I-relationship);

$K^{II}$ —similarity coefficient of objects (II-relationship).

The similarity coefficient of attribute structures (ATTR) associated with the ROOT vertex (Action) is calculated using the following formula (2):

$$K_{ATTR}^A(TA_k, TA_l) = \frac{\sum_{i=1}^{N_i} S(t_1, t_2)}{\max_{ATTR}(TA_k, TA_l)}, \quad (2)$$

**Fig. 3** Algorithm for extracting SAO structures



where  $TA_k, TA_l$  are semantic sub-trees (associated with the ROOT-vertex attribute structures) for the k-th and l-th sentences of two patents, respectively;

$\max_{ATTR}(TA_k, TA_l)$ —is the maximum number of ATTR structures for patents;

$S(t_1, t_2)$ —a function that determines whether the ATTR terms  $t_1$  and  $t_2$  coincide for the trees being compared;

$N_i$  is the number of terms for the semantic tree  $TA_k$ .

The second step is a comparison by Subject, that is, a comparison of the actant relations of I. If the vertices do not match, the similarity coefficient for this I-vertex of the first patent is equal to 0 and the comparison begins for the next I-vertex.

The third step is the comparison by Object, that is, the comparison of actant relations II. In case of a mismatch, the similarity coefficient for this II-vertex of the first patent is equal to 0 and the comparison for the next II-vertex begins.

After comparing the SAO structures, their frequency characteristics are calculated and added to the data frame, representing a term-document matrix “Patent-SAO”.



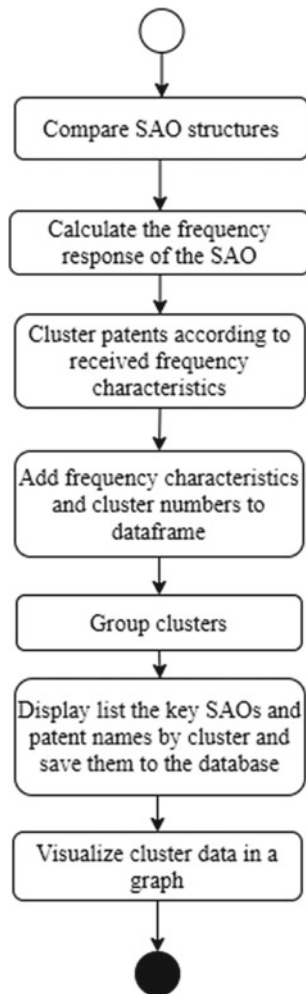
Based on the obtained frequency characteristics, patents are clustered. The “topic” function displays a list of key SAO structures and the names of patents that characterize each cluster, and this data is also stored in the database. Next, the clusters are grouped and the patents related to each cluster are visualized (Fig. 4).

Based on the cluster lists and patent names, the patent landscape for a given quarter and the number of clusters are displayed.

The module implemented according to the presented algorithms was tested for operability and the creation of an up-to-date patent landscape. The test results are shown in Figs. 5 and 6.

The patent landscape, built based on the obtained data, consists of three clusters and is shown in Fig. 7 [19].

**Fig. 4** Algorithm for the formation of the patent landscape



uniq	patent_name country	number	date	patent_description	id	features	prediction
US1054854582	Anatomical imagin...	US 10548545	20200204	An imaging system...	0 [3.72,0.0,0.0,2.0...		2
US1054854782	X-ray computed to...	US 10548547	20200204	The rigidity of a...	1 [14.88,0.0,0.0,0.0...		0
US1056213882	Method for manufa...	US 10562138	20200218	A method for manu...	2 [6.51000000000000...		0
US1056214982	Polyurethane CMP ...	US 10562149	20200218	A chemical-mechan...	3 [15.81,0.0,0.0,2.2...		0
US1054856682	System and method...	US 10548566	20200204	The system includ...	4 [2.79,0.0,0.0,2.0...		1
US1054855482	Radiation imaging...	US 10548554	20200204	Provided is a rad...	5 [0.93,0.0,0.0,0.0...		1
US1056976182	Vehicle drive app...	US 10569761	20200225	A vehicle drive a...	6 [15.81,0.0,0.0,2.2...		1
US1056212482	Welding apparatus...	US 10562124	20200218	A method providin...	7 [14.88,0.0,0.0,3.3...		1
US1055664082	Unmanned semi-sub...	US 10556640	20200211	An unmanned semi-...	8 [1.86,0.0,0.0,2.0...		2
US1056214682	Fiber based finis...	US 10562146	20200218	A finishing tool ...	9 [34.4100000000000...		0
US1056976682	Vehicle control d...	US 10569766	20200225	Disclosed herein ...	10 [0.93,0.0,0.0,0.0...		1
US1056213982	Fixtures for pipe...	US 10562139	20200218	These fixtures ai...	11 [0.37000000000000...		0
US1055641982	System and method...	US 10556419	20200211	A method for oper...	12 [0.93,0.0,0.0,0.0...		1
US1056974682	Sensor cleaning s...	US 10569746	20200225	A system includes...	13 [7.44,0.05,0.0,2.2...		1
US1056213682	Method of forming...	US 10562136	20200218	A plate is contin...	14 [13.9500000000000...		0
US1055642382	Liquid ejecting a...	US 10556423	20200211	A liquid ejecting...	15 [2.79,0.0,0.0,0.0...		1
US1056974982	Brake control dev...	US 10569749	20200225	A brake control d...	16 [2.79,0.0,0.0,0.0...		1
US1056211982	Machining system ...	US 10562119	20200218	A machining syste...	17 [4.65,0.0,0.0,2.7...		1
US1055642682	Printing apparatus	US 10556426	20200211	A printing appara...	18 [0.0,0.0,0.0,0.0...		1
US1055665482	Masks for underwa...	US 10556654	20200211	A full face divin...	19 [0.93,0.0,0.0,0.0...		2

Fig. 5 Cluster numbers for each patent

```

Top SAO per cluster:
1 cluster SAO:
[SAO(subjects=("mechanism"), action="include", objects=("mechanism"))] [SAO(subjects=("941"), action="fil
1 cluster titles:
X-ray computed tomography apparatus, Method for manufacturing rack bar, Polyurethane CMP pads having a t
2 cluster SAO:
[SAO(subjects=("941"), action="file", objects=("Jul"))] [SAO(subjects=("mechanism"), action="include", ob
2 cluster titles:
System and method for tracking signal of wire in a blood vessel, Radiation imaging apparatus, radiation
3 cluster SAO:
[SAO(subjects=("Torus"), action="define", objects=("opening"))] [SAO(subjects=("mechanism"), action="incl
3 cluster titles:
Anatomical imaging system with centipede belt drive, Unmanned semi-submarine, Masks for underwater uses,
    
```

Fig. 6 Lists of key SAO structures and cluster patent names

## 4 Conclusion

The theoretical value of this work lies in the developed methodology for analyzing patents based on the “Subject-action-object” structure and the automated system created on its basis.

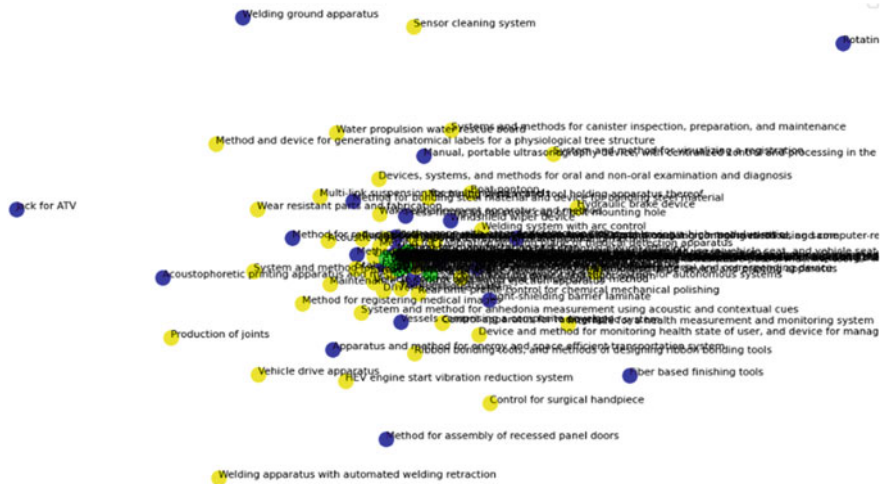


Fig. 7 The result of the module

**Acknowledgements** The reported study was funded by RSCF according to the research project 22-21-00855, RFBR and Administration of the Volgograd region according to the research projects 19-47-340007, 19-41-340016.

**References**

1. Korobkin, D.M., Fomenkov, S.A., Golovanchikov, A.B.: Method of identification of patent trends based on descriptions of technical functions. *J. Phys. Conf. Ser.* **1015**(3) (2018). <https://doi.org/10.1088/1742-6596/1015/3/032065>
2. Wang, X., Ren, H., Chen, Y., Liu, Y., Qiao, Y. and Huang, Y.: Measuring patent similarity with SAO semantic analysis. *Scientometrics.* **121**, 1–23 (2019). <https://doi.org/10.1007/s1192-019-03191-z>
3. Yufeng, D., Duo, J., Lixue, J.: Patent similarity measure based on SAO structure. *Chin. Sentence Clause Text Inf. Process.* **30**(1), 30–36 (2016)
4. Zhang, Y., Zhou, X., Porter, A.L., Gomila, J.M.V.: How to combine term clumping and technology roadmapping for newly emerging science & technology competitive intelligence: “problem & solution” pattern based semantic TRIZ tool and case study. *Scientometrics* **101**(2), 1375–1389 (2014)
5. Vasiliev, S.S., Korobkin, D.M., Kravets, A.G., Fomenkov, S.A., Kolesnikov, S.G.: Extraction of cyber-physical systems inventions’ structural elements of russian-language patents. *Stud. Syst. Springer, Decis. Control* **259**, 55–68 (2020). [https://doi.org/10.1007/978-3-030-32579-4\\_5](https://doi.org/10.1007/978-3-030-32579-4_5)
6. Park, H., Ree, J.J., Kim, K.: An SAO-based approach to patent evaluation using TRIZ evolution trends. In: 2012 IEEE 6th International Conference on Management of Innovation and Technology, ICMIT 2012 (2012). <https://doi.org/10.1109/ICMIT.2012.6225873>

7. Singh, P., Singh, S., Mishra, P.K., Garg, R.: RDD-Eclat: approaches to parallelize eclat algorithm on spark RDD framework. In: Smys S., Senjyu T., Lafata P. (eds.) Second International Conference on Computer Networks and Communication Technologies. ICCNCT 2019. Lecture Notes on Data Engineering and Communications Technologies, vol. 44. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-37051-0\\_85](https://doi.org/10.1007/978-3-030-37051-0_85)
8. Searching systems and databases. <https://www1.fips.ru/about/vptb-otdelenie-vserossiyskaya-patentno-tehnicheskaya-biblioteka/poiskovye-sistemy-i-bazy-dannykh.php>. Accessed 25 May 2021
9. Orbit Intelligence: Powerful patent searching and analysis. <https://www.questel.com/business-intelligence-software/orbit-intelligence>. Accessed 25 May 2021
10. PATENTSCOPE. <https://www.wipo.int/patentscope/en>. Accessed 25 May 2021
11. Building of patent landscapes (maps). <https://patentfamily.group/services/research/patentnyj-landshaft>. Accessed 25 May 2021
12. PatBase analyticsV2: Powerful, in-depth analysis in seconds. <https://minesoft.com/our-products/patbase-analytics-v2/>. Accessed 25 May 2021
13. STN AnaVist—Analysis and visualization software for information professionals. <https://www.cas.org/support/training/stnavivist>. Accessed 25 May 2021
14. Kim, C., Lee, H.: A patent-based approach for the identification of technology-based service opportunities. *Comput. Indus. Eng.* **144**(2020), 106464. <https://doi.org/10.1016/j.cie.2020.106464>
15. Kochura, O.A., Korobkin, D.M., Fomenkov, S.A., Kolesnikov, S.G.: Development of the patent array analysis module based on the “Problem-Solution” model. *J. Phys. Conf. Ser.* **1801**(1) (2021). <https://doi.org/10.1088/1742-6596/1801/1/012014>
16. Korobkin, D., Fomenkov, S., Zlobin, A., Shabanov, D., Golovanchikov, A.: The software for formation of technical function assessments based on the patent analysis. *Stud. Syst. Decis. Control Springer* **350**, 137–147 (2021). [https://doi.org/10.1007/978-3-030-67892-0\\_12](https://doi.org/10.1007/978-3-030-67892-0_12)
17. UDPipe [Electronic resource]. Institute of Formal and Applied Linguistics. Mode of access. <https://ufal.mff.cuni.cz/udpipe>. Accessed 10 March 2021
18. Mel'uk, I.: *Dependency Syntax Theory and Practice*. SUNY, New York (1988)
19. Korobkin, D.M., Fomenkov, S.A., Kravets, A.G.: Extraction of physical effects practical applications from patent database. In: 2017 8th International Conference on Information, Intelligence, Systems and Applications, IISA 2017 (2018)
20. No, H.J., Lim, H.: Exploration of nanobiotechnologies using patent data. *J. Intellect. Prop.* **4**(3), 109–129 (2009)

# Machine Learning of Diagnostic Neural Network for Railway Track Monitoring



Sergey Orlov , Antonina Piletskaya, Nadezhda Kusakina,  
and Andrey Tyugashev 

**Abstract** The chapter is devoted to the problem of automatic detection of defects during rail fastener inspection. The machine learning technique is proposed for the convolutional neural network as part of the diagnostic system of the rail track. The problem of the massive data analysis obtained in the form of images of rail fastenings with continuous visual monitoring of the track using four video cameras is described. It is shown that the use of a deep neural network can improve the efficiency of detecting defects in rail track elements. There is also the problem of the training sample representativeness in the defect classification. It is associated with a relatively small number of detected defect images. It is proposed to balance the training sample and the test sample using the SMOTE image synthesis methodology and to apply affine transformations to existing rail fastener images. The quality assessment of the datasets thus obtained was carried out using dimensionality reduction methods. The results of the experimental studies are presented.

**Keywords** Technical diagnostics · Railway track · Rail fastener inspection · Artificial neural network · Deep learning · Dataset balancing

## 1 Background

Rail fastening systems are an essential part of the high-speed railway. Their condition is critical to the operational safety and comfort of rail transport. Monitoring the rail fastening system conditions, identifying the looseness or absence of fasteners is the essential problem in the maintenance of the rail track [1].

---

S. Orlov (✉) · A. Piletskaya · N. Kusakina · A. Tyugashev  
Samara State Technical University, 244 Molodogvardeyskaya str, Samara 443100, Russia

N. Kusakina  
e-mail: [nadyakusakina@yandex.ru](mailto:nadyakusakina@yandex.ru)

A. Tyugashev  
e-mail: [tau797@mail.ru](mailto:tau797@mail.ru)

There is a problem in analyzing extensive data on the rail fastener technical states since there is a high density of fasteners on the rail track—up to 8 thousand units per kilometer. Monitoring railway sections 100 km long using video observation leads to the need to analyze about one million rail fastener images. Manual processing is possible only in the autonomous mode, the quality of the operator is deficient, and the number of defects missed is unacceptably large. Therefore, automated monitoring systems are being actively developed, which provide processing at the railway detector car's speed up to 120 km/h [2, 3]. Nevertheless, these results are already insufficient, as in developed countries, including Russia; there is an intensive development of a high-speed railway network.

For high-speed trains, the task was to develop autonomous diagnostic systems that are not installed on a specialized detector car but are embedded into a regular car of a regular high-speed train. Such systems must operate at speeds of up to 350 km/h and carry out continuous rail fastener inspection.

A high-speed train and railway track is a complex physical system in which a number of physical processes of various nature take place. They determine not only quality and comfort, but also the safety of transportation. For effective traffic control, it is necessary to consider a cyber-physical system in which a digital model of the interaction of all components of the “train-rail track” complex is implemented [3]. The complexity of this model makes it difficult to use the known mathematical models, which usually only describe the individual processes. To overcome these difficulties, methods of artificial intelligence are advisable to use [4, 5].

Nowadays, one of the most productive areas is the use of intelligent technologies for monitoring rail tracks. There are several methods and algorithms for big data processing in real-time, which allow obtaining information about the technical states of rail fasteners and providing a new diagnostic quality. The known solutions are based on the following methods:

- spectral analysis of vibrations of rail track elements [6–8];
- wavelet analysis [9, 10];
- the use of computer vision and artificial neural networks [11, 12].

The first two methods are commonly used in practice, but are quite complex and require significant processing time. At the same time, difficulties arise in tuning spectral and wavelet algorithms to new data with a wide variety of defective elements. At present, most research is focused on the development of diagnostic systems on artificial neural networks. The authors of this report developed and investigated several solutions based on artificial neural networks for the diagnosis of technical objects, including for monitoring the rail track [13–15].

This chapter is devoted to the development and study of the machine learning methodology of the neural network and the formation of training and test datasets. For this purpose, datasets are balanced based on the SMOTE method [16] and the use of affine transformations for rail fastening system images. The proposed methodology is based on the approach consisting in considering the diagnostic system and the monitored object as a cyber-physical system. For this, a “digital twin” of the elements of the rail track upper structure is created in the form of a fastener model image set.

## 2 Rail Track Monitoring System

Russian railways use mainly six types of rail fasteners. Among them, there are both modern types (KB, KBOP105, JBR, ARS), and quite a long time used fasteners with spikes (DO2, DO3). Examples of images of workable fasteners are presented in Fig. 1. Furthermore, Fig. 2 shows examples of fasteners with defects.

In the article [17], we proposed a diagnostic complex, which includes four video cameras for continuous recording of the visual image of both rails, an image preprocessing unit Automated WorkStation (AWS), and a convolutional neural network (Fig. 3). The decision support system DSS throughout the rail track quickly determines the technical states of the rail fasteners. It generates the necessary messages to

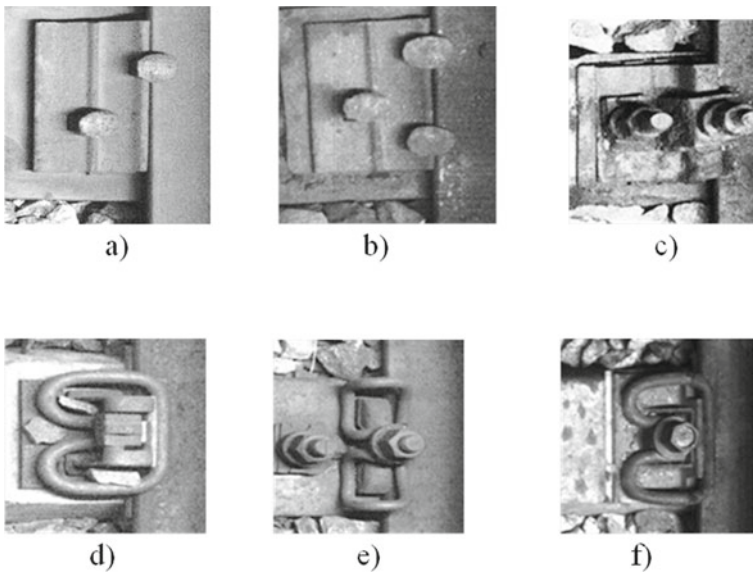


Fig. 1 Types of rail fasteners in workable states: a DO2, b DO3, c KB, d KBOP105, e JBR, f ARS

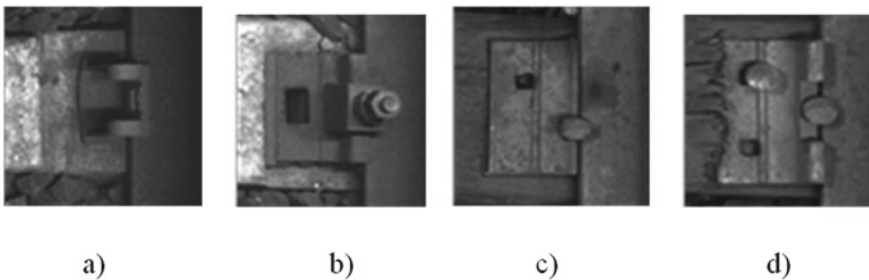
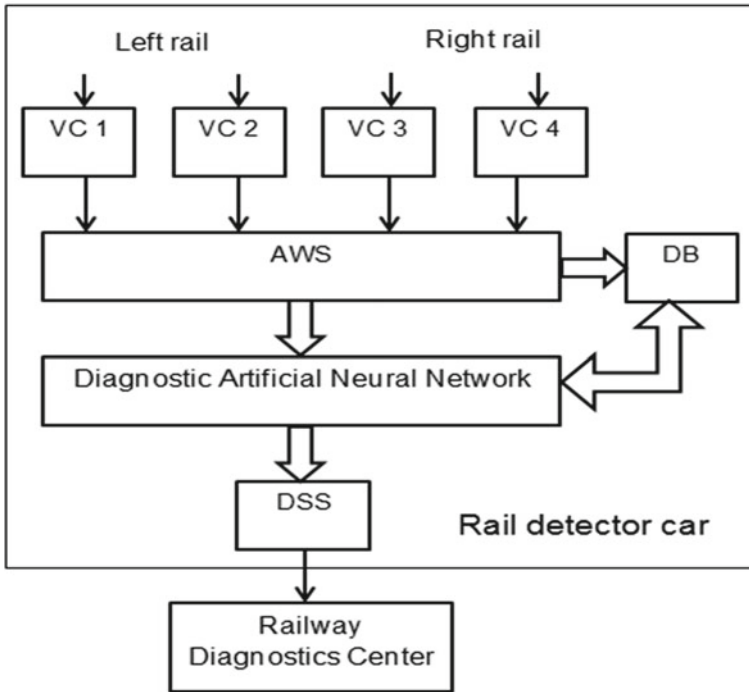


Fig. 2 Examples of rail fasteners: a ARS without railway clip, b KB no rail bolt, c DO2 no a spike, d DO3 no a spike



**Fig. 3** The general structure of the diagnostic system for rail detector car

Railway Diagnostic Center for the maintenance-of-way gang. The essential unit of the system is a deep neural network, which processes the frames of the video stream. The purpose of this neural network is to classify technical conditions and find rail fasteners with defects.

Thus, the diagnostic system performs real-time monitoring during the passing of trains on the route.

Machine learning of the diagnostic complex includes the following procedures.

1. Detection of the position on the video frame of the controlled elements of the rail track: fasteners, fish-plates, and connectors.
2. Formation of the initial image set of all items with coordinates on the observation section (50–100 km).
3. Expert assessment of the first set with the feature assignment by images of rail track elements: (a) “Fastener Type,” (b) “Defect Type,” or “Workable.”
4. Formation of the primary training dataset for the convolutional neural network.
5. Balancing the training dataset and test dataset using various methods of synthesizing samples.
6. Training a diagnostic neural network.
7. Testing a diagnostic neural network.



8. Operation of the diagnostic neural network and the collection of new images to complement the training dataset.
9. Re-training of a neural network on the updated data.

### 3 Convolutional Neural Network

This chapter discusses a neural network that classifies rail fasteners into nine classes. Six classes correspond to serviceable fasteners of the following types: DO2; DO3; KB; KBOP105; JBR; ARS. Two fastener classes are with the most common defects: KB without a bolt and DO3 without a spike. The ninth class “UR” is the set of all unrecognized and unclassified images. Therefore, the neural network has nine outputs, on which signals are interpreted as the probability of the analyzed image belonging to a specific class.

The developed diagnostic complex employs a deep convolution network with many layers [18], which is well suited for detecting the characteristic features of rail fastenings and overlays. Based on the analysis of known deep networks, the VGG-16 architecture [19] was chosen, which provided the required accuracy with a relatively simple architecture.

### 4 Balancing Training Dataset

The main problem of machine learning is caused by the relatively small number of defective fasteners (from 10 to 50) that are detected when a flaw detector car travels along rather than long sections of the railway. At the same time, there are a large number of images with serviceable fasteners. There is a problem with imbalanced classification. As a result, training a neural network leads to high accuracy (about 98%) of determining the type of a serviceable element. However, at the same time, the accuracy of defect detection is not more than 60–70%.

We propose balancing datasets using the SMOTE (Synthetic Minority Over-Sampling Technique) technique [16]. It is a technique that allows supplementing the training dataset with new synthesized samples based on existing samples. For this purpose, the “k-nearest neighbors” algorithm is used. SMOTE algorithm generates synthetic examples by working in a “feature space” rather than a “data space.” SMOTE works by selecting patterns that are close in the feature space, drawing a line between the examples in the feature space, and drawing a new sample at a point along that line.

As a result, each newly created sample contains attributes (in our task, these are pixels of an image raster), the values of which are calculated by interpolating the values of the corresponding attribute from randomly selected nearest neighbors. Figure 4 presents examples of software-synthesized fastener images with maximum likelihood.

**Fig. 4** Synthesized images of rail fasteners

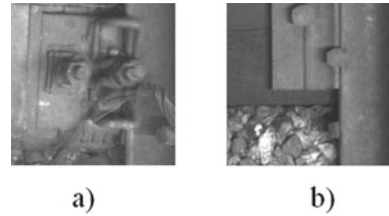


Figure 4a shows the artificial addition of a cable in the image of the KBOP105 fastener and Fig. 4b shows the addition of external elements of the embankment around the fastener.

After balancing the dataset using SMOTE, produced for the new dataset is performed one more addition. The creation of new instances from existing ones is carried out using their affine transformations, such as rotation and shift.

The general equation of the affine transformation in the form of a transition matrix of homogeneous coordinates has the way:

$$\begin{pmatrix} x' \\ 1 \end{pmatrix} = \begin{pmatrix} A & c \\ 0' & 1 \end{pmatrix} \begin{pmatrix} x \\ 1 \end{pmatrix}, 0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where  $\mathbf{A}$ —nonsingular matrix and  $\det(\mathbf{A}) \neq 0$ .

We used a rotation of each image by a random angle value (in the range from  $1^\circ$  to  $10^\circ$ ). An offset is also performed by an arbitrary value (within 10% of the image size). Such an addition to the training set was made in real-time using a computer with a GPU on an NVIDIA graphics card.

The initial dataset of rail fastening systems after the passage of the detector car and the work of the expert group are given in Table 1.

The total number of source images (samples) is 50,365. This dataset was balanced using the SMOTE technique, resulting in 161,640 samples. The obtained new dataset was divided into a training dataset and test dataset in a ratio of 80–20%. The number of samples in the training and test datasets was 129,312 and 32,328, respectively. The ninth class UR—“Unrecognized” in the amount of 20,205 samples (including 16,164 samples of the training sample and 4,041 samples of the test sample) was added to 8 classes. The dataset for nine classes in the amount of 145,476 samples was expanded with the use of affine transformations. Such an approach provided a total training dataset of 203,666 samples.

The results of balancing the dataset are presented in Fig. 5. The convolution neural network was trained on this dataset. Then, verification of the neural network on the test dataset was performed.

**Table 1** Initial dataset of rail fasteners

Fastener class	DO2	DO3	KB	KBOP105	JBR	ARS	DO3 defect	KB defect	UR
Image quantity	4114	5055	20,200	1041	3358	15,228	563	801	20,205

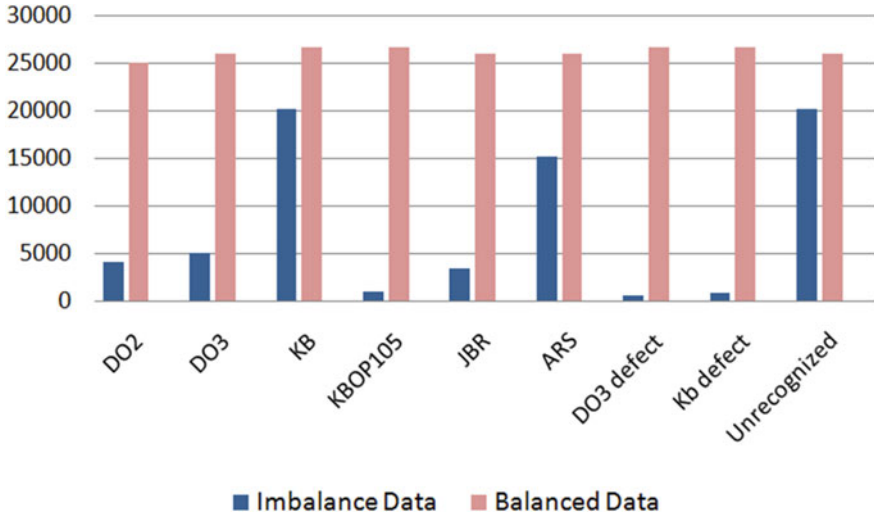


Fig. 5 Dataset balancing for machine learning diagnostic neural network

## 5 Training Dataset Quality Analysis

The fastener image is located at a certain point in the multidimensional feature space. The brightness intensities in each pixel on the corresponding image raster determine the position of this point. The number of feature space dimensions is equal to the number of fastener image pixels. In the experiments, the diagnostic system processed images of two sizes:  $100 \times 100$  pixels and  $250 \times 250$  pixels. Accordingly, the maximum dimension of the feature space is 62500.

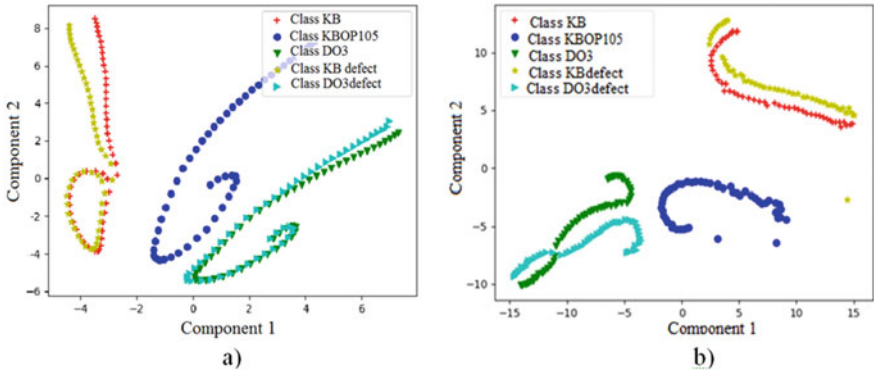
The first dimension of the feature space for analysis of the dataset must be reduced to a conveniently perceived form. For this, two methods of reducing the dimension were used: Multidimensional scaling (MDS) [20], t-Distributed Stochastic Neighbor Embedding (tSNE) [21].

In general, the problem of reducing the dimension can be written as follows:

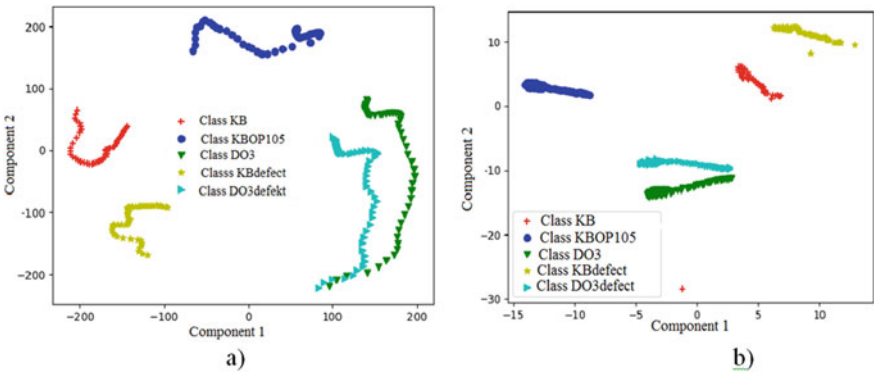
$$X = (x^{(1)}, \dots, x^{(p)}) \Rightarrow Z(X) = (z^{(1)}, \dots, z^{(r)}), r \ll p,$$

where  $(x^{(1)}, \dots, x^{(p)})$  are the coordinates that determine the position of the sample in the original multidimensional space with dimension  $p$ ,  $(z^{(1)}, \dots, z^{(r)})$  are the coordinates that determine the position of the sample in the space of reduced dimension  $r$ .

Area Figs. 6 and 7 show examples of the PCA and MDS methods application in the analysis of the training dataset for five classes of fastening systems: DO3, KB, KBOP105, DO3 without a spike, and defect KB. Figure 6 presents the results of



**Fig. 6** Diagrams of the results of the feature dimension reducing before balancing the dataset: **a** MDS method; **b** t-SNE method



**Fig. 7** Diagrams of the results of the feature dimension reducing after balancing the dataset: **a** MDS method; **b** t-SNE method

reducing the dimension to the training dataset before the balancing procedure. After balancing the dataset for these five classes, separation diagrams are obtained, which are shown in Fig. 7. For clarity, the diagram shows only two principal components.

## 6 Diagnostic System Tests and Neural Network Accuracy Assessment

Laboratory tests of an artificial neural network were carried out under stationary conditions on a computer complex with an Intel Core i7 7820X processor and an NVIDIA GeForce RTX 2080Ti graphics card with NVIDIA Turing architecture. The

Actual	Predicted									
	DO2	DO3	KB	KBOP105	JBR	ARS	UR	DO3 d	KB d	
DO2	0,9741	0,0123	0,0000	0,0000	0,0003	0,0000	0,0059	0,0071	0,0003	
DO3	0,0370	0,9510	0,0000	0,0020	0,0000	0,0027	0,0052	0,0020	0,0001	
KB	0,0000	0,0000	0,9643	0,0015	0,0030	0,0010	0,0122	0,0000	0,0180	
KBOP105	0,0016	0,0010	0,0058	0,9684	0,0073	0,0064	0,0058	0,0001	0,0036	
JBR	0,0005	0,0000	0,0000	0,0010	0,9400	0,0085	0,0500	0,0000	0,0000	
ARS	0,0000	0,0000	0,0032	0,0045	0,0060	0,9740	0,0110	0,0000	0,0013	
UR	0,0130	0,0048	0,0050	0,0049	0,0016	0,0065	0,9444	0,0124	0,0074	
DO3 d	0,0312	0,0022	0,0000	0,0005	0,0020	0,0001	0,0038	0,9634	0,0006	
Kb d	0,0007	0,0000	0,0190	0,0017	0,0032	0,0000	0,0000	0,0000	0,9771	

Fig. 8 Confusion matrix  $M_{CM}$  based on the passing of detector car

classification accuracy of serviceable fasteners and defects reached 97.4%. Field tests were carried out in a detector car on a section of the Kuibyshev railway with a length of 85 km.

Confusion matrix  $M_{CM}$ , obtained after processing the results of the classification using a neural network during the passing of a detector car, is shown in Fig. 8. The rows of the  $M_{CM}$  matrix correspond to the actual classes of fastener states, and the columns correspond to the states predicted by the neural network. Thus, the diagonal values are characterizing correct classification results, and the elements in each row compare to misclassification for the type of fasteners. The sum of the  $M_{CM}$  matrix elements in any row is 1. The area of predictions, shaded in blue, refers to errors of the second kind, i.e., “False Negative.”

## 7 Conclusion

It should be noted that, according to the requirements of traffic safety on the railway, defect skipping is very critical. Thus, incorrect classification of a useful fastener (“False Positives” or an error of the first kind) does not pose a danger of an emergency. In the worst case, it requires additional analysis, possibly with the assistance of the operator to refine results. Whereas classifying a defect as a good bond or assigning a given image with a defect to unrecognized states (“Skipping defect/False Negatives” or an error of the second kind) is an undesirable result.

Further development of the diagnostic system is supposed to be carried out in the direction of improving the architecture of the deep convolution network (for example, the use of VGG-19 or ResNet networks). The second direction is the use of affine transformations of compression and expansion of rail fastenings images for the formation of more representative datasets. In this case, the neural network will

be able to classify the railway track elements even if they are distorted during the video recording.

## References

1. The concept of development of systems for diagnostics and monitoring of objects of railway track facilities for the period until 2025. <https://rulaws.ru/acts/Rasporyazhenie-OAO-RZHD-ot-27.04.2016-N-777r/>. Accessed 29 Nov 2019
2. Feng, H., Jiang, Z., Xie, F., Yang, P., Shi, J., Chen, L.: Automatic fastener classification and defect detection in vision-based railway inspection systems. *IEEE Trans. Instrum. Measur.* **63**(4), 877–888 (2014)
3. Lee, E.A., Seshia, S.A.: *Introduction to Embedded Systems, A Cyber-Physical Systems Approach*, 2nd edn. The MIT Press, USA (2017)
4. Faghieh-Roohi, S., Hajizadeh, S., Nunez, A., Babuska, R., De Schutter, B.: Deep convolutional neural networks for detection of rail surface defects. In: *Proceedings of the 2016 International Joint Conference on Neural Networks (IJCNN)*, pp. 2584–2589 (2016)
5. Norvig, P., Russell, S.: *Artificial Intelligence: A Modern Approach*, 3rd edn. Pearson (2010)
6. Valsan, V.N., Patil, C.Y., Patekari, J.M.: Non-contact rail track parameter measurement. In: *2015 International Conference on Industrial Instrumentation and Control (ICIC)*. IEEE Xplore (2015)
7. Vágner, J., Zelenka, J., Hába, A., Kohout, M., Havlíek, P.: Stationary device for vibrodiagnostics of passing vehicle. *Vibroeng. Proc.* **6**, 98–103 (2015)
8. Yi, H., Kai, L.: *Inspection and Monitoring Technologies of Transmission Lines with Remote Sensing*. Academic Press (2017)
9. Wei, J., Liu, C., Ren, T., Liu, H., Zhou, W.: Online condition monitoring of a rail fastening system on high-speed railways based on wavelet packet analysis. *Sensors* **17**(2), 318 (2017)
10. Yue, G., Xu, Z., Wang, L.: Vibration analysis for slab track at different train speeds using Bayes wavelet denoising. In: *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, **231**(8), 892–901 (2016)
11. Karakose, M., Yaman, O., Akin, E., Baygin, M., Murat, K.: A new computer vision based method for rail track detection and fault diagnosis in railways. *Int. J. Mech. Eng. Robot. Res.* **6**(1), 22–27 (2017)
12. Gibert, X., Patel, V.M., Chellappa, R.: Deep multi-task learning for railway track inspection. *IEEE Trans. Intell. Transp. Syst.* **18**(1), 153–164 (2017)
13. Orlov, S.P., Vasilchenko, A.N.: Intelligent measuring system for testing and failure analysis of electronic devices. In: *Proceedings of the 2016 XIX IEEE International Conference on Soft Computing and Measurements*, vol.1, pp. 401–403. IEEE Xplore (2016)
14. Orlov, S.P., Girin, R.V., Uytova, O.Y.: Artificial neural network for technical diagnostics of control systems by thermography. In: *Proceedings of the 2018 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*, pp. 1–4. IEEE Xplore (2019)
15. Orlov, S., Girin, R.: Intelligent technologies in the diagnostics using object’s visual images. In: *Cyber-Physical Systems: Advances in Design & Modelling. Studies in Systems, Decision and Control*, vol. 259, pp. 301–312. Springer (2020)
16. Chawla, N.V., Bower, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE. Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
17. Orlov, S., Girin, R., Piletskaya, A.: Intelligent information processing system for monitoring rail tracks. In: *Proceedings of 3rd International Conference on Control in Technical Systems (CTS)*. IEEE Xplore (2020)
18. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: *NIPS’12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, vol. 1, pp. 1097–1105 (2012)

19. VGG. <https://www.robots.ox.ac.uk/~vgg/publications/2015/Simonyan15/>. Accessed 12 May 2019
20. Borg, I., Groenen, P.: Modern Multidimensional Scaling: Theory and Applications, 2nd edn. Springer, New York, NY (2005)
21. Van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008)

# Neural Network Analysis of Electromagnetic Field Effect on the Thermofluctuation Characteristics Measurement in the Power Cable Insulating Materials



N. K. Poluyanovich, M. N. Dubyago , N. V. Azarov, and A. V. Ogrenichev

**Abstract** The work is devoted to real-time throughput assessment and temperature forecasting of the power cable line (PCL) core based on the temperature monitoring system data, considering the influence on the PCL of factors caused by the magnetic component of the electromagnetic field. The problem of electromagnetic compatibility (EMC) is associated with ensuring the proper operation of a set of primary (power) and secondary circuits (electronic devices). Electromagnetic field influence on the measurement of thermal processes in PCL insulating materials is considered. The experimental study results of cable core temperature forecasting are presented with the help of developed INS, considering the influence of electromagnetic radiation (noise) at different levels. The separation of useful temperature signals among various types of noise is shown by the signal selection method using wavelet transform. Comparisons of forecast values with actual values (about 1%) suggest the selected network model adequacy and its applicability in practice for the reliable operation of the cable power supply system for consumers. It is shown that the effect of the implementation of the proposed principles of organizing monitoring systems contributes to the transition of the domestic industry to advanced digital, intelligent production technologies, robotic systems, new materials, and design methods, which corresponds to the direction of the Russian Federation state scientific and technical policy in the field of instrumentation and mechanical engineering.

---

N. K. Poluyanovich (✉) · M. N. Dubyago · N. V. Azarov · A. V. Ogrenichev  
Southern Federal University, Taganrog, Russian Federation  
e-mail: [nik1-58@mail.ru](mailto:nik1-58@mail.ru)

M. N. Dubyago  
e-mail: [w\\_m88@mail.ru](mailto:w_m88@mail.ru)

N. V. Azarov  
e-mail: [s.t.a.l.k.e.r999@yandex.ru](mailto:s.t.a.l.k.e.r999@yandex.ru)

A. V. Ogrenichev  
e-mail: [ogrenicheval@mail.ru](mailto:ogrenicheval@mail.ru)



**Keywords** Neural networks · Magnetic interference · Insulation materials · Wavelet analysis · Cable systems · Electromagnetic radiation · Temperature field analysis

## 1 Introduction

A specific cable insulation feature is aging, i.e. deterioration of electrical characteristics during operation. Thermoplastic polyethylene cable insulation has significant drawbacks—creep and a sharp deterioration in mechanical properties (up to a loss of dimensional stability) at temperatures close to the melting temperature;—internal stresses, “frozen in insulation” during its manufacture, manifest themselves at elevated operating temperatures, leading to noticeable shrinkage, and in some cases, to the insulation cracking;—it heats up due to dielectric losses. With an increase in the insulation temperature, chemical reactions occurring in the insulation materials are accelerated, amplified by internal inhomogeneities. This reduces the electrical properties of the insulation and leads to thermal breakdown of the insulation and failure of electrical equipment [1]. One solution to this problem is to monitor the cable temperature during operation.

The power supply reliability of consumers depends on monitoring the throughput and condition of the PCL by a non-destructive diagnostic method. Modern cable networks use cross-linked polyethylene (XLPE) as the main insulation material. The main functions for continuous monitoring of the PCL current thermal mode include the following functions:—fixing the excess of the permissible cable temperature in time;—prevention of current overloads of cable lines arising in the lines;—forecasting of the load permissible for the cable line if the power cable reaches its maximum operating temperature;—ensuring the continuity of the power supply process to consumers and reducing the likelihood of emergency events.

From the point of heat engineering view, PCL is not a complex system, however, environmental factors and electromagnetic processes complicate the task of analyzing thermal regimes by several orders of magnitude. This significantly complicates the work and reduces the reliability of non-destructive methods of diagnostics and monitoring of the state and operation of the PCL. The problem that prevents the widespread use of non-destructive methods for PCL insulation through diagnostics by registration and analysis of signals from sensors is their low noise immunity. Non-destructive methods [thermal fluctuation, partial discharge method, etc.] use highly sensitive sensors of current (voltage), zero sequences, temperature, etc. Thus, there are a lot of high-frequency pulses in high-voltage networks, in terms of parameters close to pulses from PD in PCL insulation. And there is a very high probability of registering signals “similar to pulses from PD”, directed to the sensor (mounted at the end of the cable line) from the outside, or neighboring PCLs. This significantly reduces the reliability of the diagnostic conclusions of the PCL monitoring system for the PD.

The growth of loads in the city cable distribution network, an increase in the density of cable lines lead to the need to assess their electromagnetic compatibility. Low-frequency electromagnetic fields (EMF) generated by high-voltage power lines can create significant interference that disrupts the normal functioning of electrical and electronic devices [2]. Cable lines made of shielded cables with molecular cross-linked polyethylene insulation and laid in the ground can create significant magnetic fields (MF), and in some cases exceed the norms permissible under the conditions of electromagnetic compatibility (EMC). Therefore, the problem of the effect of an electromagnetic field on monitoring temperature processes in solving the problems of the throughput of high-voltage PCL is urgent. The EMC task is associated with ensuring the proper operation of a set of primary (power) and secondary circuits (electronic devices) requires:—an assessment of the interference effect on the system;—calculating the degree of electromagnetic interference influence;—calculation of the EMF effect of various frequency ranges;—calculation of the MF induction effective value of single-core PCL to limit it to the maximum permissible level [3]. The EMC problem is also very relevant in connection with the widespread introduction of microprocessor technology in the secondary circuits. The study of electromagnetic processes in high-voltage power cable networks is an important task associated with maintaining the required level of insulation heating to ensure an uninterrupted power supply to consumers.

Due to these factors and to solve the problems of the increasing volume of energy consumption associated with the development and emergence of new energy-intensive industries, it is necessary to create specialized cyber-physical models (CPM), in which both the work of PCL throughput forecasting and the forecasting of thermal fluctuation processes in insulating materials are simulated [4]. Examples of CPS include monitoring, process control systems [5]. The key disadvantage of existing systems is the lack of effective forecasting of incidents and events that occur during the CPS infrastructure functioning [6, 7]. In addition, existing systems do not fully utilize the possibilities of self-learning based on the available data on the states of objects. In turn, this is due to the insufficiently effective operation of monitoring systems.

The architecture of the developed CPS is the implementation of the information technology concept for the integration of computing resources into the power grid. CPS is organized as a multi-level complex system, consisting of hardware, computational and physical elements, which continuously receives data on the operating parameters of the power cable network of consumers' power supply, and uses them to further optimize the control of the PCL throughput. It integrates computing, communication, and information storage capabilities with real-time PCL operation monitoring and cable system management to ensure efficient, reliable, and safe power supply to consumers. In it, the physical and software components are closely interconnected. A distinctive feature of a cyber-physical system is a significant amount of processed data ( $U_c$ ,  $I_c$ ,  $I_{tnp}$ ,  $T_{o.s.}$ ,  $T_{PCL}$ , etc.). This fact required the development of different approaches to organizing the architecture of all the key elements of the developed CPS. And so, CPS is a multi-level computing platform with

many interconnected elements such as:—physical (transformers, PCL, etc.);—hardware (current, voltage, temperature sensors, etc.);—an information system, which includes a database (search and data processing), information processing devices. The processed information goes to a computer that has access to the network either via a wifi network or via a GSN network. The information is recorded on a local disk and then sent to the data collection and processing center of the automated process control system of the energy supplying organization. This will allow monitoring and managing sections of the power grid, maintaining power quality indicators at a given level. At the same time, the control of CPS elements can be carried out both by the end-user and by intelligent algorithms.

The study aims are to increase the reliability and energy efficiency of power systems through the use of new non-destructive methods and algorithms for monitoring and diagnosing the PCL resource based on artificial neural networks for short-term and medium-term forecasting of temperature modes of the current-carrying core of power cable lines.

## 2 Review of Literary Sources

There are known works on assessing the cable electromagnetic field (EMF) influence according to the electromagnetic compatibility criteria [1], the optimal cable laying configuration from the position of reducing their electromagnetic field [8], work on reducing the electromagnetic field level by the development of the screen [8]. It seems expedient to assess the EMF influence considering: the CL load nature, the probable short-circuit currents (SC), and the adopted settings of relay protection and automation [9–13]. In [1], a joint solution of the electromagnetic field and heat conduction equations is given. Various screen grounding schemes are considered here, however, the absence of an attached electrical circuit forces us to use simplified grounding conditions and does not allow estimating losses while using the screen and armor at the same time. Further development of the finite element analysis method of the cable thermal state was carried out in [9], where a joint solution of the unsteady equations of the electromagnetic field and thermal conductivity was obtained for a short powerful current pulse (lightning strike).

## 3 Permissible Rated Cables Current Load

The electromagnetic compatibility issues of the electric power equipment operation play an important role in the CL and CS design. Moreover, the level of electromagnetic interference, the source of which is the current in the cable core, is the higher, the greater the throughput, i.e., the core larger the cross-section and, accordingly, the load

current of each phase. When high-voltage equipment operates in space, electromagnetic fields of various frequency ranges arise, which can cause an electromagnetic compatibility violation of the primary (power) and secondary circuits.

The equations of a quasi-stationary alternating electromagnetic field in the frequency domain with eddy currents are written concerning the complex vector magnetic potential  $A$ , which in the two-dimensional plane-parallel approximation has only one nonzero component  $A = Az$ :

$$\frac{\partial}{\partial x} \left( \frac{1}{\mu} \frac{\partial A}{\partial x} \right) + \frac{\partial}{\partial y} \left( \frac{1}{\mu} \frac{\partial A}{\partial y} \right) = -j_{ext} + i\omega\gamma A \quad (1)$$

where  $\mu$ —magnetic permeability, H/m;  $\gamma$ —specific conductivity, S/m;  $\omega$ —circular frequency, rad/s;  $j_{ext}$ —external current density, A/m<sup>2</sup>;  $A$ —vector magnetic potential, Wb/m.

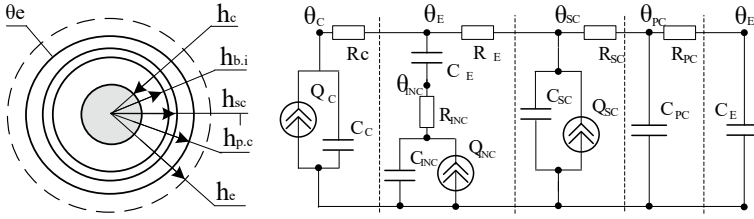
There is a difference in the process of CL insulation damage which is energized (ie, under the influence of the CL electric field), but without load and loaded CL (ie, under the influence of the CL electromagnetic field) [7, 14].

The problem of maximum current load estimating is the inverse problem of the thermal field calculating. The direct problem uses a known heat source; it is used to calculate the temperature distribution at all points of the model. The inverse problem identifies the heat source (cable current load) by the setpoint temperature at the model control point. The inverse problem is solved by enumerating solutions to several direct problems, possibly using interpolation and/or a search algorithm [15]. The results of calculations carried out by following per under IEC 60,287 [16] are taken as an initial approximation for the conductor's current effective value. The permissible current load of cables for alternating voltage can be obtained from the formula for the excess of the core temperature over the environment temperature [17].

$$\Delta\theta = \left( I^2 R + \frac{1}{2} W_d \right) \cdot T_1 + [I^2 R(1 + \lambda_1) + W_d] \cdot n T_2 + \\ + [I^2 R \cdot (1 + \lambda_1 + \lambda_2) + W_d] \cdot n(T_3 + T_4) \quad (2)$$

## 4 Development of the Thermal Model of the XLPE Cable Line

To form a mathematical model of a cable power transmission line (CPTL) with XLPE insulation of a single-core cable APvPu g-1 × 240/25-10, we represent each layer in the form of a thermal layer, Fig. 1a, where  $h_c - h_{p,c}$  is the outer cable-layers radii,  $\theta_{b,i}$ —the casing temperature, °C. The equivalent cable circuit is shown in general in Fig. 1b, where  $\theta_c - \theta_{sc}$ —the core heat release, inclusions, and screen, which are



**Fig. 1** a—cable structure; b—thermal equivalent circuit for numerical calculation of temperature

**Table 1** Cable parameters

Material parameters	Material				
	Aluminium	XLPE—insulation	Copper	Air	Soil
Specific heat conductivity (W/(m · K))	237	0,32	400	0,0259	1
Density (kg/m <sup>3</sup> )	2700	952	8700	1.2	2000
Specific heat capacity (J/(kg · K))	920	1900	385	1005	850
Electrical conductivity (S/m)	31,75·106	1·10 <sup>-9</sup>	5·107	—	—

analogs of current sources;  $\theta_e$ —the temperature of the air-ground interface, which is an analog of the EMF source.

The authors have developed a simulation model of the cable, which allows considering thermo-fluctuation processes taking into account dielectric losses in the insulation caused by PD.

The material’s physical properties and the geometric dimensions of the cable APvPu g—1 × 240/25–10 elements are summarized in Table 1.

The mathematical model of the XLPE-insulated cable line is formed for the cable cross-section shown in Fig. 2. The system of equations for the layer-by-layer determination of the cable line temperature regimes is represented by the system of Eq. (3) [18]. In these equations  $\theta(h)$ —temperatures of PCL section homogeneous layers;  $h_c, h_{b,i}, h_{sc}, h_{p,c}, h_e$ —distances from the cable homogeneous area center (core, basic insulation, screen, protective casing, and environment, respectively);  $\lambda_e, \lambda_{b,i}, \lambda_{sc}, \lambda_c$ —the specific thermal conductivity of the environment, polyethylene, screen, and core, respectively;  $I_c, I_{sc}$ —current of the core and in the screen, respectively;  $\varkappa$ —coefficient of thermal conductivity;  $\delta_c = I_c/S_c; \delta_{sc} = I_{sc}/S_{sc}; S_c, S_{sc}$ —cross-sectional areas of the core and screen;  $\gamma_c, \gamma_{sc}$ —electrical conductivity of the core and screen.

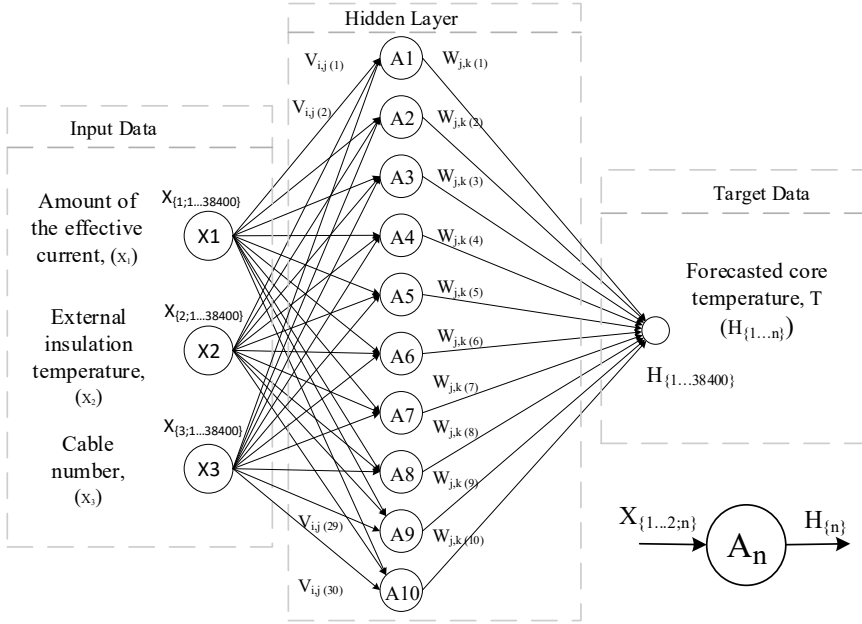


Fig. 2 The neural network structure used for the forecasting

$$\theta(h) = \begin{cases} \theta_e + \left( \frac{I_C^2}{\gamma_C S_C} + \frac{I_{SC}^2}{\gamma_{SC} S_{SC}} \right) / 2\pi h \chi, (h = h_e) \\ \theta(h_e) + \ln\left(\frac{h_e}{h}\right) \left( \frac{I_C^2}{\gamma_C S_C} + \frac{I_{SC}^2}{\gamma_{SC} S_{SC}} \right) / 2\pi \lambda_e, (h_{P.S.} \leq h \leq h_e) \\ \theta(h_{P.S.}) + \ln\left(\frac{h_{P.S.}}{h}\right) \left( \frac{I_C^2}{\gamma_C S_C} + \frac{I_{SC}^2}{\gamma_{SC} S_{SC}} \right) / 2\pi \lambda_{B.C.}, (h_{SC} \leq h \leq h_{P.S.}) \\ \theta(h_{SC}) + \ln\left(\frac{h_{SC}}{h}\right) \left( \frac{I_C^2}{\gamma_C S_C} - \frac{j_{SC}^2 h_{B.C.}^2 \pi}{\gamma_{SC}} \right) / 2\pi \lambda_{SC} + \\ + j_{SC}^2 (h_{SC}^2 - h^2) / 4\lambda_{SC} \gamma_{SC}, (h_{B.C.} \leq h \leq h_{SC}) \\ \theta(h_{B.C.}) + \ln\left(\frac{h_{B.I.}}{h}\right) \frac{I_C^2}{2\pi \lambda_{B.C.}} + \frac{j_{B.C.}^2 (h_{B.C.}^2 - h^2)}{4\lambda_{B.C.} \gamma_{B.C.}}, (m_1 + 2r_V \leq h \leq h_{B.C.}) \\ \theta(h_{B.}) + \frac{\left( \frac{I_C^2}{\gamma_C S_C} + \frac{I_{INC}^2}{\gamma_{INC} S_{INC}} \right) \ln\left(\frac{m_1}{h_C}\right)}{2\pi \lambda_{INC}} + \frac{I_{INC}^2 (h_C^2 - h^2)}{\gamma_{INC} S_{INC} \lambda_{INC}} + \\ + \frac{I_C^2 \ln(h_{B.C.}/m_1)}{2\pi \lambda_C \gamma_C S_C}, (m_1 \leq h \leq m_1 + h_{INC}) \\ \theta(h_{B.C.}) + \ln\left(\frac{h_{B.C.}}{h}\right) \frac{I_C^2}{2\pi \lambda_{B.C.}} + \frac{j_{B.C.}^2 (h_{B.C.}^2 - h^2)}{4\lambda_{B.C.} \gamma_{B.C.}}, (h_C \leq h \leq m_1) \\ \theta(h_C) + \frac{j_C^2 (h_C^2 - h^2)}{4\lambda_C \gamma_C}, (0 \leq h \leq h_C) \end{cases} \quad (3)$$

Known electrical load forecasting methods do not work well with noisy and incomplete data. The most widespread forecast method is based on specific electricity consumption rates with expert correction (correlation-regression analysis). Polynomial expansions can also be successfully used as a model for approximating the load graph and its forecast [19]. If it is necessary to forecast for a long-time-interval, then, in such cases, autoregressive methods can be used [20]. All these methods have significant drawbacks:

- their application requires the analyst’s direct participation;
- differ in the calculations complexity;
- are characterized by insufficient forecast accuracy;
- can only be applied to a specific type of forecast;
- differ in sensitivity to input data.

Due to the power grid instability when the operating mode parameters change, the throughput forecasting process of power supply networks and ensuring the operating parameters becomes a difficult task. Forecasting systems are built based on artificial intelligence using machine learning methods. The most promising is the method of artificial neural networks (ANN), and for their training can be used, among other things, open databases for monitoring energy systems and weather phenomena [16].

## 5 The Signal Noise Removal Procedure

To clean the signal from noise, the Daubechies wavelet was used, since it is orthogonal and can reconstruct the signal. The model of a noisy signal is usually assumed to be additive:  $s(n) = f(n) + k \cdot e(n)$  with a uniform step in the argument  $n$ , where  $f(n)$ —the useful information component,  $e(n)$ —the noise signal with an average zero value.

To construct wavelets, we will use the stretching equation and the wavelet equation:

$$\phi(t) = \sqrt{3} \sum_k h_k \phi(2t - k) \quad (4)$$

$$\psi(t) = \sqrt{3} \sum_k g_k \phi(2t - k) \quad (5)$$

The compactness of the functions supports  $\phi$  and  $\psi$  can be achieved if a finite number  $h_n \neq 0^\circ$  is chosen in such a way as to achieve the wavelet orthogonality and smoothness. For the Fourier domain, the orthogonality and smoothness condition is as follows:

$$|m_0(\omega)^2| + |m_0(\omega + \pi)|^2 = 1 \quad (6)$$

where  $|m_0(\omega)| = \sum_n \frac{h_n e^{-in\omega}}{\sqrt{2}}$ —exponential polynomial, subject to moments

$$\frac{d^l \psi \omega}{d\omega^l} \Big|_{\omega=0} = 0 \tag{7}$$

for  $l = 0, 1, \dots, N - 1$  assuming the form

$$m_0(\omega) \propto \left( \frac{1 + e^{i\omega}}{2} \right)^N \tag{8}$$

If we assume that  $M_0(\omega) = |m_0(\omega)|^2$ —is a  $\cos \omega$  polynomial, then the zero moment condition gives.

$$M_0(\omega) = \cos^{2N} \frac{\omega}{2} L(\omega), \text{ where } L(\omega) = P \sin^2 \frac{\omega}{2}$$

To find the coefficients  $h_n$ , it is necessary to obtain  $m_0$  by selecting the form of the polynomial P. From the orthogonality condition and the zero moment conditions, it follows that

$$P(y) = (1 - y)^{-N} (1 - y^N P(1 - y)) \tag{9}$$

where y polynomial function.

Expanding  $(1 - y)^{-N}$  to the  $N - 1$  order, we obtain an explicit form of the polynomial:

$$P(y) = (1 - y)^{-N} (1 - y^N P(1 - y)) = \sum_{k=0}^{N-1} \binom{N+k-1}{k} y^k \tag{10}$$

By spectral factorization, we can extract the roots  $m_0$  from P:

$$m_0(\omega) = const \left( \frac{z + 1}{2} \right)^N \prod_{j=1}^{N-1} (z - z_j) \tag{11}$$

The sought wavelet coefficients  $h_j/2$  will be the coefficients of  $z_j$  in the reverse order. Also, to construct this type of wavelet, a cascade algorithm is used. It allows constructing a scaling function  $\varphi$  pointwise from known coefficients  $h_n$ . At each algorithm step, the function  $\varphi$  is specified along the t axis. After that, knowing  $\varphi$  and  $h_n$ , the function of the wavelet  $\psi$  itself is found.



## 6 Description of the Architecture, Parameters and Choice of ANN

During solving the problem of forecasting thermal processes in PCL, considering the noise level of the received signals, the neural network technological cycle was determined. The ANN architecture (Fig. 2) for forecasting PCL thermal fluctuation processes, which is a multilayer perceptron, has been developed. The following were determined as the initial parameters:—cable core current ( $I_c$ );—the cable surface temperature ( $\theta_{p,c}$ );—environment temperature ( $\theta_e$ ).

The number of input and output ANN elements is determined, the number of ANN hidden layers is determined, the number of neurons in the hidden layers is calculated, Fig. 2. The neural network equations of this architecture can be represented as follows (12), where  $y_1$ —the output of the neural network;  $f_j()$ —activation function for the  $j$ -th neuron of the hidden layer;  $n$ —the number of neurons in the hidden layer;  $h$ —the number of entry nodes;  $v_{ij}$ —the weight connecting the  $i$ -th input node with the  $j$ -th hidden neuron;  $w_{jk}$ —the weight connecting the hidden neuron  $j$ -th node with the output layer  $k$ -th neuron;  $Q_j$ —displacement of the  $j$ -th hidden neuron;  $T_k$ —displacement of the output layer neuron.

$$\left\{ \begin{array}{l} g_j = f\left(\sum_{i=1}^n v_{ij}x_i + Q_j\right) \\ y_1 = f\left(\sum_{j=1}^h w_{jk}g_j + T_k\right) \\ w_{jk}(t + 1) = w_{jk}(t) + a \frac{\partial E}{\partial w_{jk}(t)} \\ v_{ij}(t + 1) = v_{ij}(t) + a \frac{\partial E}{\partial v_{ij}(t)} \end{array} \right. \quad (12)$$

The analysis of the various NN activation function’s influence on the forecast error of PCL thermal fluctuation processes is carried out. It has been established that the minimum error in forecasting thermal processes in power cable networks is the NN with the logsig activation function in the hidden layer and purelin in the output layer. Networks of this type are to some extent a universal structure for many problems of approximation and forecasting.

A comparative analysis of the training accuracy for various training algorithms is carried out, Table 2.

**Table 2** NN training

NN training algorithm	Average forecast error, at 80% of the sample	
	$\varepsilon$ , °C,	$\varepsilon$ , %,
Leavenberg-Marquardt	0.87	2.2
Bayesian regularization	1.6	4.21
Gradient descent	2.13	5.6

The research results have shown that of the considered NN training algorithms, the Lie-Venberg-Marquardt has the highest accuracy. The network is trained every time “from scratch” the weights obtained during the initial training of the network are not used for repeated training.

In the research course by measuring the thermal characteristics of the studied PCL samples insulating layers, an experimental database (BDNN (i, j)) was compiled for training the neural network, which includes indicators of serviceable and aged samples. TraiNN (i) is a training sample that includes data on serviceable experimental cable samples. TestNN (j) is a test set that includes data on aged experimental samples.

## 7 Experimental Research Results

As an analysis result of available software tools for designing and developing ANNs for solving problems of forecasting the PCL residual resource, the choice was made in favor of the Neural Network Toolbox software package as part of MATLAB. The noise-cleaning process was performed using the Wavelet Toolbox utility. The voltage of the temperature dependences, taken with the help of temperature sensors for several samples of the cable Apv Pu g–1 × 240/25–10, were investigated.

As an input signal for cleaning, we used the results of measuring the temperature dynamics on the power cable protective casing  $\theta_{p,c}$  and the environment  $\theta_e$  in a wide range of power cable core currents variation ranging from 450, A to 650, A, Fig. 2a. The wavelet transformation graphs of the noisy signal  $e(n)$  of the temperature dependences  $\theta_{p,c}$  on the PCL protective casing versus time for the cable samples under study are obtained. The noise thresholds for processing the Daubechies wavelet of cable samples 5, 7, 9 are presented in Table 3.

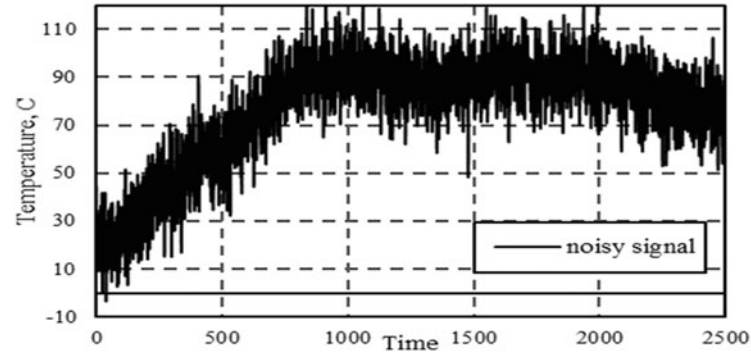
The graphs of the experimental dependence  $\theta_{p,c}(t)$  PCL protective casing temperature on time—noise signal  $e(n)$  Fig. 3a and cleared by noise  $f(n)$  Fig. 3b for one of the samples under study (cable No. 7), built using NN.

**Table 3** Noise thresholds for the cable samples processing

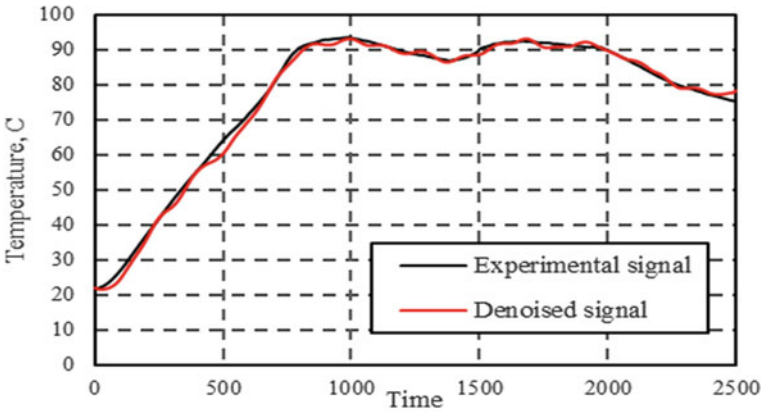
Decomposition rate N	Threshold detailing coefficients		
	Sample №5	Sample №7	Sample №9
6	30.736	23.387	23.661
5	27.316	25.034	21.006
4	26.052	30.930	31.581
3	27.045	30.841	24.900
2	34.610	35,832	29.913
1	39.349	31.952	34.181

**Table 4** The error between experimental and cleaned signal

Test sample №	Average temperature, $\theta_{p,c}, ^\circ\text{C}$		Error	
	Initial, $e(n)$	Cleared, $f(n)$	$t, ^\circ\text{C}, \text{ev}$	$t, ^\circ\text{C}, \%$
5	61,2	61,3	0,6	1,03
7	65.58	65.49	0,9	1,08
9	77.22	77.16	0,8	1,06



a



b

**Fig. 3** a—experimental  $e(n)$  and b—noise-free temperature dependence  $f(n) \theta_{p,c}(t)$

The results of the wavelet cleaning of signals from the cable core temperature sensors ( $\theta_c$ ) and temperature forecasting using the ANN for the cable sample, considering the influence of electromagnetic interference, Fig. 3, prove the need to develop a more perfect method for assessing the PCL throughput  $AP_v P_u g - \times 240/25 - 10$  in real-time, considering the EMF influence.

The error between the experimental temperature dependence  $\theta_{p,c}(t)$ —noisy signal  $e(n)$ —and the wavelet cleaned signal  $f(n)$  is presented in Table 4.

The received data deviation from the ANN training sample data, in the presence of magnetic interference on the cable (considering them as an input parameter of the neural network), increases the error by 1.3 degrees (or 0.8%). Cleaning the signals taken from the sensors from interference showed that, at different noise levels, the error of the reconstructed signal was about 1%. The conducted comparisons of the forecasted values with the actual ones allow us to speak about the selected network model adequacy and its applicability in practice for the reliable operation of the consumers' cable power supply system.

## 8 Conclusion

The neural network has been developed to determine the temperature regime of the power cable current-carrying conductor. The error evaluation in the forecasted and experimental cable samples temperatures showed that the average forecast error was 2.6 °C, which is quite an acceptable result. The model makes it possible to assess the current state of insulation and forecast the PCL residual resource. Studies have shown that much more current can be passed through the cable line without exceeding the permissible wire temperature. The results obtained can be applied in the expert system for forecasting the cable power supply systems' technical state.

The experimental study results of the signals wavelet purification from the cable core temperature sensors ( $\theta_c$ ), using the ANN for a cable sample, considering the electromagnetic interference influence, showed that, at different noise levels, the reconstructed signal error was approximately 1%. The forecasted values conducted comparisons with the actual ones, allow us to speak about the selected network model adequacy, and its applicability in practice for the reliable operation of the power supply cable system to consumers.

The system's introduction using temperature monitoring will allow practically online monitoring of each cable line state, including its real throughput, which means that much more current can be passed without exceeding the permissible cable temperature.

## References

1. Titkov, V.V.: To the assessment of the thermal regime of a three-phase line made of XLPE cable. *Cable-News* **10**, 47–51 (2009)
2. Sidorov, A.I., Okraïnskaya, I.S.: Electromagnetic fields near extra-high voltage electrical installations. Chelyabinsk: YUr-GU, 204 p (2008)
3. Rozov, V.Yu., Reutskiy, S.Yu., Piliugina, O.Yu.: The method of calculation of the magnetic field of three-phase power lines. *Tekhnichna Elektrodynamika*, no. 5, pp. 11–13 (Rus) (2014)

4. Khaitan, et al.: Design techniques and applications of cyber physical systems: a survey. *IEEE Syst. J.* (2014)
5. Lee, E.A., Seshia, S.A.: Introduction to Embedded Systems—A Cyber-Physical Systems Approach. LeeSeshia.org, Introduction to Embedded Systems—The Cyber-Physical Systems Approach (2011)
6. Terracciano, M., Purushothaman, S.: Thermal analysis of cables in unfilled troughs: investigation of the IEC standard and a methodical approach for cable rating. *IEEE Trans. Power Delivery* 27(3), 1423–1431 (2012)
7. Anders, G.J., Napieralski, A., Orlikowski, M., Zubert, M.: Advanced modeling techniques for dynamic feeder rating systems. *IEEE Trans. Ind. Appl.* 39(3), 619–626 (2003)
8. Ippolito, M.G., Puccio, A., Ala, G., Ganci, S.: Attenuation of low frequency magnetic fields produced by HV underground power cables. In: *IEEE Conference Publications in 50th International Universities Power Engineering Conference (UPEC)*, pp. 1–5. <https://doi.org/10.1109/UPEC.2015.7339774> (2015)
9. Hernandez Jimenez, V.J., Castronuovo, E.D.: Optimal geometric configurations for mitigation of magnetic fields of underground power lines. *IEEE Conference Publications. IEEE Eindhoven PowerTech.2015*, pp. 1–6 (2015). <https://doi.org/10.1109/PTC.2015.7232457>
10. Karady, G.G., Nunez, C.V., Raghavan, R.: The feasibility of magnetic field reduction by phase relationship optimization in cable systems. *IEEE Trans. Power Delivery* 13(2), 647–654 (1998). <https://doi.org/10.1109/61.660956>
11. Del-Pino-Lopez, J.C.: Magnetic field shielding optimization in underground power cable duct banks. *Electric Power Syst. Res.* 114, 21–27 (2014). <https://doi.org/10.1016/j.epsr.2014.04.001>
12. Greshnyakov, G.V.: Comprehensive assessment of the technical and operational characteristics of XLPE—medium and high voltage cable systems: Ph.D. thesis, St. Petersburg, 236 p (2018)
13. Dubitsky, S.D., Korovkin, N.V., Babkov, E.A.: Thermal resistance of a lightning protection cable with an optical fiber to a direct lightning strike. *Electric. Eng. News.* 4(70) (2011)
14. Greshnyakov, G.V., Naryshkin, E.V.: Pulse low-inductance high-voltage cable. *Power Electron.* 4, 42–46 (2009)
15. GOST R IEC 60287-1-1-2009. Electric cables. Calculation of the current rating. Part 1–1. Current rating equations (100% load factor) and calculation of losses. General. Moscow: Standartinform Publications (2009)
16. Akhmetyanov, R.R., Delegodina, L.A., Kopylova, N.P., et al.: Problems of forecasting energy consumption in the integrated ASKUE of the Novosibirsk Scientific Center. *Energoberezhnie.* 1, 42–47 (2007)
17. Akhmetyanov, R.R., Delegodina, L.A., Kopylova, N.P., Lutsenko, B.N., Sobstel, G.M., Cheydo, G.P.: Multiplicative model of seasonal energy consumption of enterprise. *Avtometriya* 44(3), 106–118 (2008)
18. Korzhov, A.V., Sidorov, A.I., E., Yurchenko, Nikolaevsky, A.B.: Mathematical model of damageability of insulation of power cable lines of urban electrical networks. *Electric Stations* 8, 40–47 (2008)
19. Cipolla-Ficarra, F.V.: Handbook of research on interactive information quality in expanding social network communications. Latin Association of Human-Computer Interaction, Spain & International Association of Interactive Communication, Italy, 449 p (2015)
20. Machado, V.M.: Magnetic field mitigation shielding of underground power cables. *IEEE Trans. Magn.* 48(2), 707–710 (2012). <https://doi.org/10.1109/TMAG.2011.2174775>

# **Fuzzy Models and Algorithms**

# Algorithm for Configuring Sugeno-Type Fuzzy Inference Systems Based on the Nearest Neighbor Method for Use in Cyber-Physical Systems



Mikhail Golosovskiy, Alexey Bogomolov, and Mikhail Balandov

**Abstract** The chapter presents an algorithm for configuring fuzzy inference systems of the Sugeno type of zero-order based on statistical data. The algorithm is based on the allocation of a set of reference points in the space of input variables, the values of the output variable which are calculated using the nearest neighbor method. The method consists in finding the nearest points to the reference points in a space of smaller dimensions and using the weighted average for the final calculations. The results of the study of the algorithm, showing the effectiveness of its application for the synthesis of intelligent information systems for various purposes, are presented.

**Keywords** Intelligent information system · Fuzzy inference system · Fuzzy inference system turning · Sugeno type system · Weighted arithmetic mean · Membership function

## 1 Introduction

Fuzzy inference systems (FIS) form the basis of intelligent information systems for various purposes [1–4]. In the management of cyber-physical systems, FIS began to be used after substantiating the possibilities of the synthesis of fuzzy regulators by the example of an effective solution to the problem of controlling an inverted pendulum [1]. The next reason for using fuzzy inference systems in cyber-physical systems is the ability of FIS to approximate a continuous function with a priori specified accuracy  $f: X \rightarrow Y$ , where  $(X)$  is compact [2, 5, 6]. At the same time, the main limitation of the use of FIS in practice is that the number of rules necessary for the synthesis of a complete FIS increases exponentially and is limited from above by the value  $kn$  (where  $n$ —number of input variables,  $k$ —maximum number of

---

M. Golosovskiy (✉) · A. Bogomolov · M. Balandov  
Saint Petersburg Institute of Informatics, Russian Academy of Sciences, 39, line 14 of Vasilevsky Island, 199178 Saint Petersburg, Russia  
e-mail: [golosovskiy@yandex.ru](mailto:golosovskiy@yandex.ru)

M. Balandov  
e-mail: [mihail01@yandex.ru](mailto:mihail01@yandex.ru)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022  
A. G. Kravets et al. (eds.), *Cyber-Physical Systems: Intelligent Models and Algorithms*, Studies in Systems, Decision and Control 417,  
[https://doi.org/10.1007/978-3-030-95116-0\\_7](https://doi.org/10.1007/978-3-030-95116-0_7)

membership functions (MF) of input variables). According to the definition given in [2], a fuzzy inference fuzzy system is complete if with each input state  $x^* = (x^*_1, \dots, x^*_n)$ , belonging to the  $X$ , FIS can link some output state  $y^*$  from  $Y$ .

Many authors [7–10] have investigated methods and algorithms for setting up FIS, based on dividing the domain of determining the approximated function of the MF of input variables into  $n$ -dimensional hyper-rectangular regions, where the degree of dimension  $n$  corresponds to the number of input variables. At the same time, the well-known works emphasize the theoretical possibilities of implementing a FIS turning, but insufficient attention is paid directly to the algorithms for FIS turning in cases where the FIS contains a large (more than 100) number of production rules, which makes it difficult to configure the FIS manually [11–14].

It is known that many cyber-physical systems use Sugeno-type FIS [15–18]. The above has determined the relevance of developing an algorithm for configuring Sugeno-type FIS with a completely consistent set of rules based on statistical data linking the arguments and values of the approximated function [19, 20].

## 2 Problem Statement

Lifecycle management of software development in modern IT infrastructure involves the use of test models and algorithms for organizing software testing [8–10]. Any software testing is limited by the time and resources allocated for testing, as well as the duration of the experts' work, therefore, they cannot guarantee an exhaustively complete software check for compliance with the requirements for its functions and characteristics [11, 12, 21–23].

The algorithm is designed to configure zero-order Sugeno-type FIS [19, 24, 25], in which each MF has one and only one point at which its value is maximum and equal to 1 (this point is called the core of the MF).

The domain of definition of each input variable is equally discretely divided into  $m_i$  segments of the same length by reference points  $x_{ij}$  ( $i$ —input variable number,  $j$ —reference point number, the total number of reference points for  $i$ -th input variable is equal to  $m_i + 1$ ). For each reference point  $x_{ij}$  is put in accordance MF of  $i$ -th input variable. The kernel of MF is located at the reference point. When setting the FP, there are no restrictions on their form, but we consider the condition fulfilled:

$$\mu_{ij}(x) \approx 0 \text{ при } x \in [x_{j1}, x_{ij}-1] \text{ и } x \in [x_{ij}+1, x_{im}+1], \quad (1)$$

where:  $\mu_{ij}(x)$ — value of the  $j$ -th MF of the  $i$ -th input variable,  $m$ —the number of segments into which divided the definition area of  $i$ -th input variable by the reference points.



With such a partition, the domain of the function definition, approximated by the FIS, will be divided by the reference points (FP cores) into n-dimensional hyper-rectangles, where n is the number of input variables (if there is only one input variable, the domain of definition is divided into segments) [2].

Next, we will consider FIS systems with triangular MF of input variables, since such functions represent the simplest and most frequently used type of MF when ensuring strict compliance with the condition (1).

A zero-order Sugeno-type FIS consists of rules R:

$$R_k : IF (x_1 \text{ is } A_1^k), \dots, AND (x_n \text{ is } A_n^k) THEN y \text{ is } b_k \tag{2}$$

where k is the number of the production rule, n is the number of input variables, and A is the term set of the i-th input variable, the degree of membership to which is calculated using the corresponding membership function  $\mu_{ik}(x)$ ,  $b_k$ —the value of the conclusion of the k-th rule. For a Sugeno-type FIS of zero-order, the conclusion of the "if-then" inference rule is a real number. the procedure for defuzzification (obtaining a crisp value of the output variable) is performed by the formula [2–4]:

$$y = \frac{\sum_{k=1}^N w_k b_k}{\sum_{k=1}^N w_k}, \tag{3}$$

where y—FIS output value;  $w_k$ —degree of fulfillment of the condition IF of the k-th rule, N—number of rules.

In this case, the turning of the FIS is changing the values of the input variables [ $x'1, x'2, x'n$ ] the output value of the system from the original  $y'$  on the new value  $y'$ . It is provided by changing the conclusions of the FIS rules, performed according to the formula [6]:

$$b'_k = b_k - w_k \frac{\sum_{k=1}^N w_k b_k - y' \sum_{k=1}^N w_k}{\sum_{k=1}^N w_k^2}, \tag{4}$$

where  $b'_k$ —adjusted value of the conclusion of the k-th rule. In the proposed algorithm the adjustment according to formula (4) will be made for the reference points corresponding to the MF cores and located at the vertices of n-dimensional hyper-rectangles.

The use of automated tools and automated test scripts reduces the time spent on organizing software tests.

### 3 Formal Description of the Algorithm

The principle of proposed in the chapter algorithm is based on the calculation of the values of the approximated function at reference points, based on the statistical values of this function at  $m$  nearest points. The calculated values of the approximated function at the reference points are used for the subsequent configuration of the FIS.

The algorithm for configuring the FIS system is represented by the following sequence of operations:

1. Take the first reference point.
2. Set the value of the number of nearest points equal to  $m = n + 1$ .
3. For the reference point select  $m$  points of the statistical data with the minimum distance  $d$  calculated by the formula:

$$d = \sqrt{(x_{1j}^* - x_{1r})^2 + (x_{2j}^* - x_{2r})^2 + \dots + (x_{ij}^* - x_{ir})^2}, \quad (5)$$

where  $i$  is the number of the input variable,  $j$  is the number of the reference point for the  $i$ -th input variable, and  $r$  is the number of the point of the statistical data based on which it is necessary to restore the functional dependence. The number of points  $m$  is chosen by one greater than the number of input variables.

1. If the number of points with  $d = 0$  is greater than  $m$ , then all such points are selected, and a value equal to the number of selected points is written in  $m$ . The algorithm goes to step 7.
2. Normalized the distance between the reference point and selected points in relation to the maximum  $d_{max}$  value. To perform the normalization operation: for each selected point with the number  $l$  from 1 to  $m$  calculate  $d_l = d_l/d_{max}$ .
3. For each  $l$ -th point calculate weight  $v_l$  by the formula:

$$v_l = \frac{1}{d_l + 0.01}, \quad (6)$$

where 0,01—an arbitrary coefficient entered to avoid division by zero and to set the maximum weight value. The basic principle of weight generation is to make the nearest points have the most weight and the weight of the most distant points as close to 1 as possible.

1. If all points with a distance  $d = 0$ , were selected, then calculate the value for setting up the fuzzy inference system at the reference point  $y'$  as the arithmetic means of the values of the points:

$$y' = \frac{\sum_{l=1}^m d_l}{m}, \quad (7)$$

2. Else calculate it as the weighted arithmetic means by the formula:

$$y' = \frac{\sum_{l=1}^m v_l d_l}{\sum_{l=1}^m v_l}, \tag{8}$$

3. Adjust the value of the output of the FIS at each reference point to the value of the approximated function  $y'$ , calculated in step 7 using the formula (4).
4. Is the turning done for all reference points? If yes,—stop. Else take the next reference point and go to step 3.

### 4 Research of the Effectiveness of the Developed Algorithm

To study the efficiency of the proposed algorithm, and in particular, the ability of the proposed algorithm to adjust the aperiodic functions with several local maxima and minima, an arbitrary function of two variables with several maxima and minima is used (the term “Initial function” will be used):

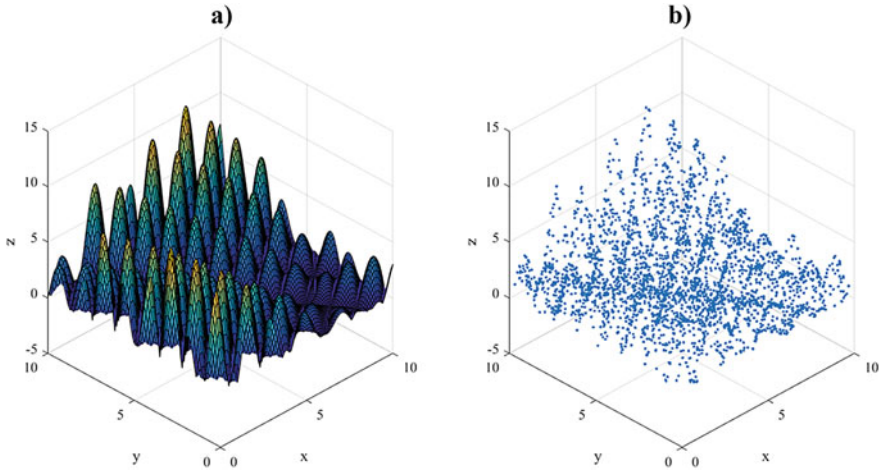
$$f(x_1, x_2) = |\sin(0.882 * x_1 + 1.2135 * x_2 - 7.5) * \cos(1.764 * x_2 - 2.427 * x_1 - 15) * \sqrt{|2.1406 * x_2^2 - 1.556 * x_1^2 - 0.031 * x_1 * x_2 - 31.432 * x_2 + 4.937 * x_1 + 112.5|}| \tag{9}$$

Checking the FIS settings is performed on noisy data. Therefore, to configure the FIS, we do not take points with the exact values of the function, but with values generated according to a random distribution law with a mathematical expectation that changes in accordance with the Initial function (9). For this, a random set of pairs of values of the input variables is generated  $(x_{1i}, x_{2i})_i$  with ranges of values:

$$x_{1i} \in [0, 10] \text{ and } x_{2i} \in [0, 10],$$

where  $i$  is the number of the generated value of the input variable. For each generated pair of input variables, the value  $y_i$  is generated according to the law of normal distribution with the mathematical expectation calculated for the value  $(x_{1i}, x_{2i})_i$  according to the formula (9) and the standard (mean square) deviation equal to 1.3. Figure 1 shows a graph of the original function and a graphical representation of the coordinates of the 3000 generated points with the average value given by the original function.

The comparison of the FIS configured using the proposed algorithm with the FIS configured based on the values of the original function at the reference points was performed according to the following indicators:



**Fig. 1** On the left is a graph of the original function, on the right is a graphical representation of the values for the 3000 generated points with the average value set by the original function

1. The maximum and average error values of the configured FIS for a different number of MF in input variables.
2. The average error value of the configured FIS from the number of points based on which it was configured.

The maximum error value  $\varepsilon_{max}$  was calculated using the formula:

$$\varepsilon_{max} = \max_{i=1...K} |h(x_{1i}, x_{2i}) - f(x_{1i}, x_{2i})|, \tag{10}$$

the average error value  $\varepsilon$  is calculated using the formula:

$$\varepsilon = \frac{1}{K} \sum_{i=1}^K |h(x_{1i}, x_{2i}) - f(x_{1i}, x_{2i})|. \tag{11}$$

In formulas (10)–(11)  $h(\times 1, \times 2)$ —calculation of the value using the configured FIS,  $f(\times 1, \times 2)$ —calculating a value of Initial function using the formula (9),  $K$ —the number of points used to perform the calculation. Additionally, the algorithm was compared with the system setup using the ANFIS neuro-fuzzy model. The parameters of the ANFIS model are shown in the Table 1.

To test the algorithm, depending on the number of MF,  $K = 10,000$  was used, and the number of points for adjusting a given area was  $N = 6$ .

To compare the error values depending on the number of MF, the number of MF of each input variable  $x$  and  $y$  were taken the same and increased in the interval [3, 40] with a step of 1. A graphical representation of the calculation results for the FIS (hereinafter FIS-a) configured using the proposed algorithm, for the FIS (hereinafter FIS-b) configured based on the Initial function (according to formula (9)) and for

**Table 1** Parameters of the ANFIS model

Parameter title	Parameter value
Number of epochs	10
Target error value on the test sample	0
Initial step	0,01
Step increase	0,9
Step Reduction	1,1

the FIS (hereinafter FIS-c) configured using ANFIS is shown in Fig. 2. The results of the calculations are shown in Table 2.

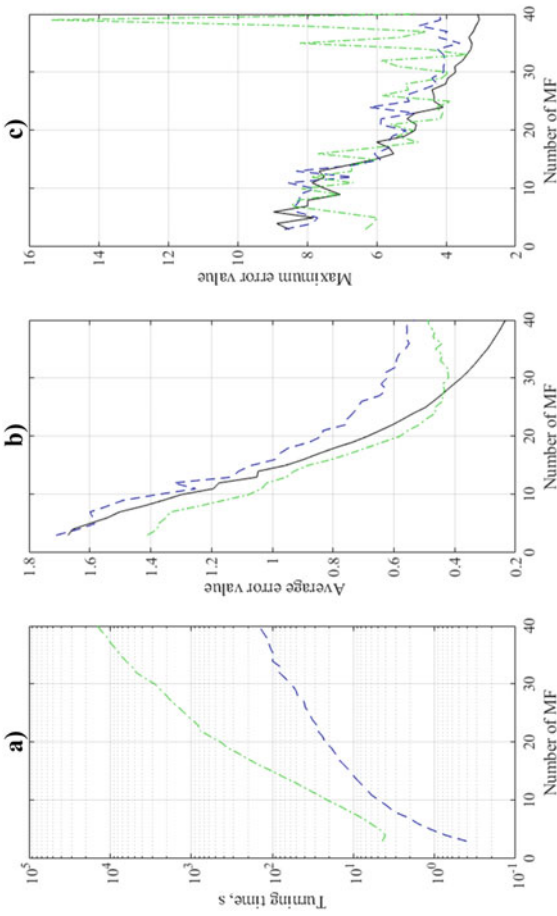
The obtained results allow us to conclude that with an increase in the number of MF, and as a result with an increase in the number of reference points, the average error of the FIS configured using the proposed algorithm decreases. The value of the average error of the FIS configured using the proposed algorithm is comparable to the error of the FIS configured based on the Initial function. At the same time, the maximum error values of the FIS configured using the proposed method can be several times higher than the maximum error value of the FIS configured based on the exact value of the function, which can be explained by the occurrence of individual “peaks” of maxima or minima corresponding to outliers—points with a strong deviation from the mathematical expectation.

Examples of surfaces obtained using FIS with 15 MF of each input variable configured using the algorithm proposed in the chapter based on 3000 points and based on the Initial function are shown in Fig. 3.

A graphical representation of the calculation results for the proposed algorithm and the FIS configured based on the exact values of the function (according to formula (8)) is shown in Fig. 4. The calculation results are shown in Table 3.

To study the dependence of the average error value of the configured FIS on the number of points on the basis of which the turning is made, FIS were used, in which the number of MF for each input variable changed in the interval [3, 30] with a step of 1. The number of random points for tuning was taken in the interval [4, 2000] with a step of 200, and in the interval [2000, 10000] with a step of 1000.

Based on their results, it follows that the average value of the function approximation error decreases with an increase in the number of tuning points and the number of MF. At the same time, it is always greater than the error value of the FIS, configured based on the exact value of the function, although it asymptotically approaches it. The error value decreases non-linearly. For the example considered, an increase in the number of MF greater than 17 with the number of construction points greater than 1600 gives a decrease in the error value of no more than 10% relative to the maximum and minimum error values. And with an increase in the number of MF, the error value may increase, which is associated with a decrease in the number of adjustment points located in the area between the reference points.



**Fig. 2** Graphs of the dependence of the execution time and the error value on the number of MF for: on the left of the execution time, in the center of the average, and on the right of the maximum value of the error value (the dashed line shows the dependencies for FIS-b, the solid line for FIS-a, the dotted line for FIS-c)

**Table 2** The values of the maximum, minimum, and average error values of the configured FIS, depending on the number of MF

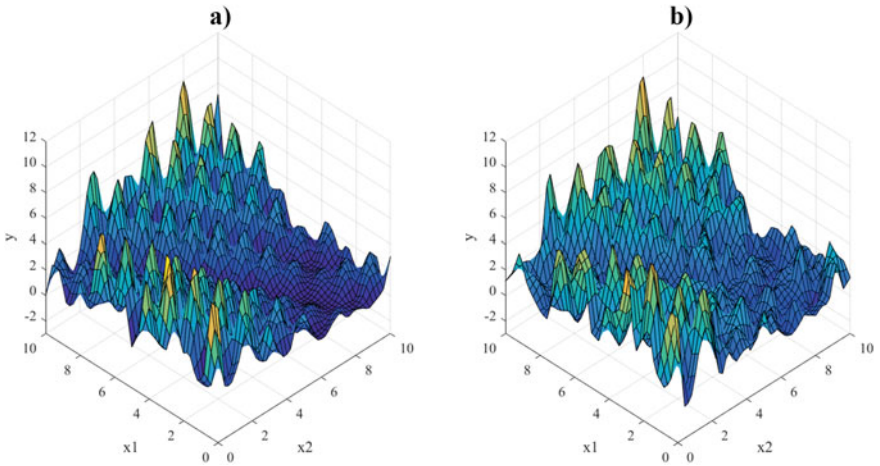
Number MF	Turning time, s			Average error value $\epsilon$			Maximum error value $\epsilon_{max}$		
	FIS-a	FIS-c	FIS-b	FIS-a	FIS-b	FIS-c	FIS-a	FIS-b	FIS-c
3	0,401	4,414	1,6717	1,7111	1,6717	1,4520	8,6495	8,5097	6,9306
4	0,720	4,027	1,6574	1,6400	1,6574	1,3975	7,8502	8,8720	6,1914
5	1,098	4,714	1,6018	1,5841	1,6018	1,3889	7,6924	7,8202	6,5091
6	1,593	6,094	1,5428	1,5896	1,5428	1,3765	8,3012	8,9631	6,4486
7	2,074	7,858	1,5025	1,5992	1,5025	1,3659	8,3384	7,9827	8,0571
8	3,069	10,742	1,4253	1,5424	1,4253	1,3273	8,2127	7,9776	7,9069
9	3,795	14,628	1,3593	1,4909	1,3593	1,2861	8,2241	7,0440	7,1791
10	4,782	20,556	1,2988	1,3789	1,2988	1,2183	7,7742	7,4793	6,5377
11	6,057	27,630	1,1934	1,2509	1,1934	1,1643	8,5429	7,8561	6,6929
12	6,959	37,709	1,1753	1,3255	1,1753	1,1339	6,6938	7,5009	7,3648
13	8,233	50,828	1,0506	1,1327	1,0506	1,0669	8,3104	7,6519	6,2862
14	9,651	71,854	1,0465	1,1108	1,0465	1,0353	6,6926	6,8502	6,3757
15	10,924	98,746	0,9574	1,0700	0,9574	0,9948	5,8881	6,0719	6,0395
16	12,787	140,888	0,8988	0,9923	0,8988	0,9406	6,0681	5,5097	7,5716
17	15,047	191,551	0,8492	0,9776	0,8492	0,8917	5,6491	5,6356	6,0577
18	16,596	254,131	0,7961	0,9468	0,7961	0,8388	5,6557	5,9954	5,6213
19	18,613	349,137	0,7361	0,8783	0,7361	0,7781	5,5533	5,1999	6,0433
20	22,226	429,067	0,6878	0,8381	0,6878	0,7301	5,1436	4,9016	5,0920
21	24,027	581,233	0,6460	0,8331	0,6460	0,6877	5,8841	4,8553	4,7483

(continued)

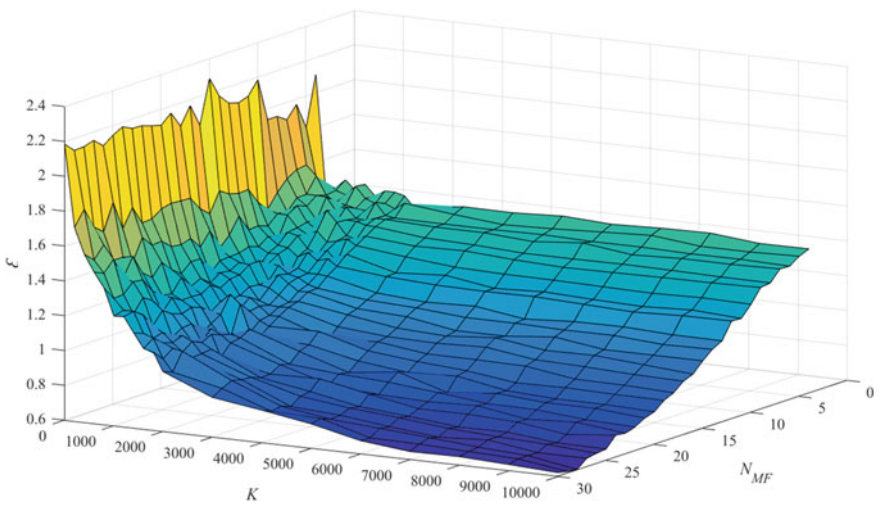
**Table 2** (continued)

Number MF	Turning time, s			Average error value $\varepsilon$			Maximum error value $\varepsilon_{max}$		
	FIS-a	FIS-c		FIS-a	FIS-b	FIS-c	FIS-a	FIS-b	FIS-c
22	27,521	772,898		0,7587	0,6049	0,6330	5,8707	5,1251	4,9687
23	28,601	828,127		0,7534	0,5690	0,6088	4,8893	4,8922	4,9052
24	32,491	1002,569		0,7343	0,5348	0,5753	6,2154	4,0834	5,1993
25	34,954	1201,972		0,7193	0,4960	0,5581	5,0527	4,3409	4,7861
26	39,168	1428,517		0,7067	0,4713	0,5495	5,1898	4,3483	4,2905
27	40,638	1744,700		0,6474	0,4469	0,5222	4,6928	4,4042	4,2763
28	48,451	2017,159		0,6237	0,4252	0,5047	4,2516	4,0149	4,5346
29	51,541	2381,771		0,6425	0,4027	0,5130	4,4131	3,9457	3,9940
30	58,094	2792,764		0,6234	0,3794	0,5320	4,0747	3,7275	3,9661
31	69,091	3720,980		0,6267	0,3594	0,4949	4,0736	3,7501	5,8739
32	82,350	4766,656		0,5909	0,3415	0,5029	4,0285	3,5270	3,9540
33	80,046	5396,679		0,5928	0,3248	0,5221	4,0728	3,4045	4,7360
34	100,170	6417,390		0,5849	0,3097	0,5356	4,2528	3,2817	4,5139
35	95,004	7441,297		0,5625	0,2930	0,5076	3,6009	3,2584	6,6351
36	104,054	8434,728		0,5498	0,2798	0,5210	4,2133	3,3420	4,5811
37	110,481	9750,247		0,5596	0,2678	0,5368	3,9595	3,2187	5,8125
38	117,537	10,816,811		0,5570	0,2553	0,5606	4,7792	3,1413	7,3459
39	131,195	12,631,510		0,5577	0,2441	0,5246	4,1563	3,0457	4,8927
40	146,308	14,289,072		0,5336	0,2336	0,5396	4,1937	3,0872	8,7088





**Fig. 3** Examples of surfaces obtained using the FIS with 25 MF for each input variable: on the left, configured based on the Initial function, on the right, configured using the proposed algorithm based on 3000 points



**Fig. 4** Graphs of the dependence of the average error value of the tuned model on the number of turning points and the number of MF of the input variables (expanded by 90° relative to the z-axis).  $K$  is the number of adjustment points,  $N_{MF}$  is the number of membership functions,  $\epsilon$  is the average error value

**Table 3** Values of the average error value of the tuned FIS from the number of FP ( $N$  FP) and the number of tuning points

N	Number of adjustment points																
	MF	4	204	604	1004	1404	1604	1804	2000	3000	4000	5000	6000	7000	8000	9000	10,000
3	2,105	1,446	1,520	1,508	1,483	1,482	1,467	1,407	1,443	1,441	1,455	1,441	1,441	1,450	1,454	1,427	1,429
4	1,821	1,547	1,479	1,515	1,428	1,436	1,411	1,450	1,418	1,392	1,401	1,401	1,414	1,427	1,397	1,397	1,405
5	1,971	1,630	1,469	1,426	1,408	1,460	1,392	1,383	1,381	1,381	1,385	1,378	1,378	1,379	1,372	1,374	1,366
6	1,900	1,622	1,486	1,409	1,380	1,376	1,394	1,394	1,382	1,382	1,363	1,389	1,353	1,353	1,364	1,363	1,369
7	1,938	1,588	1,443	1,432	1,390	1,388	1,372	1,370	1,339	1,331	1,315	1,315	1,315	1,317	1,314	1,309	1,307
8	1,939	1,527	1,465	1,340	1,394	1,325	1,309	1,334	1,309	1,315	1,297	1,285	1,288	1,288	1,292	1,294	1,299
9	2,186	1,450	1,429	1,331	1,300	1,311	1,330	1,299	1,266	1,240	1,232	1,230	1,225	1,220	1,220	1,227	1,221
10	2,110	1,508	1,308	1,342	1,246	1,277	1,216	1,266	1,193	1,212	1,169	1,181	1,170	1,170	1,162	1,175	1,158
11	2,090	1,600	1,343	1,291	1,276	1,236	1,254	1,216	1,190	1,174	1,170	1,172	1,161	1,161	1,149	1,160	1,162
12	2,107	1,595	1,303	1,273	1,206	1,182	1,184	1,179	1,115	1,094	1,103	1,062	1,080	1,080	1,082	1,069	1,067
13	2,167	1,611	1,321	1,237	1,190	1,177	1,143	1,128	1,093	1,080	1,057	1,056	1,050	1,050	1,051	1,031	1,036
14	2,285	1,647	1,303	1,272	1,204	1,160	1,122	1,125	1,062	1,047	1,000	1,001	1,001	1,028	0,988	0,989	0,988
15	2,036	1,534	1,300	1,267	1,148	1,113	1,077	1,089	1,045	1,001	0,987	1,010	0,970	0,970	0,973	0,952	0,931
16	2,166	1,581	1,295	1,228	1,092	1,073	1,069	1,073	0,994	0,970	0,964	0,974	0,935	0,935	0,927	0,914	0,908
17	2,067	1,660	1,241	1,221	1,130	1,086	1,039	1,040	0,942	0,959	0,921	0,920	0,916	0,916	0,874	0,871	0,861
18	2,160	1,663	1,343	1,188	1,081	1,076	1,028	1,001	1,004	0,925	0,896	0,876	0,839	0,839	0,843	0,837	0,835
19	2,109	1,668	1,346	1,224	1,184	1,134	0,979	0,961	0,946	0,885	0,852	0,863	0,813	0,813	0,798	0,798	0,788
20	2,124	1,654	1,294	1,190	1,087	0,960	0,986	1,021	0,926	0,838	0,873	0,787	0,815	0,815	0,794	0,793	0,771
21	2,143	1,614	1,336	1,194	1,052	0,994	1,005	0,990	0,871	0,848	0,833	0,806	0,775	0,775	0,772	0,748	0,738
22	2,146	1,610	1,396	1,192	1,013	1,049	1,019	0,979	0,887	0,806	0,780	0,772	0,781	0,781	0,736	0,751	0,720

(continued)

**Table 3** (continued)

N	Number of adjustment points																
	MF	4	204	604	1004	1404	1604	1804	2000	3000	4000	5000	6000	7000	8000	9000	10,000
23	2,173	1,716	1,329	1,226	1,088	0,990	1,041	0,918	0,851	0,857	0,801	0,750	0,751	0,738	0,699	0,728	
24	2,143	1,614	1,372	1,159	1,083	1,015	1,012	0,981	0,849	0,803	0,748	0,725	0,696	0,699	0,670	0,666	
25	2,101	1,782	1,486	1,245	1,050	1,064	1,036	0,972	0,870	0,793	0,758	0,735	0,706	0,690	0,679	0,660	
26	2,151	1,633	1,381	1,281	1,094	1,042	0,994	0,969	0,875	0,753	0,735	0,702	0,674	0,691	0,650	0,650	
27	2,133	1,679	1,506	1,205	1,078	1,059	1,035	1,023	0,888	0,774	0,731	0,711	0,689	0,667	0,651	0,604	
28	2,130	1,799	1,445	1,246	1,126	1,105	0,992	0,922	0,861	0,762	0,753	0,675	0,671	0,661	0,651	0,611	
29	2,184	1,713	1,468	1,229	1,146	1,063	1,048	0,943	0,828	0,787	0,741	0,677	0,647	0,650	0,638	0,623	

## 5 Conclusion

The algorithm considered in the chapter allows us to configure fuzzy inference systems based on statistical data. The advantage of the proposed algorithm is a relatively high speed compared to ANFIS. This can be important for setting up a FIS system with a large number of input variables and membership functions. As a further development of the topic, it is planned to consider the development of algorithms for changing the size of the intervals between reference points to achieve a more accurate adjustment in areas with a large volume of statistical data. An additional direction can be highlighted by considering the possibility of optimizing the algorithm by configuring the FIS system based on the use of the window smoothing method near the reference points.

This work was supported by a grant from the President of the Russian Federation for state support of leading scientific schools of the Russian Federation (NSh-2553.2020.8).

## References

1. Yamakawa, T.: Stabilization of an inverted pendulum by a high-speed fuzzy logic controller hardware system. *Fuzzy Sets Syst.* **32**(2), 161–180 (1989). [https://doi.org/10.1016/0165-0114\(89\)90252-2](https://doi.org/10.1016/0165-0114(89)90252-2)
2. Piegat, A.: *Fuzzy Modeling and Control*, p. 798. Springer, Berlin (2013)
3. Kosko, B.: Fuzzy systems as universal approximators. *IEEE Trans. Comput.* **43**(11), 1329–1333 (1994)
4. Kosko, B.: Global stability of generalized additive fuzzy systems. *IEEE Trans. Syst., Man, Cybern. Part C: Appl. Rev.* **28**(3), 441–452 (1998)
5. Manentia, F., Rossia, F., Goryunov, A., Dyadik, A., Kozin, K., Nadezhdin, I., Mikhalevich, S.: Fuzzy adaptive control system of a non-stationary plant with closed-loop passive identifier. *Resour.-Efficient Technol.* **1**(1), 10–18 (2015)
6. Golosovskiy, M.: Algorithm for local configuration of fuzzy inference systems of the Mamdani type with preservation of interpretability of production rules. *Large-Scale Syst. Control* **74**, 6–22 (2018)
7. Balunov, A.I., Smirnov, M.A., Boykov, S.Y.: Software application for modeling the fractionation process based on the principle of maximum entropy. In: Kravets A.G., Bolshakov A.A., Shcherbakov, M.V., (eds.), *Studies in Systems, Decision and Control. Cyber-Physical Systems: Digital Technologies and Applications*, vol 350, pp. 63–70. Springer (2021). <https://doi.org/10.1007/978-3-030-67892-0>
8. Kravets, A.G., Salnikova, N.A., Shestopalova, E.L.: Development of a module for predictive modeling of technological development trends. In: Kravets, A.G., Bolshakov, A.A., Shcherbakov, M.V., (eds.), *Studies in Systems, Decision and Control. Cyber-Physical Systems: Digital Technologies and Applications*, vol 350, pp. 125–136. Springer (2021). <https://doi.org/10.1007/978-3-030-67892-0>
9. Korobkin, D., Fomenkov, S., Fomenkova, M., Vayngolts, I., Kravets, A.: The software for computation the criteria-based assessments of the morphological features of technical systems. In: Kravets, A.G., Bolshakov, A.A., Shcherbakov, M.V., (eds.), *Studies in Systems, Decision and Control. Cyber-Physical Systems: Digital Technologies and Applications*, vol 350, pp. 161–172. Springer (2021). <https://doi.org/10.1007/978-3-030-67892-0>

10. Bychkov, E.V., Bogomolov, A.V., Kotlovanov, K.Y.: Stochastic mathematical model of internal waves. *Bull. South Ural State Univ.. Ser.: Math. Model., Program. Comput. Softw.* **13**(2), 33–42 (2020). <https://doi.org/10.14529/mmp200203>
11. Larkin, E., Akimenko, T., Bogomolov, A., Krestovnikov, K.: Mathematical model for evaluating fault tolerance of on-board equipment of mobile robot. *Smart Innov., Syst. Technol.* **187**, 383–393 (2021). [https://doi.org/10.1007/978-981-15-5580-0\\_31](https://doi.org/10.1007/978-981-15-5580-0_31)
12. Larkin, E., Bogomolov, A., Antonov, M.: Modeling of Increased Rigidity of Industrial Manipulator. *Lecture Notes in Computer Science*, vol. 12336, LNAI, pp. 170–178 (2020). [https://doi.org/10.1007/978-3-030-60337-3\\_17](https://doi.org/10.1007/978-3-030-60337-3_17)
13. Davydenko, A., Sai, C., Shcherbakov, M.: Forecast evaluation techniques for i4.0 systems. In: Kravets, A.G., Bolshakov, A.A., Shcherbakov, M.V. (eds.), *Studies in Systems, Decision and Control. Cyber-Physical Systems: Modelling and Intelligent Control*, vol. 338, pp. 79–102. Springer (2021). <https://doi.org/10.1007/978-3-030-66077-2>
14. Golosovskiy, M.S., Bogomolov, A.V., Terebov, D.S., Evtushenko, E.V.: Algorithm to adjust fuzzy inference system of Mamdani type. *Bull. South Ural State Univ.. Ser.: Math. Mech. Phys.* **10**(3), 19–29 (2018). <https://doi.org/10.14529/mmph180303>
15. Brabant, Q., Couceiro, M., Dubois, D., Prade, H., Rico, A.: Learning rule sets and Sugeno integrals for monotonic classification problems. *Fuzzy Sets Syst.* **401**, 4–37 (2020). <https://doi.org/10.1016/j.fss.2020.01.006>
16. Ben-Ari, M., Mondada, F.: Fuzzy logic control. In: *Elements of Robotics*. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-62533-1\\_11](https://doi.org/10.1007/978-3-319-62533-1_11)
17. Zadeh, L.A.: Fuzzy sets. *Inf. Control* **8**, 338–353 (1965). [https://doi.org/10.1016/S0019-9958\(65\)90241-X](https://doi.org/10.1016/S0019-9958(65)90241-X)
18. Zadeh, L.A.: Is there a need for fuzzy logic? In: *Information Sciences*, vol. 178, pp. 2751–2779 (2008)
19. Sugeno, M., Terano, T., Asai, K.: *Applied Fuzzy Systems*. Morgan Kaufmann Publishers, Burlington (1994)
20. Nobile, M.S., et al.: Fuzzy Self-Tuning PSO: a settings-free algorithm for global optimization. *Swarm Evol. Comput.* **39**, 70–85 (2018)
21. Jain, D.K., Kumar, A., Sharma, V.: Tweet recommender model using adaptive neuro-fuzzy inference system. *Fut. Gener. Comput. Syst.* **112**, 996–1009 (2020)
22. Bounabi, M., Moutaouakil, K., Satori, K.: Association models to select the best rules for fuzzy inference system. *Adv. Intell. Syst. Comput.* **1076**, 349–357. (2020)
23. Zimmermann, H.: *Fuzzy Set Theory and Its Applications*. Kluwer Academic Publishers, Boston (2001)
24. Buckley, J.J., Jowers, L.J.: *Simulating Continuous Fuzzy Systems*, p. 202. Springer, Berlin (2006)
25. Tobin, D.S., Golosovsky, M.S., Bogomolov, A.V.: Technology for ensuring the accuracy of information during network examinations. In: *Communications in Computer and Information Science*, pp. 623–632 (2020). <https://doi.org/10.25559/SITITO.16.202003.623-632>

# Construction of Membership Functions for Fuzzy Management of Security Information and Events



Igor Kotenko  and Igor Parashchuk 

**Abstract** The object of the research in the chapter is the decision-making processes in special security information and event management (SIEM) systems designed for (near) real-time analysis of security threats to modern cyber-physical systems. Some mathematical methods for constructing a membership function as applied to decision-making problems within the framework of fuzzy management of security information and events of cyber-physical systems are studied. The values of the membership function are determined for the decision-making problems on whether a particular computer attack belongs to a fuzzy set of dangerous attacks (a set of attacks of a high level of danger). We consider fuzzy algorithms for solving this problem using a variant of constructing membership functions according to a probabilistic scheme and a method based on the representation of membership functions as functions of the probability density of clear random boundaries between terms of a linguistic variable. At the same time, both the method of constructing membership functions based on the analysis of probability density functions and the method using a simple probabilistic scheme, do not have high mathematical and computational complexity but allow taking into account the uncertainty (fuzziness) of the observed and controlled security parameters, which provides an increase in the reliability of control for security information and events within the framework of security fuzzy management for systems of this class.

**Keywords** Cyber-physical system · Security information and event management system · Membership function · Fuzzy control · Probability · Computer attack · Security parameters · Threats

---

I. Kotenko (✉) · I. Parashchuk  
St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences (SPIIRAS), 39, 14 Liniya, St. Petersburg 199178, Russia  
e-mail: [ivkote@comsec.spb.ru](mailto:ivkote@comsec.spb.ru)

I. Parashchuk  
e-mail: [parashchuk@comsec.spb.ru](mailto:parashchuk@comsec.spb.ru)

## 1 Introduction

Unlike classical controlled engineering systems, for example, systems of mechanical engineering, energy, and industrial production, modern cyber-physical systems (CPSs) are large and multicomponent engineering objects implemented based on seamless integration of computational algorithms and embedded physical components. The use of CPSs allows increasing the adaptability, scalability, fault tolerance, and security of engineering systems, as well as the ergonomics (convenience) of their use. It is generally accepted that in systems of this class, the cybernetic and physical components are closely integrated at all scales and all levels within a single information space using sensors and sensors.

CPSs allow organizational and technical interconnection of heterogeneous discrete and continuous subsystems, objects, and processes, integrate the cybernetic component, computer hardware, and software technologies. Within the framework of the CPS, forces, means and processes are combined. These systems are based on the integration of cyber, technological, physical (resource), and information spaces. In other words, these are systems in which resources, technologies, computing elements, and elements of physical nature are interconnected, serving as both consumers and sources of information. The process of functioning of the CPS is based on the exchange of information, therefore, an important area of research—the management of events and incidents of information security—continues to be relevant.

For this purpose, special systems for managing security information and events (Security Information and Event Management, SIEM) are being created, designed to analyze in real-time security events (threats) emanating from network devices and applications, and allowing one to respond to these events (threats) in near real-time. Their task is to prevent significant damage to the integrity, confidentiality, and availability of data [1]. The tasks that SIEM systems solve are as follows: collecting, summarizing, and storing event logs and other security information from a variety of heterogeneous sources; automatic notification and visualization of warnings; use of tools for analysis of incidents and event analysis; analysis and processing of events using data mining methods; preparation of data for investigations (forensic).

The main advantage of SIEM systems is the detection of security threats and attacks at the early stages of their manifestation. In addition, SIEM systems provide forecasting of the behavior of CPSs in the course of a negative impact, which allows the timely development of adequate measures to counter attacks [1, 2].

The processes for monitoring and managing security events and incidents implemented in modern SIEM systems have a complicated hierarchy. But they are traditionally based on decision-making and decision support procedures. The procedures and algorithms for decision-making and its support for the management of security information and events are aimed at implementation at all stages of collecting and analyzing data from digital network content control systems to detect and counteract CPS security threats. At the same time, the tasks of collecting and processing data on security events from distributed sensors, tasks of processing big

data for preliminary analysis of security events, and tasks of assessing security and many others are solved [1–3].

At the same time, the analysis of restrictions on data reliability for making such decisions and solving the listed problems leads to the need to search for a sufficiently rigorous analytical description of the process of developing control actions in a fuzzy environment. In our opinion, these constraints can be taken into account most correctly within the framework of fuzzy control methods considered in several modern works [4–25]. In these works, it is noted that one of the modern approaches to the implementation of such tasks, to obtaining adequate and reliable results of processing big data arrays, the results of decision-making and its support, are the methods and algorithms of fuzzy control.

The chapter proposes the fuzzy algorithms of construction of membership functions for fuzzy management of security information and events and the method for the representation of membership functions as functions of the probability density of clear random boundaries between terms of a linguistic variable. These results allow taking into account the fuzziness of the observed and controlled security parameters, which provides an increase in the reliability of control for security information and events within the framework of security fuzzy management.

The chapter is organized as follows. The related work is outlined in the second section. The algorithms for constructing membership functions of fuzzy sets are suggested in the third section. The method for constructing membership functions based on the analysis of functions of probability density and discussion are considered in the fourth section. The fifth section is the conclusion and plans for further research.

## 2 Related Work

Technological and methodological approaches used in solving the problems of security management and decision-making on security information and event management are known. The work [3] is devoted to the problems of developing procedures for making such decisions. The proposed algorithms for making decisions on the control and management of information security differ in approach, depth of analysis, and set of controlled parameters. However, the methodological approaches and decision-making algorithms considered are complex from the point of view of mathematical implementation and are not unified for various specific decision-making processes. They are not able to take into account the uncertainty (fuzziness) of the observed and controlled parameters during decision-making. They are not adapted to the tasks of the so-called fuzzy control.

The work [4] considers the issues of fuzzy identification of systems and its application to modelling and control. Here, a generalized management model is proposed in abstract terms, which does not allow it to be fully used to develop control actions to identify specific threats and prevent them within the SIEM framework.

In [5–7] it is argued that fuzzy control algorithms can be implemented using the results of the analysis of a variety of expert opinions. But these algorithms require



significant expenditures for collecting statistics of the initial data that characterize large amounts of expert knowledge. This knowledge serves, in turn, to identify the states of parameters and security indicators in the interests of the efficient operation of SIEM systems. Such approaches to the construction of fuzzy control algorithms will not be able to meet the user's needs for unification and usability (simplicity) of use; they are complex in terms of describing large amounts of expert knowledge.

Thus, in [8–11], applications are proposed for control procedures based on fuzzy clustering algorithms, or on algorithms for constructing “decision trees”. But the methods used in these works for determining the membership functions of fuzzy sets are ineffective and not always reliable due to the need to collect and process a set of related data, which is not always possible in practice.

There are known approaches to solving fuzzy control problems and formulating membership functions based on aggregation (geometric and non-geometric) functions [12, 13]. But these approaches are focused mainly on the intuitionistic fuzzy set, which contains both the membership function and the function of non-membership of some arbitrary element in the fuzzy set. For such sets, intuitionistic indices of fuzziness are determined, which does not quite correspond to the tasks faced by SIEM. The work [14] develops the theory of intuitionistic fuzzy sets. This work is devoted to the so-called Pythagorean fuzzy sets, not directly related to the Pythagorean theorem, but using the postulates proved by Pythagoras (including operations with complex numbers) in the formulation and search for membership functions. This method is not bad for analyzing aggregation of satisfaction with criteria and for comparing alternatives, but the formulation of membership functions for solving control problems based on Pythagorean fuzzy algorithms is difficult and time-consuming.

The works [15, 16], as well as, in part, the work [9], analyze and propose to use control algorithms taking into account unreliable initial data. This data can be identified based on the mathematics of fuzzy sets and artificial neural networks. At the same time, the use of such an approach for solving the problems analyzed in our chapter is redundant and requires large computational and time costs (for building and training a neural network).

In [17, 18] an approach to solving problems of fuzzy control and finding values of membership functions based on probabilistic analysis is considered. But this approach is not always accurate, uncontested, does not take into account the values of the probability density functions.

In [19, 20] an interesting approach to the practical application of the algorithms for processing fuzzy sets in problems related to risk management is suggested. Fuzzy risk analysis is simple, accessible, but does not cover the entire range of problems arising in the management of the security information and events (threats) for complex CPS.

The papers [21, 22] are devoted to the possibilities of mathematical descriptions of control processes based on fuzzy logic. Such approaches will be considered in our chapter, these will be options for constructing membership functions of fuzzy sets used in the problems of security information and event management. They are the basis for fuzzy control, but these works do not consider specific management decisions taking into account the values of the probability density. For example, they do not consider the probability density for the lower and upper thresholds for

a particular variable to belong to the entire set of variables, by changing which information and security events are controlled.

The approach considered in [23] is partially free of these shortcomings. However, the fuzzy control model proposed in this work, although it is based on membership functions determined based on alternative identification methods, taking into account uncertainty, is focused on the development of managerial decisions in conditions of local stationarity.

The closest (in physical and mathematical essence) to the idea proposed in this chapter are the papers [24] and [25], where several interacting and complementary approaches to identifying membership functions in the interests of fuzzy management of security information and events are proposed.

Thus, the analysis of related works shows that fuzzy control algorithms are highly versatile, and the combination of simple probabilistic methods for constructing membership functions with methods for their identification based on the analysis of probability density functions opens up ample opportunities for the study of such complex processes as security information and event management.

In other words, it is possible to recognize the unconditional relevance, as well as to assume the theoretical significance and practical possibility of solving the problem of formulating new approaches to decision-making on fuzzy management of security information and events, based on algorithms for constructing membership functions of fuzzy sets.

### 3 Theoretical Part

Let us consider algorithms for constructing membership functions of fuzzy sets as applied to decision-making tasks for security information and event management. Let us say that the terms of a fuzzy linguistic variable are introduced, i.e. qualitative (not quantitative) values of the logical–linguistic variable  $x$ , which characterizes the fuzzy judgments (opinions) of experts and decision-makers, for example, about the current level (degree) of the danger of a particular attack type in CPS: “the level of danger of a particular attack type  $x$  is small” and “the severity level of a particular attack type  $x$  is large”.

It is obvious that any membership function  $\mu(x)$  characterizing (within the framework of making a decision on security information and event management) the current level of danger of a particular attack type for the CPS security is subjective. Although it can reflect the opinion of not one person, a whole team or a group of experts.

To make an informed decision on security information and event management, it is necessary to determine the value of the membership functions characterizing the level (degree) of the danger of a particular attack type. One of the simplest fuzzy algorithms for solving this problem is a variant of constructing membership functions according to a probabilistic scheme [24]. The essence of this approach is as follows. Each of the  $n$  experts is supposed to answer the question of whether the variable  $x$

belongs to the set A – the set defining the term “high level of danger of a particular attack type for the CPS security”. If  $n_1$  experts answer the question in the affirmative, and  $n_2$ —in the negative, then they consider that

$$\mu(x) = \frac{n_1}{(n_1 + n_2)}, \tag{1}$$

and  $n_1 + n_2 = n$ .

There are more complex algorithms, theoretical principles, and practical procedures for constructing membership functions.

Sometimes important concepts of the norm and universal scale are introduced. These theoretical calculations show that set-theoretic operations (identification of membership functions based on a probabilistic set of operations on fuzzy sets) are correct only when membership functions are measured on a scale of relations. If the measurements are performed only in the order scale, then the operations of identification of membership functions based on the minimax will be correct.

There are four basic classification features of algorithms (methods) for constructing membership functions of fuzzy sets [24]: the assumed form of the domain of definition of a fuzzy set—numeric, including discrete ( $a$ ) or continuous ( $b$ ), and non-numeric ( $c$ ); the method of expert survey used—individual ( $d_1$ ) or group ( $d_2$ ); type of used expert information—ordinal ( $e_1$ ) or cardinal ( $e_2$ ); interpretation of expert survey data—probabilistic ( $D$ ) or deterministic ( $N$ ).

The basic theoretical methods for constructing membership functions often include an algorithm of the type  $\langle a, d_1, e_2, N \rangle$ . It is based on a quantitative comparison of the degrees of belongingness by an individual decision-maker.

So, for problems with the participation of an individual decision-maker, an algorithm of the type  $\langle b, d_1, e_2, N \rangle$  is proposed. It is an algorithm for the parametric determination of the membership function of fuzzy sets. Following this algorithm, the form of the function is set axiomatically, and its parameters are directly estimated by the decision-maker. For example, for the case of a triangular shape of the membership function, the decision-maker indicates its parameters  $u_1, u_2, u_3$ , at which it takes on the unit and zero values, i.e.

$$\mu_{\tilde{A}}(u_2) = 1, \tag{2}$$

but for all  $u \leq u_1, u \geq u_3$  there is

$$\mu_{\tilde{A}}(u) = 0. \tag{3}$$

It should be noted that the parametric representation of the membership functions is compact and provides ease of constructing them in practice. At the same time, such a parametric representation of membership functions is associated with the study (and

proof) of the adequacy of the forms used (triangular, trapezoidal, bell-shaped, etc.) and the corresponding analytical descriptions of such membership functions.

A similar approach is used in algorithms based on the use of a standard set of graphs of membership functions. The decision-maker chooses the most suitable graph from the standard set, and then, in a dialogue with the computer, finds out the parameters of this graph and, if necessary, corrects them.

In addition, methods of psychological scaling can also be used to construct membership functions of fuzzy sets. There is an algorithm for constructing membership functions based on the procedure  $\langle b, d_1, e_2, N \rangle$  and focused on calculating the terms of a linguistic variable with a numerical domain based on the equal division method. In this case, the decision-maker is presented with several pairs of points in turn. At each presentation, the decision-maker must name a point for which the degree of membership is in the middle between the degrees of membership of the points included in the presented pair.

One of the most commonly used fuzzy algorithms for constructing membership functions also belongs to the class  $\langle b, d_1, e_2, N \rangle$ . It is based on the representation of membership functions as functions of the probability density of clear random boundaries between the terms of a linguistic variable.

The theoretical analysis of the considered methods allows us to hypothesize that effective approaches to the construction of membership functions in decision-making problems for security information and event management can be found by combining simple probabilistic methods (according to a probabilistic scheme) and a method for constructing membership functions based on the analysis of functions of the probability density. To test this hypothesis, it is proposed to consider the methodological (practical) application of this approach (method) in more detail.

## 4 Methodological Part and Discussion

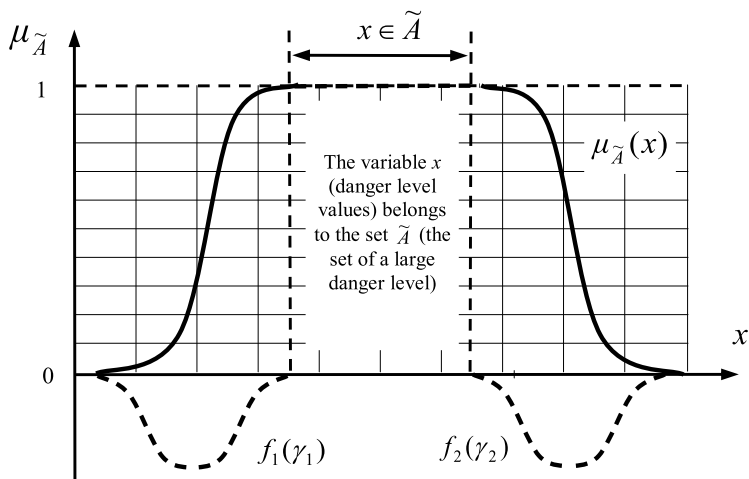
Let us consider an algorithm for constructing membership functions based on the analysis of functions of probability density as applied to decision-making problems for security information and event management, for example, to the problem of decision-making about belongingness of a particular computer attack to a fuzzy set of dangerous attacks (a set of attacks of a high level of danger).

Let some set  $A$  have the physical meaning of the term of the set of values of the linguistic variable "high level of danger of a particular attack type for information security" and is described by an interval  $(\gamma_1, \gamma_2)$ .

In this case, if object  $x$  is the level of danger of a particular attack type  $x > \gamma_1$  and  $x < \gamma_2$ , then  $x \in A$ , otherwise  $x \notin A$ .

If  $\gamma_1$  and  $\gamma_2$  are random variables, then  $A$  is a fuzzy set  $\tilde{A}$ , since there are objects (values of the danger level)  $x$ , relative to which it is impossible to unambiguously assert whether they belong to the set  $A$  (the set of a high danger level) or not.

Let  $f_1(\gamma_1)$  and  $f_2(\gamma_2)$  be the probability density functions for the lower and upper thresholds of membership of the variable  $x$  (the value of the danger level) to the



**Fig. 1** Graphical interpretation of the construction of the membership function

set  $A$  (the set of the high danger level), respectively (Fig. 1). With a probabilistic interpretation of the membership function, we have.

$$\mu_{\tilde{A}}(x) = P(x \in A). \tag{4}$$

Taking into account the introduced notation, we obtain.

$$\mu_{\tilde{A}}(x) = P(\gamma_1 < x < \gamma_2). \tag{5}$$

With the independence of random variables  $\gamma_1$  and  $\gamma_2$ , we find the value of the membership function for our decision-making problem on whether a particular computer attack belongs to a fuzzy set of dangerous attacks (a set of attacks of a high level of danger).

$$\mu_{\tilde{A}}(x) = P(x > \gamma_1) P(x < \gamma_2). \tag{6}$$

We denote by  $F_1(x)$  (see Fig. 1) the probability measure (boundaries).

$$F_1(x) \Leftrightarrow \int_{-\infty}^x f_1(\gamma) d\gamma. \tag{7}$$

and by  $F_2(x)$ —the probability measure (boundaries)

$$F_2(x) \Leftrightarrow \int_{-\infty}^x f_2(\gamma_2) d\gamma \tag{8}$$

By definition  $F_1(x) = P(\gamma_1 < x)$ .

Then we get the final value of the membership function for the decision-making problem within the framework of fuzzy management of security information and events—does a specific computer attack of level  $x$  belong to a fuzzy set of dangerous attacks (a set of attacks of a high level of danger):  $\mu_{\tilde{A}}(x) = F_1(x) [1 - F_2(x)]$ .

The further development and modification of this approach is an algorithm for the simultaneous determination of membership functions of all basic terms of a linguistic variable based on a survey of a group of experts—the type of the technique  $\langle b, d2, e1, D \rangle$ .

In essence, it is an interactive software system for constructing membership functions in terms of linguistic variables based on the considered methods and requirements for a set of such functions.

This approach implies the construction of membership functions in a dialogue mode with a computer-based on the use of a standard set of graphs of membership functions. When implementing this approach, it is necessary to observe some rules that must be followed: the terms of the set  $T(x)$  must be ordered; membership functions of extreme terms should not have the form of bell-shaped functions; each term (concept) must have at least one typical object, i.e.

$$\exists x_i: \mu_{\tilde{A}}(x_i) = 1. \tag{9}$$

for any two terms, the condition

$$0 < \max_x \mu_{\tilde{C}_i \cap \tilde{C}_j}(x) < 1 \tag{10}$$

is true, where  $\tilde{C}_i$  and  $\tilde{C}_j$  are fuzzy sets, the intersection of which gives us the values of the compatibility function.

In other words, at least one concept ( $\mu(x) > 0$ ) corresponds to each  $x$ , and no compatibility function can be completely nested in another compatibility function of a given linguistic variable.

Thus, using the method of representing membership functions as functions of the probability density of clear random boundaries between the terms of a linguistic variable, as well as based on a standard set of graphs of membership functions, it is possible to obtain the values of these functions, for example, for the statement “the value  $x$  is small”. The meaning of this statement for our practical task is to determine the current level (degree) of the danger of a particular attack type for the CPS information security—“the level of danger of a particular attack type  $x$  is small”. To determine the current level (degree) of the danger of a particular attack type for

the CPS information security—“the level of danger of a particular attack type  $x$  is large”—the problem of finding membership functions is solved similarly.

## 5 Conclusion

The considered method shows that it is possible to determine membership functions in such decision-making problems within the framework of fuzzy management of security information and events, not only with the help of procedures for direct questioning of experts but also based on distribution functions  $F_1(x)$  and  $F_2(x)$  with further use of the expression obtained for  $\mu_{\hat{A}}(x)$ . Moreover, the functions  $F_1(x)$  and  $F_2(x)$  themselves can be built based on either statistical data or an expert survey. Thus, several methods of constructing a membership function as applied to decision-making problems within the framework of fuzzy management of security information and events in CPSs are considered. These methods do not have great mathematical and computational complexity, but they allow to take into account the uncertainty (fuzziness) of the observed and controllable security parameters, which makes it possible to increase the reliability of security information and event management of modern CPSs.

Practical application of the proposed methods for constructing the membership function in fuzzy control problems is possible both within the framework of research work and in the systems of automated control of information security for complex information-telecommunication and cyber-physical systems. The direction of further research can be the development of methods that take into account not only the fuzzy but also the contradictory (incomplete) nature of the initial data characterizing the controlled parameters of systems of this class.

**Acknowledgements** This research is being supported by the grant of RSF #21-71-20078 in SPC RAS.

## References

1. Miller, D.R., Harris, S., Vandyke, S.: Security Information and Event Management (SIEM) Implementation. McGrawHill, New York (2011)
2. Kotenko I., Parashchuk I.: Determining the parameters of the mathematical model of the process of searching for harmful information. Cyber-physical systems: industry 4.0 challenges. In: Kravets et al. A.G., (eds.), Studies in Systems, Decision and Control, vol. 260. Springer Nature Switzerland AG, Cham (2020)
3. Maimbo C.: Exploring the Applicability of SIEM Technology in IT Security: Masters thesis. Auckland, Auckland University of Technology (2014)
4. Takagi, T., Sugeno, M.: Fuzzy identification of systems and its application to modelling and control. IEEE Trans. Syst. Man Cybern. **15**, 116–132 (1985)
5. Carlsson C., Fedrizzi M., Fuller R. (2004) Fuzzy Logic in Management, vol. 2004, p. 288. Springer, Boston (1985)

6. Buckley J.J., Eslami E.: *An Introduction to Fuzzy Logic and Fuzzy Sets*, vol. 2002, p. 285. Springer, Berlin (2002)
7. De Barros, L.C., Bassanezi, R.C., Lodwick, W.A.: *A First Course in Fuzzy Logic, Fuzzy Dynamical Systems, and Biomathematics. Theory and Applications*, vol. 2017, 304. Springer, Berlin (2017)
8. Ramya, K., Teekaraman, Y., Ramesh Kumar, K.A.: Fuzzy-based energy management system with decision tree algorithm for power security system. *Int. J. Comput. Intell. Syst.* **12**(2), 1173–1178 (2019)
9. Gaikwad, D., Jagtap, S., Thakare, K., Budhawant, V.: Anomaly based intrusion detection system using artificial neural network and fuzzy clustering. *Int. J. Eng. Sci.* **1**, 1–6 (2012)
10. Hooda, D.S., Raich, V.: *Fuzzy Logic Models and Fuzzy Control. An Introduction*, p. 409. Oxford, Alpha Science International Ltd (2017)
11. Zhang, W.-R.: The road from fuzzy sets to definable causality and bipolar quantum intelligence - to the memory of Lotfi A. Zadeh. *J. Intell. Fuzzy Syst.* **36**(4), 3019–3032 (2019)
12. Beliakov, G., Pradera, A., Calvo, T.: *Aggregation Functions: A Guide for Practitioners. Studies in Fuzziness and Soft Computing*. Springer, Berlin GmbH **2007**, 39–122 (2007)
13. Xu, Z., Yager, R.R.: Some geometric aggregation operators based on intuitionistic fuzzy sets. *Int. J. Gener. Syst.* **35**, 417–433 (2006)
14. Yager, R.R.: Properties and applications of Pythagorean fuzzy sets. *Imprecision and Uncertainty in Information Representation and Processing*. Springer Nature Switzerland AG, Cham **2015**, 119–136 (2015)
15. Kotenko, I.V., Parashchuk, I.B., Omar T.K.: Neuro-fuzzy models in tasks of intelligent data processing for detection and counteraction of inappropriate, dubious and harmful information. II International Scientific and Practical Conference on the Fuzzy Technologies in the Industry (FTI 2018), Ulyanovsk, 23–25 October 2018. CEUR Workshop Proceedings (CEUR-WS), vol. 2258, pp. 116–125 (2018)
16. Parashchuk I.B., Doynikova E.V.: The architecture of subsystem for eliminating an uncertainty in assessment of information objects semantic content based on the methods of incomplete, inconsistent and fuzzy knowledge processing. In: Kotenko, I., et al. (eds.), *The XIII Science Conference of Intelligent Distributed Computing (IDC-2019)*. Studies in Computational Intelligence, vol. 868 (SCI 868), pp. 294–301. Springer Nature Switzerland AG, Cham (2020)
17. Singpurwalla, N.D., Booker, J.M.: Membership functions and probability measures of fuzzy sets. *J. Am. Stat. Assoc.* **99**(467), 867–877 (2004)
18. Mingtian, F., Zuping, Z., Chengmin, W.: Optimization of annual generator maintenance scheduling. Selection of fuzzy membership function. In: *Mathematical Models and Algorithms for Power System Optimization. Modeling Technology for Practical Engineering Problems*, pp. 49–80. Academic, New York (2019)
19. Li-Hua, F., Gao-Yuan, L.: Analysis on fuzzy risk of landfall typhoon in Zhejiang province of China. *Math. Comput. Simul.* **79**(11), 3258–3266 (2009)
20. Qiang, Z., Jianzhong, Z., Chao, Z., Jun, G., Weiping, D., Mengqi, Y., Li L.: Fuzzy risk analysis of flood disasters based on diffused-interior-outer-set model. *Expert Syst. Appl.* **39**(6), 6213–6220 (2012)
21. Bojadziev, G., Bojadziev, M.: *Fuzzy Logic for Business, Finance, and Management*, 2nd edn., p. 252. World Scientific Publishing Co. Pte. Ltd., Singapore (2007) (2007)
22. Chen, G., Trung Tat, P.: *Introduction to Fuzzy Sets, Fuzzy Logic, and Fuzzy Control Systems*, p. 328. CRC Press LLC, Boca Raton (2008)
23. Shahbazova, S.N., Sugeno, M., Kacprzyk, J.: *Recent Developments in Fuzzy Logic and Fuzzy Sets: Dedicated to Lotfi A. Zadeh.*, 1st edn., p. 220. Springer Nature Switzerland AG, Cham (2020)



24. Parashchuk, I.B., Bobrik, I.P.: Fuzzy Sets in Problems of Analysis of Communication Networks, p. 80. St. Petersburg, VUS (2001)
25. Kotenko, I., Parashchuk, I.: Decomposition and formulation of system of features of harmful information based on fuzzy cxsrelationships. In: The 2019 International Russian Automation Conference (RusAutoCon). IEEE Xplore Digital Library: Browse Conferences. 2019, vol. 8867588, pp. 1–5 (2019)

# Building a Fuzzy Model to Automate the Search for Product Names in Procurement Documentation for a Shipbuilding Plant



Lilia Emaletdinova, Aigul Kabirova, and Ivan Kulagin

**Abstract** The chapter discusses a method of building a search model for lexically differently recorded attributes of products and equipment in procurement documentation for a shipbuilding plant based on Tsukamoto's fuzzy inference. Three attributes are considered as the basis for the search: the names of the item and the document for the purchase, as well as the designation of the industry code of the industry, which can be recorded differently by different users. The input data for the model are three linguistic variables: the degree of similarity of the product name, delivery document, and the degree of similarity of the designation of the industry code. These variables are described using fuzzy sets with triangular membership functions. The output of the model is the degree of coincidence of the required record with the names in the procurement documentation. The degree of coincidence is described by fuzzy sets with Gaussian membership functions. The fuzzy inference algorithm is implemented in PHP 5.6, PHP Framework Laravel. The procurement documentation database is implemented in the MySQL system. The results of the algorithm are presented.

**Keywords** Fuzzy logic · Tsukamoto's fuzzy logical inference · Membership functions · Lexical similarity · Procurement documentation

## 1 Introduction

In the modern world, fuzzy logic is used to solve many problems in various scientific fields: in medicine [1–4], in object control [5–9], in robotics [10–13], pattern recognition [14–16], forecasting [17, 18] and etc. The principles of fuzzy inference can be applied to automate the search for product names in procurement documentation for a shipbuilding plant.

In [19], the results of the development of a subsystem for the preparation of procurement documentation for the AIS “Mercury” for JSC “Zelenodolskoye PKB” were presented. Preparation of procurement documentation is an integral part of

---

L. Emaletdinova (✉) · A. Kabirova · I. Kulagin  
Kazan National Research Technical University Named After A.N. Tupolev-KAI, Kazan, Russia

the design and construction of ships and vessels. However, in the course of its use, problems arose associated with the complexity of finding lexically similarly recorded products and equipment in the procurement documentation.

This problem arises due to the fact that the database of attributes of products and equipment is replenished by specialists from various design departments, who can enter information about the same product (equipment) into the database, but with minor differences in the recording of the name and/or attributes of products, and equipment. Verification and correction of similar records are carried out in the preparation of procurement documentation. To automate this process and reduce the time spent on creating procurement documentation in this work it is proposed to use the search model for lexically differently recorded attributes of products and equipment based on Tsukamoto’s fuzzy logical inference.

## 2 Problem Statement

There is a database of attributes of products and equipment, which contains the following data (Fig. 1):

- (A) Position number;
- (B) Product name (100 characters) and Industry Sector Code (ISC) (10 characters);
- (C) Name of the delivery document (50 characters);

№	Product name	Type Index	Name of the delivery document	Document type code	Quantity unit	Code quantity unit	Quantity	Mass unit	Code mass unit	Mass unit	Provider	Original holder
1	2	3	4	5	6	7	8	9	10	11	12	13
<b>Loudspeakers</b>												
1	Loudspeakers with cylindrical glass, two cranes br. DN 6, PN 4 kgf / cm2 OST 5.5419-79 Industry Sector Code: 2942174027		ITSHL.494521.004-08			796	1		166	1.97	JSC "Askold", Arsenyev Primorsky Territory	JSC TstSS KB "Armas", St. Petersburg
2	Loudspeakers with flat corrugated glass, made of carbon steel DN 10 PN 1.0 MPa Industry Sector Code: 2942151011		ITSHL.494521.002-01 (598-03.083-01)			796	8		166	7.2	JSC "Askold", Arsenyev Primorsky Territory	JSC TstSS KB "Armas", St. Petersburg
<b>Taps</b>												
3	Ball valve "lever" 1/2 "bunk bed / bunk bed. Industry Sector Code:					796	1		166	0.2	Trading networks	
<b>Valves</b>												
4	Flanged bronze non-return shut-off valve DN 50, PN 1.0 MPa (10 kgf / cm2) Industry Sector Code: 2912144045		ITSHL.491915.018-02 (522-35.4151-02)			796	1		166	18	JSC "Askold", Arsenyev Primorsky Territory	JSC TstSS KB "Armas", St. Petersburg
5	Non-return shut-off valve, angle choke, bronze DN 25 PN 10.0 MPa (100 kgf / cm2) Industry Sector Code: 2912434033		ITSHL.491911.003-06 (522-01.500-06)			796	2		166	2.9	JSC "Plant" Burevestnik", Gatchina, Leningrad region.	JSC TstSS KB "Armas", St. Petersburg
6	Non-return shut-off valve angle flanged bronze DN 50 PN 2.5 MPa (25 kgf / cm2) Industry Sector Code:		ITSHL.491915.011-02			796	2		166	16	JSC "Zelenodolsk plant named after Gorky", Zelenodolsk	JSC "Zelenodolsk plant named after Gorky", Zelenodolsk

Fig. 1 Fragment of the procurement documentation database

- (D) Fields responsible for quantitative indicators (pieces, tons, kilograms, etc.)
- (E) Provider (100 characters);
- (F) Original holder (50 characters);

It is required to automate the search for information about a product (equipment) in the procurement documentation using fuzzy logical inference and generate a report for a specialist in the procurement documentation department containing information about the product (product name, name of the delivery document, designation of ISC), suspected of duplication, for further analysis and adjustments.

### 3 Method for Solving the Problem

To solve the problem of automating the search, it is proposed to use Tsukamoto's fuzzy logical inference. Let's introduce the following linguistic variables.

$x_1$ —«the degree of similarity of the product name»;

$x_2$ —«the degree of similarity of the delivery document name»;

$x_3$ —«the degree of similarity of the industry sector code designation»;

$p$ —«the degree of coincidence of the information about product/equipment on procurement documentation».

Let's introduce fuzzy sets:

$A_1^1 = \{x_1, \mu_{A_1}(x_1)\}$ —«small degree of similarity of the product name»;

$A_2^1 = \{x_1, \mu_{A_2}(x_1)\}$ —«average degree of similarity of the product name»;

$A_3^1 = \{x_1, \mu_{A_3}(x_1)\}$ —«large degree of similarity of the product name»;

$A_1^2 = \{x_2, \mu_{A_1}(x_2)\}$ —«small degree of similarity of the delivery document name»;

$A_2^2 = \{x_2, \mu_{A_2}(x_2)\}$ —«average degree of similarity of the delivery document name»;

$A_3^2 = \{x_2, \mu_{A_3}(x_2)\}$ —«large degree of similarity of the delivery document name»;

$A_1^3 = \{x_3, \mu_{A_1}(x_3)\}$ —«small degree of similarity of the industry sector code designation»;

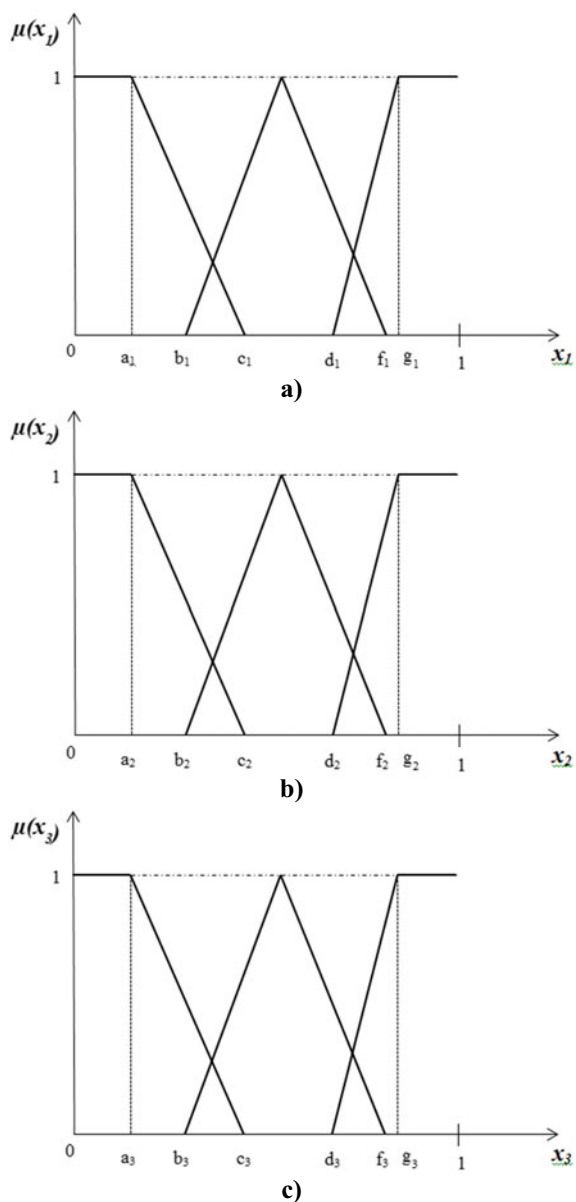
$A_2^3 = \{x_3, \mu_{A_2}(x_3)\}$ —«average degree of similarity of the industry sector code designation»;

$A_3^3 = \{x_3, \mu_{A_3}(x_3)\}$ —«large degree of similarity of the industry sector code designation»;

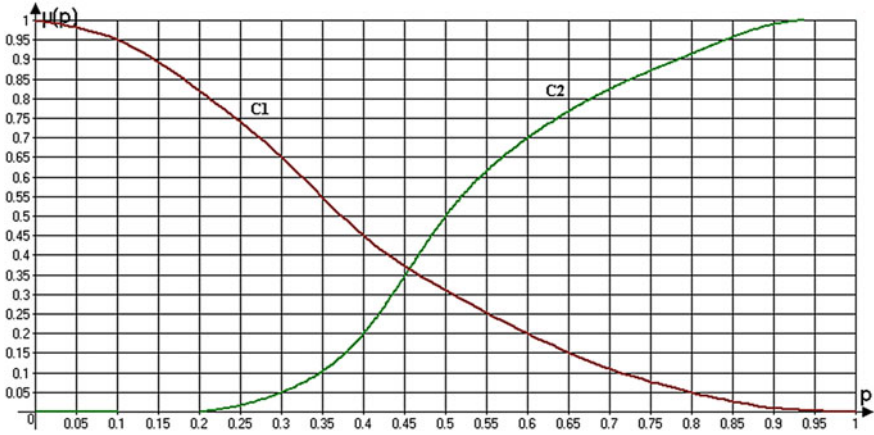
$c_1 = \{p, \mu_{c_1}(p)\}$ —«small degree of coincidence of the information about product/equipment on procurement documentation»;

$c_2 = \{p, \mu_{c_2}(p)\}$ —«large degree of coincidence of the information about product/equipment on procurement documentation».

Input variables  $x_1, x_2, x_3$  are described using fuzzy sets with triangular membership functions (Fig. 2). And the output variable is described by fuzzy sets with Gaussian membership functions (Fig. 3).



**Fig. 2** Membership functions for input variables: **a**  $x_1$ —«the degree of similarity of the product name»; **b**  $x_2$ —«the degree of similarity of the delivery document name»; **c**  $x_3$ —«the degree of similarity of the industry sector code designation»



**Fig. 3** Membership functions of the output variable  $p$ —“the degree of coincidence of the information about product/equipment on procurement documentation”

Based on expert knowledge, a decision-making model is built with 54 rules of fuzzy inference:

R1 : If( $x_1 = 'A_1^1'$ ) and ( $x_2 = 'A_1^2'$ ) and ( $x_3 = 'A_1^3'$ ), THEN  $p_1 = c_1$

R2 : If( $x_1 = 'A_1^1'$ ) and ( $x_2 = 'A_1^2'$ ) and ( $x_3 = 'A_1^3'$ ), THEN  $p_2 = c_2$

R3 : If( $x_1 = 'A_1^1'$ ) and ( $x_2 = 'A_1^2'$ ) and ( $x_3 = 'A_2^2'$ ), THEN  $p_3 = c_1$

R4 : If( $x_1 = 'A_1^1'$ ) and ( $x_2 = 'A_1^2'$ ) and ( $x_3 = 'A_2^3'$ ), THEN  $p_4 = c_2$

R5 : If( $x_1 = 'A_1^1'$ ) and ( $x_2 = 'A_1^2'$ ) and ( $x_3 = 'A_3^3'$ ), THEN  $p_5 = c_1$

R6 : If( $x_1 = 'A_1^1'$ ) and ( $x_2 = 'A_1^2'$ ) and ( $x_3 = 'A_3^3'$ ), THEN  $p_6 = c_2$

R7 : If( $x_1 = 'A_1^1'$ ) and ( $x_2 = 'A_2^2'$ ) and ( $x_3 = 'A_1^3'$ ), THEN  $p_7 = c_1$

R8 : If( $x_1 = 'A_1^1'$ ) and ( $x_2 = 'A_2^2'$ ) and ( $x_3 = 'A_1^3'$ ), THEN  $p_8 = c_2$

...

R53 : If( $x_1 = 'A_3^1'$ ) and ( $x_2 = 'A_3^2'$ ) and ( $x_3 = 'A_3^3'$ ), THEN  $p_{53} = c_1$

R54 : If( $x_1 = 'A_3^1'$ ) and ( $x_2 = 'A_3^2'$ ) and ( $x_3 = 'A_3^3'$ ), THEN  $p_{54} = c_2$

Tsukamoto’s fuzzy inference algorithm, adapted to solve the problem, is as follows:

1. Determines the degree of truth  $\mu(x_1^0), \mu(x_2^0), \mu(x_3^0)$  for the premises of each rule [20].
2. Find the “cut-off” levels  $\alpha_k, k = \overline{1, 54}$  for the premises of each of the rules using the operation “minimum”.

3. For each  $k$ -th rule ( $k = 1, 3, 5, \dots, 53$ ) a clear value of  $p_1$  is determined as a solution to the equation  $\alpha_k = \mu_{c_1}(p_1)$ . The resulting value  $p_1$  determines the value of the output variable of  $k$ -th rule  $p_1^k = p_1$ .
4. For each  $k$ -th rule ( $k = 2, 4, 6, \dots, 54$ ) a clear value of  $p_2$  is determined as a solution to the equation  $\alpha_k = \mu_{c_2}(p_2)$ . The resulting value  $p_2$  determines the value of the output variable of  $k$ -th rule  $p_2^k = p_2$ .
5. Define the clear value of the output variable as:

$$p = \frac{\sum_{k=1, k=k+2}^{53} \alpha_k \cdot p_1^k + \sum_{k=2, k=k+2}^{54} \alpha_k \cdot p_2^k}{\sum_{k=1}^{54} \alpha_k}$$

## 4 Preparation of Initial Data

To determine the values of variables  $x_1, x_2, x_3$ , where  $x_1 \in [0, 1], x_2 \in [0, 1], x_3 \in [0, 1]$  in the work, it is proposed to use the method that determines the degree of similarity of words, based on the use of a matrix, proposed in [21]. This method has been improved to provide the ability to compare strings (phrases). To be able to work with phrases, it is necessary to normalize the string, bring all characters to the same case, and split the string into separate words, which can already be compared by means of the existing algorithm.

The degree of similarity of word combinations is determined by the formula:

$$K_{similarity\ string} = \frac{2 \cdot \sum_{i=1}^n K_{word}}{L_{S_1} + L_{S_2}}$$

$K_{word}$ — it is a measure of the similarity of a word from a string  $S_1$  (original string) with a word from a string  $S_2$  (compared string), which is determined using a matrix that determines the sequential coincidence of characters in words;

$L_{S_1}$ — the number of words per line  $S_1$ ;

$L_{S_2}$ — the number of words per line  $S_2$ ;

$n = L_{S_1}$ .

To assess the speed of the algorithm execution a comparison with the operation of the sequential character comparison algorithm was carried out. Comparison of the algorithms showed that the algorithm for determining the similarity of rows using a matrix works more efficiently. Table 1 shows a short example of the results of the algorithms.

**Table 1** A fragment of the results of comparing strings by two algorithms to determine the degree of complexity of the product name  $x_1$

Source string	Comparing string	Algorithm using matrix		Algorithm of the sequential character comparison	
		Comparison result $x_1 \in [0, 1]$	Time comparison	Comparison result $x_1 \in [0, 1]$	Time comparison
Main switchboard	Main switchboard	1	0.000427	1	0.00133
Main switchboard	Loop-through switch Post	0.19	0.000424	0.32	0.00173
Main switchboard	U-shaped bridge 75–143	0.28	0.000425	0.34	0.00123
Main switchboard	Switchboard	0.9	0.000422	0.72	0.00113

### 5 The Results of Tsukamoto’s Fuzzy Inference Algorithm

Tsukamoto’s fuzzy inference algorithm, adapted to the task at hand, was implemented using the PHP 5.6 programming language, PHP Framework Laravel with a MySQL database connection. The developed AIS “Mercury” module allows a specialist in the procurement documentation department to automatically identify duplicate products in the documentation.

Figure 4 shows an example of a report on identified duplicates in the procurement documentation:

Pos	Product name	Name of the delivery document	Industry code	Similarity level (from 0 to 1)
1	Angle non-return shut-off valve, bronze DN 10 PN 10.0 MPa	ITSHL.491911.004-01 (522-01.499-01)	2912434020	
	Angle non-return shut-off valve, bronze DN 10 PN 10.0 MPa	ITSHL.491911.004-01 (522-01.499-01)	2912434020	1
	Angle non-return shut-off valve, bronze DN 10 PN 10.0 MPa ITSHL.191111.016TU	ITSHL.491911.004-01 (522-01.499-01)	2912434020	0.971
	Angle non-return shut-off valve, bronze DN 10 PN 10.0 MPa ITSHL.191111.016	ITSHL.491911.004-01 (522-01.499-01)	2912434020	0.974
	Angle non-return shut-off valve, bronze	ITSHL.491911.004-01 (522-01.499-01)	2912434020	0.924
	Non-return shut-off valve, angle, bronze DN 10 PN 10.0 MPa	ITSHL.491911.004-01 (522-01.499-01)	2912434020	0.96

**Fig. 4** Fragment of the result of Tsukamoto’s fuzzy inference algorithm



The report displays information about the product, and then a list of duplicates indicating the degree of coincidence of information about the product/equipment in the procurement documentation.

According to the submitted report, a specialist of the procurement documentation department can make a brief analysis to confirm the conclusions of the system and by means of specialized functionality integrated into the interface of the AIS "Mercury" can make an automated replacement of the product in the procurement documentation in accordance with the recommendations of the system.

## 6 Conclusion

Tsukamoto's fuzzy inference algorithm is used to identify lexically differently recorded attributes of products and equipment in the procurement documentation. This allows you to automate the process of creating procurement documentation, which leads to a reduction in the time spent on its creation. The algorithm was implemented and implemented in the AIS "Mercury" at JSC "Zelenodolskoye PKB".

## References

1. Cleophas, T.J., Zwinderman, A.H.: Fuzzy Modeling. Machine Learning in Medicine. Springer, Dordrecht. (2013). [https://doi.org/10.1007/978-94-007-5824-7\\_19](https://doi.org/10.1007/978-94-007-5824-7_19)
2. Helgason, C.M., Jobe, T.H.: Fuzzy logic and continuous cellular automata in warfarin dosing of stroke patients. *Curr. Treat. Options Cardiovasc. Med.* **7**, 211–218 (2005). <https://doi.org/10.1007/s11936-005-0049-4>
3. Jason Bates, H.T., Young, M.P.: Applying fuzzy logic to medical decision making in the intensive care unit. *Am. J. Respirat. Crit. Care Med.* **167**(7), 948–952 (2003). <https://doi.org/10.1164/rccm.200207-777CP>
4. Phuong, N.H., Kreinovich, V.: Fuzzy logic and its applications in medicine. *Int. J. Med. Informat.* **62**(2–3), 165–173 (2001)
5. Da, F.: Fuzzy neural network sliding mode control for long delay time systems based on fuzzy prediction. *Neural Comput. Appl.* **17**, 531–539 (2008). <https://doi.org/10.1007/s00521-007-0130-x>
6. Engell, S., Heckenthaler, T.: Fuzzy control. In: Berber, R., (eds.), *Methods of Model Based Process Control. NATO ASI Series (Series E: Applied Sciences)*, vol. 293. Springer, Dordrecht (1995). [https://doi.org/10.1007/978-94-011-0135-6\\_29](https://doi.org/10.1007/978-94-011-0135-6_29)
7. Naranjo, J.E., Sotelo, M.A., Gonzalez, C., Garcia, R., Pedro, T.D.: Using fuzzy logic in automated vehicle control. *IEEE Intell. Syst.* **22**(1), 36–45 (2007). <https://doi.org/10.1109/MIS.2007.18>
8. Wen, Z.M., Zhou, S.D., Wang, M.: Fuzzy control for the obstacle avoidance of a quadrotor UAV. *Appl. Mech. Mater.* (2017). <https://doi.org/10.4028/www.scientific.net/AMM.775.307>
9. Zhang, Z., Hu, S.: Fuzzy control of attitude of four - rotor UAV. *Green Energy and Sustainable Development I: Proceedings of the International Conference on Green Energy and Sustainable Development.* (2017). <https://doi.org/10.1063/1.4992883>

10. Al-Hadithi, B.M., Matía, F., Jiménez, A.: Fuzzy controller for robot manipulators. In: Melin, P., Castillo, O., Aguilar, L.T., Kacprzyk, J., Pedrycz, W. (eds.), *Foundations of Fuzzy Logic and Soft Computing. IFSA 2007. Lecture Notes in Computer Science*, vol 4529. Springer, Berlin (2007). [https://doi.org/10.1007/978-3-540-72950-1\\_68](https://doi.org/10.1007/978-3-540-72950-1_68)
11. Amer, A.F., Sallam, E.A., Elawady, W.M.: A new adaptive fuzzy sliding mode control using fuzzy self-tuning for 3 DOF planar robot manipulators. *Appl. Intell.* **37**, 613 (2012). <https://doi.org/10.1007/s10489-011-0276-1>
12. Chen, C.H., Wang, C.C., Wang, Y.T., Wang, P.T.: Fuzzy logic controller design for intelligent robots. *Math. Probl. Eng.* (2017). <https://doi.org/10.1155/2017/8984713>
13. Van, M., Kang, H.J., Suh, Y.S.: A novel fuzzy second-order sliding mode observer-controller for a T-S fuzzy system with an application for robot control. *Int. J. Precis. Eng. Manuf.* **14**, 1703–1711 (2013). <https://doi.org/10.1007/s12541-013-0229-1>
14. Melin, P., Felix, C., Acosta, M.L.: Pattern recognition using fuzzy logic and neural networks. In: *Proceedings of the International Conference on Artificial Intelligence, IC-AI '03*, June 23–26, vol. 1. Las Vegas, Nevada (2003)
15. Novak, V., Hurtik P., Habiballa, H., Stepnicka, M.: Recognition of damaged letters based on mathematical fuzzy logic analysis. *J. Appl. Logic* **13**(2), part A, 94–104 (2015)
16. Ross, T.J.: *Fuzzy Logic with Engineering Applications*, 3rd edn., p. 607. University of New Mexico, Mexico (2010)
17. Janarthanan, R., Balamurali, R., Annapoorani, A., Vimala, V.: Prediction of rainfall using fuzzy logic. In: *Materials Today: Proceedings*, 2021, vol. 37, part 2, pp. 959–963 (2021)
18. Mohsen, R.A., Abbassi, B.: Prediction of greenhouse gas emissions from Ontario's solid waste landfills using fuzzy logic based model. *Waste Manage.* **102**, 743–750 (2020)
19. Kulagin, I.M., Emaletdinova, L.Y.: Functionality of the automated information system preparation documents for a Shipyard. In: *Proceedings of the International Conference Mathematical Methods in Technique and Technologies – MMTT*, vol. 6, p. 5 (2020)
20. Kruglov, V.V., Borisov, V.V.: *Artificial Neural Networks. Theory and Practice*, 2nd edn., Stereotype, M.: Hotline – Telecom, p. 382 (2002)
21. Kalyanathaya, K. P., Akila, D., Suseendren, G.: A fuzzy approach to approximate string matching for text retrieval in NLP. *J. Comput. Inf. Syst. USA* **15**(3), 26–32 (2019)

# Synthesis of a Neuro-Fuzzy Model in the Structure of an Automatic Rolling Stock Control System



Konstantin Yurenko , Pavel Kharchenko, Evgeniy Fandeev, and Ivan Yurenko

**Abstract** The problem of constructing a reference kinematic trajectory (RKT) of the movement of rolling stock on the example of the movement of a train is considered. The generated RKT is used in the circuit for regulating the travel time of the automatic control system when calculating the optimal control actions. The structure of a fuzzy neural network (FNN) is proposed, which includes two ANFIS networks that simulate the dependence of the values of the parameters of the optimal trajectory—the stabilization speed and the coordinate of the transition to the run-out during the movement along the stage on the specified travel time and mass of the train. The network training is based on the results of numerical solutions of the optimization problem of train driving modes. A numerical experiment was performed to compare the results of train motion modeling based on the RKT obtained by numerical methods and with the FNN data. Its results confirm the effectiveness of the use of this fuzzy neural network in the optimization of train traffic control in real-time. The simulation was performed in the MATLAB environment.

**Keywords** Rolling stock · Optimal control · Auto-driving · Simulation model · Fuzzy neural network · Reference kinematic trajectory

---

K. Yurenko (✉) · E. Fandeev  
Platov South-Russian State Polytechnic University (NPI), Novocherkassk, Russia  
e-mail: [ki-yurenko@yandex.ru](mailto:ki-yurenko@yandex.ru)

E. Fandeev  
e-mail: [eif@inbox.ru](mailto:eif@inbox.ru)

P. Kharchenko  
JSC Russian Railways, Traction Directorate, Ekaterinburg, Russia  
e-mail: [lrk-9@mail.ru](mailto:lrk-9@mail.ru)

I. Yurenko  
LLC Scientific and Technical Center CYBERINTELLS, Novocherkassk, Russia  
e-mail: [ikyur@yandex.ru](mailto:ikyur@yandex.ru)

## 1 Introduction

The intellectualization of control systems (CS) of complex technical objects and technological processes is currently one of the main directions for improving the quality of management, energy efficiency and reducing the risks associated with the human factor. As you know, a “complex” object is characterized by such features as the lack of a complete mathematical description, non-determinism (stochasticity), and non-stationarity [1, 2]. The presence of at least one of these features allows you to assign the control object (CU) to the category of "complex". With regard to complex dynamic CU, which include, in particular, railway rolling stock, the use of traditional methods of CS synthesis is associated with great difficulties and often does not provide the required quality of management. This applies not only to terminal systems, which are characterized by the task of transferring an object from the initial state to a given one in accordance with certain optimality criteria (control quality functional) but also to regulators that must perform robust control of complex non-stationary objects. In this case, developers often turn to modern information technologies for knowledge processing, which allows them to build intelligent ones. These technologies are characterized by the fact that they make it possible to work with formalized knowledge (expert systems, fuzzy logic), or use human-specific learning and thinking techniques: artificial neural networks and genetic algorithms [3–7].

To develop intelligent control systems, these methods should be combined with the achievements of modern control theory. Structurally, such CS contain additional blocks that perform system processing of knowledge based on the above information technologies. At the same time, they can be implemented as an add-on to a conventional CS or directly included in a closed control loop [2, 8–12]. Currently, intelligent control systems are used in many applied fields, in particular, in robotics, mechatronics, industry, and transport [13–15].

In the tasks of automating the control of rolling stock, in particular, the movement of trains, the use of intelligent control methods can be effective in various circuits of the CS. In [16] the problem of improving the automatic train operation (ATO) system based on the predictive fuzzy control method is considered. The authors of the article propose a method and device for automatic control of train brakes based on the use of a logical–linguistic model (LLM) using the Takagi–Sugeno fuzzy inference system. In [17], an improvement of the control system with pneumatic brakes of a freight train was performed, which uses the Mamdani fuzzy output algorithm to calculate the parameters of the relay controller. Another example, when fuzzy logic is used in the task of train braking control, is presented in the article [18]. One of the ways to improve the automatic train control system is to include intelligent functional blocks in its structure. Next will consider the problem of synthesizing a neuro-fuzzy model for computing the parameters of optimal train driving modes.

## 2 Structure of the Automatic Train Operation System (ATO)

The automatic train operation system (auto-driver) is designed to automate train movement control, including starting and accelerating, selecting the train driving mode on stages, braking when speed limits are exceeded, targeted braking at platforms, and collecting and processing data on train movement [19–21]. In the remote control system shown in Fig. 1, the internal loop controls the speed of movement, and the external loop controls the travel time. The latter is proposed to be built on the basis of LLM [22].

One of the options for implementing the travel time controller based on the LLM in the circuit of the automatic train driving system is shown in Fig. 1, where the following designations are adopted:  $t, s, v$ —travel time, path, speed;  $T_z$ —for-given travel time;  $t_{res}$ —travel time reserve;  $dist$ —distance to the end of the run;  $v_z$  and  $R$ —set speed and operating mode;  $u$ —control effect on the CS;  $P_u$ —train driving program, which is often called the “reference” trajectory in works on optimal traffic control. Next, we will use the term “reference kinematic trajectory” (RKT).

In general, the structure of the optimal RKT of the train moves along the stage has the form shown in Fig. 2, where  $v_s$  is the speed of stabilization;  $s_c$  is the coordinate of the transition to the run-out;  $s_r(v_r)$  is the coordinate or speed of the beginning of regenerative braking and includes the sections of acceleration, speed stabilization, run-out, regenerative and pneumatic braking. Depending on the speed constraints and the track profile, the optimal sequence of switching modes may differ from the one shown.

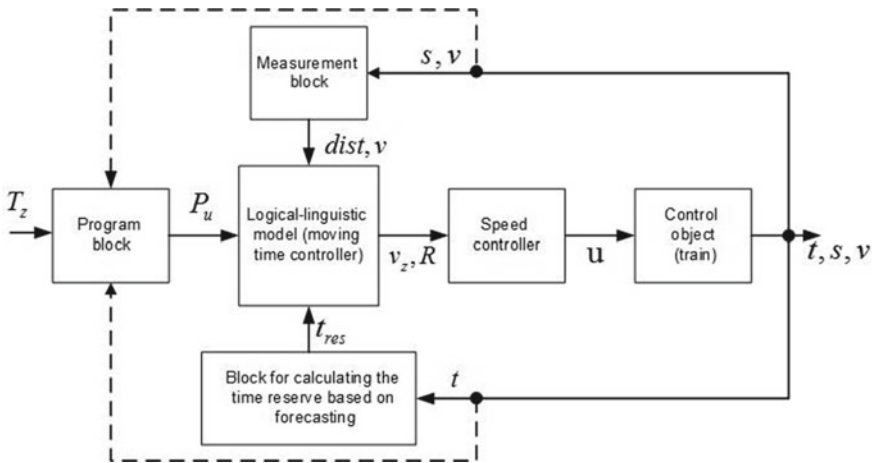


Fig. 1 Functional diagram of the automatic train driving system with LLM

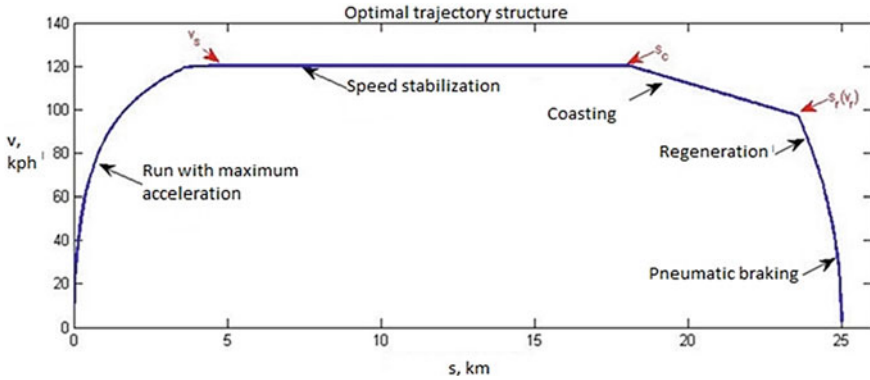


Fig. 2 Structure of the optimal reference kinematic trajectory of the train

### 3 Purpose of the Research

To construct an optimal RKT, in general, it is necessary to solve the problem of optimal control of the train movement. When presenting the model of the control object—the train, as an autonomous system (clearly independent of the time) it can be formalized as follows:

$$\left\{ \begin{array}{l} \frac{dt}{ds} = \frac{1}{v}; \\ \frac{dv}{ds} = \frac{1 + \gamma}{v} \left[ \frac{F_{\Sigma}(u) - W_{\Sigma}(v) - W_i(s) - W_a}{m_{\pi} + \sum m_B} \right]; \\ F_{\Sigma}(u)|_{u=(u_1, u_2, u_3)} = [F(u_1) + R(u_2) + B(u_3)]|_{u_1 \cdot u_2 \cdot u_3 \equiv 0}; \\ J(u) = A_e = \int_0^{T_x} \left( \frac{F(u_1)v}{\eta_T} - R(u_2)v\eta_r \right) dt |_{u_1 \cdot u_2 \equiv 0} \rightarrow \min, \end{array} \right.$$

where  $\gamma$  is the coefficient of inertia of the rotating parts;  $F_{\Sigma}$  is the total controlled force;  $F$ ,  $R$ ,  $B$  is the traction, regenerative and mechanical braking forces;  $W_{\Sigma}$ ,  $W_i$ ,  $W_a$ —the main, additional (from the slopes and curves of the track) and additional resistance to the movement of the train;  $\mathbf{u} = (u_1, u_2, u_3)$ —traction control, electrodynamic and mechanical brakes  $\eta_T$  and  $\eta_r$ —the efficiency of the electric locomotive in the traction mode and recovery;  $A_e$ —the total energy consumption for traction.

Various methods are known for constructing the optimal train trajectory [19, 21]. In [22], to find the  $v_s$  stabilization rate, a block for approximating the optimal  $v_s$  values determined by simulation modeling for the minimum (corresponding to the maximum permissible speed along the stretch) travel time  $T_{\min}$  and conditionally maximum  $T_{\max}$  is proposed. As the results of computational experiments have shown, this approach, combined with the use of LLM, can be used to control commuter trains

with small mass deviations from the nominal values and relatively short stretches. For passenger trains (and even more so for freight trains, when the number of cars and, consequently, the mass of which varies in a wide range), the  $v_s$  and  $s_c$  parameters for each stage significantly depend on both  $T_z$ , and the mass of the train.

These parameters can be determined using computational methods for solving the optimization problem, for example, [19–21], and the general nature of this dependence is revealed. In this case, the problem arises of constructing a model of such dependence for using the calculation of the control signals  $v_z$  and  $R$  in real-time onboard the locomotive.

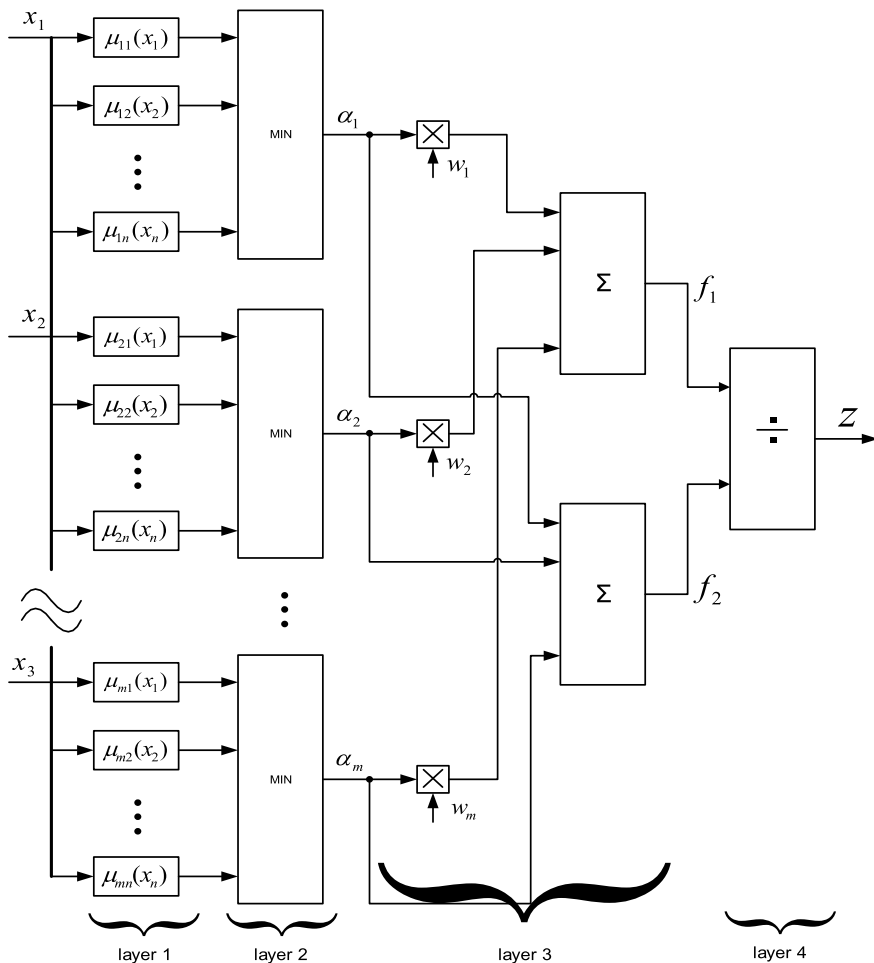
For this purpose, we propose to use the apparatus of FNN (Fuzzy Neural Networks), also known as Adaptive Neuro-Fuzzy Inference Systems (ANFIS) [2, 23].

## 4 Building a Fuzzy Neural Network

A fuzzy neural network is a multi-layer NN in which the layers function as elements of a fuzzy inference system. The neurons of such a network are characterized by a set of parameters, which are adjusted during the training process, as in ordinary FNN. On the one hand, the ANFIS hybrid network is an NN with a single output and several inputs that represent fuzzy linguistic variables, and on the other hand, it is a system of fuzzy inference of the Sugeno type of zero or first order, in which each of the rules of fuzzy productions has a constant weight equal to 1 [23].

In contrast to mathematical methods, which require precise or statistical regularities to be modeled at each step, fuzzy systems offer a completely different approach, thanks to which the model of optimal parameters of the train trajectory operates at a high level of abstraction and uses a minimum number of regularities. Thus, it becomes possible to obtain knowledge in the form of connections between the influencing factors and the parameters of the optimal trajectory, as a result of which it becomes possible to implement fuzzy rules based on neural networks. The advantages of the FNN over the conventional NN are that, in contrast to the conventional NN, a clear representation of the knowledge contained in the unclear rules is used, as well as the ability to establish a correspondence between the mathematical representation of the fuzzy inference procedure and the network structure. Figure 3 shows, as an example, the FNN based on the zero-order Sugeno algorithm [2].

Layer 1 performs fuzzification, non-linear functions  $\mu_{rj}(x_j)$ , where  $r$  is the number of the production rule,  $j$  is the number of the components of the input vector  $\bar{x}$  corresponding to the membership functions of the prerequisites of the rules. The configurable parameters of this layer are the parameters of the membership functions used. Layer 2 calculates the resulting membership functions of the premises of fuzzy rules. Layer 3, consisting of two neurons, performs summation and weighted summation of the output signals of layer 2. The parameters of this layer are the weight coefficients  $w_r$ .



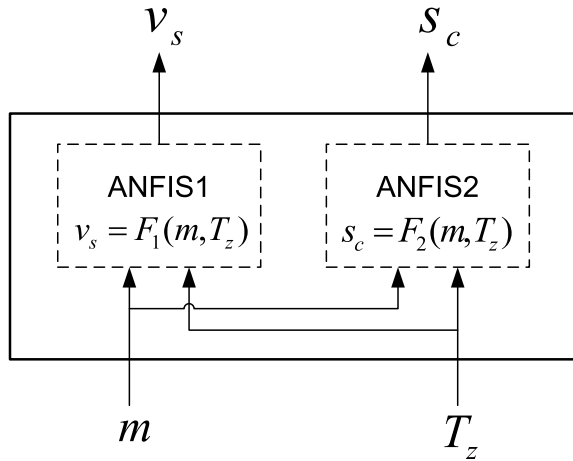
**Fig. 3** Fuzzy neural network based on the zero-order Sugeno algorithm

Layer 4 implements the division operation  $z = f_1/f_2$ . To model the dependence  $(v_s, s_c) = F(m, T_z)$ , it is proposed to use a structure consisting of two ANFIS networks, each of which models, respectively, the dependencies  $v_s = F_1(m, T_z)$  and  $s_c = F_2(m, T_z)$ , (Fig. 4).

When using the specified NNS to study the dependence of the parameters of the optimal trajectory of the train: the speed of stabilization  $v_s$  and the coordinates of the transition to the run-out  $s_c$  from the mass of the train  $m$  and the set travel time  $T_z$ , which allows you to draw clear conclusions in the conditions of fuzziness characteristic of the operation of railway rolling stock.



**Fig. 4** The structure of the FNN, which simulates the dependence of the parameters of the optimal trajectory of the train on its mass and a given travel time



When constructing the FNN, a training sample of data on the influencing factors and optimal values of the trajectory parameters is used to determine the parameters of the membership functions that best correspond to a certain system of fuzzy inference based on known NN training procedures (for example, the method of error backpropagation). The process of forming the ANFIS network structure involves choosing the type and number of membership functions (MF) for each influencing factor and the type of membership function at the output.

The development and research of the FNN were carried out using the tools of the Fuzzy Logic Toolbox package of the MATLAB “computer mathematics” environment—editors of hybrid systems ANFIS Editor (Fuzzy-Neuro Designer), fuzzy inference systems FIS Editor, Membership Function Editor functions, rules of the fuzzy inference system Rule Editor, as well as programs for viewing rules of fuzzy inference (Rule Viewer) and the surface of the fuzzy inference system (Surface Viewer).

Three types of data (samples) are used for the construction of the FNN: training data (Training) for the construction of a hybrid network; Testing data (Testing) for testing the built hybrid network in order to check the quality of its functioning, as well as Checking data (Checking) for the verification of the FNN in order to find out the fact of retraining. Such data were obtained preliminarily when solving the problem of optimizing the modes of train movement by numerical methods [20, 21].

As an example, such a problem was solved for a 20-km stretch for different train masses and given travel times along the track  $T_z$ . The sample sizes for training were 90 values of influencing factors ( $m$ —for 12, 14, 16, 18, and 20 cars in a train;  $T_z$ —with an interval of 2 min in the range from 12 to 20 min) and the corresponding parameters of the optimal trajectory  $v_s$  and  $s_c$ ; for testing—45 values for the same ranges of the initial data, and for testing—36 values for a train with 15 cars). A fragment of the initial data is presented in Table 1.

The formation of the FNN involves the following sequence of steps:

**Table 1** Sample fragments for training, testing, and verifying the ANFIS network

No.	Number of cars	Mass of train $m$ tons	Set travel time $T_z$ , min									
			12	13	14	15	16	17	18	19	20	
<i>Sample fragment for training and testing-values <math>v_s</math>, kph and <math>s_c</math>, km</i>												
1	12	840	117/12.8	105/13.6	90/12.6	90/13.6	82/14.8	81/13	80/12	80/11	80/10.4	
2	14	960	120/12.2	104/14.6	97/13.8	90/13.8	82/15	80/13.6	81/11.8	81/10.8	80/10.4	
3	16	1080	120/12.6	111/11.8	99/13.2	91/13.6	84/14.2	80/13.8	80/12.2	81/10.8	80/10.4	
4	18	1200	120/13	114/11	104/11.4	90/14.2	85/13.8	80/13.8	80/12.2	80/11.2	80/10.4	
5	20	1320	120/13.6	115/11	103/12	98/11.2	91/11.6	85/12	80/12.2	80/11.2	80/10.4	
<i>Sample fragment to check-values <math>v_s</math> and <math>s_c</math></i>												
6	15	1020	120/12.4	110/12	99/13	90/14	82/15.2	80/13.6	80/12.2	81/10.8	80/10.4	

1. Input the initial data (training, testing, and verification samples).
2. Select the type and number of MF at the input, as well as such functions at the output.
3. Setting the number of network training cycles  $N_L$  and selecting the FNN training method (when using the Fuzzy-Neuro Designer tool, you can choose between the error backpropagation method and the hybrid method, which is a combination of the least-squares method and the method of decreasing the back gradient).
4. FNN training. Evaluation of the quality of training by calculating the average square error  $E_{RMS}$  (root mean square error):

$$E_{RMS}^{v_s} = \sqrt{\frac{1}{n} \sum_{i=1}^n (v_s^{fact} - v_s^{sim})^2}; E_{RMS}^{s_c} = \sqrt{\frac{1}{n} \sum_{i=1}^n (s_c^{fact} - s_c^{sim})^2},$$

where  $n$  is the sample size;  $v_s^{fact}$  and  $s_c^{fact}$ —the actual values, and  $v_s^{sim}$  and  $s_c^{sim}$ —the values of the optimal motion parameters modeled using the NNS. The errors in the following steps 5 and 6 are determined using similar formulas.

5. Testing the network, calculating the error from the testing sample.
6. Checking the FNN, calculating the error from the test sample.
7. Based on the analysis of the received errors, it can be: (a) returned to steps 2 or 3 to select other network parameters and/or training cycles; (b) further configuration of the parameters of the built and trained network using the MF editors and the rules of the fuzzy inference system; (c) completion of training, settings, and transition to the network operation stage.

## 5 Results of Computational Experiments

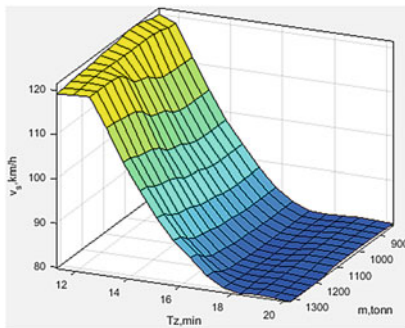
As a result of the above steps, two FNN—ANFIS1 and ANFIS2, were formed, trained, and configured to implement the structure shown in Fig. 4. The main results of the FNN design are presented in Table 2, and the response surfaces of the fuzzy inference systems corresponding to each of the networks are shown in Fig. 5.

As an example of the capabilities of the developed FNN for determining the parameters of the optimal trajectory of movement, we will compare the results of simulation modeling obtained using numerical optimization methods and generated at the output of the created FNN—Table 3. When conducting a computational experiment, the specified travel time  $T_z$ .

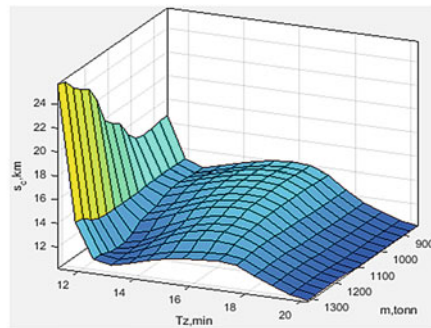
As can be seen from the table, the differences in the parameters of the travel time and energy consumption do not exceed 1%.

**Table 2** Results of design and quality assessment of ANFIS1 and ANFIS2 networks for modeling optimal train movement parameters

1	2	3		4	5
FNN	Quantity training cycles	Network error: training/monitoring/testing		MF type number of membership functions on the 1st and 2nd inputs	Type of output accessory functions
ANFIS1	300	$E_{RMS}^{v_s}$ , kph	0,95	“generalized bell”	Point map (set-singleton)
			1,17	5	
			1,33	5	
ANFIS2	360	$E_{RMS}^{s_c}$ , kph	0,43	“generalized bell”	Point map (set-singleton)
			0,47	5	
			0,62	5	



(a)



(b)

**Fig. 5** Response surfaces of fuzzy systems ANFIS1 (a) and ANFIS2 (b)

**Table 3** Comparison of the results obtained using the developed FNN and the numerical solution of the optimization problem of train movement

No.	Conditions of a computational experiment		Optimal trajectory parameters	FNN	Numerical optimization
1	Length of the track, km	20	Speed of Stabilization $v_s$ , kph	89,5	90
2	Set travel time $T_z$ , min	15	The coordinate of the transition to the run-out $s_c$ , km	13,7	13,4
3	Acceptable deviation from the schedule, sec	30	Actual travel time $t_x$ , sec	927	929
4	Train weight $m$ , tons (number of wagons)	1140 (17)	Power consumption for traction $A$ , kW·h	267,9	267,4

## 6 Conclusions

1. A method is proposed for modeling the dependence of the parameters of the optimal train trajectory—the speed of stabilization and the coordinate of the transition to the run-out on the mass of the train and the specified travel time along the stretch using the FNN, the structure of which includes two ANFIS networks trained on the basis of the results of numerical optimization of train driving modes.
2. The accuracy of the calculation of these dependencies achieved during training and network setup allows the use of the developed and trained FNN in the loop of the automatic train control system running time.
3. It is advisable to continue research in order to develop this approach for the cases of crossings with complex configurations of speed limits and track profile, as well as for freight trains, the mass of which can vary over a wide range.

## References

1. Rastrigin, L.A.: *Modern Principles of Management of Complex Objects*, p. 232. Moscow, Sov. Radio (1980)
2. Uskov, A.A., Kuzmin, A.V.: *Intelligent Control Technologies. Artificial Neural Networks and Fuzzy Logic*, p. 143. Moscow, Hotline-Telecom (2004)
3. Vizureanu, P.: *Advances in Expert Systems*, p. 128. InTeOp (2012)
4. Ross, T.J.: *Fuzzy Logic with Engineering Applications*, p. 580. Wiley (2016)
5. Mendel J.M.: *Uncertain Rule-Based Fuzzy Systems: Introduction and New Directions*. Springer, 699 p (2017)
6. Keller, J.M., Liu, D., Fogel, D.B.: *Fundamentals of Computational Intelligence. Neural Networks, Fuzzy Systems, and Evolutionary Computation*, p. 364. Wiley-IEEE Press (2016)
7. Simon, D.: *Evolutionary Optimization Algorithms. Biologically-Inspired and Population-Based Approaches to Computer Intelligence*, p. 784. Wiley (2013)
8. Hooda, D.S., Raich, V.: *Fuzzy Logic Models and Fuzzy Control. An Introduction*, p. 408. Oxford, U.K., Alpha Science International Ltd. (2017)
9. Mendel, J.M., Hagaras, H., Tan, W.-W., Melek, W.W., Ying, H.: *Introduction To Type-2 Fuzzy Logic Control: Theory and Applications*, p. 377. Wiley-IEEE Press (2014)
10. Kayacan, E., Khanesar, M.A.: *Fuzzy Neural Networks for Real Time Control Applications: Concepts, Modeling and Algorithms for Fast Learning*, p. 264. Butterworth-Heinemann (2015)
11. Siddique, N.: *Intelligent Control A Hybrid Approach Based on Fuzzy Logic, Neural Networks and Genetic Algorithms*, p. 292. Springer (2014)
12. Melin, P., Castillo, O.: *Design of Intelligent Systems Based on Fuzzy Logic, Neural Networks and Nature-Inspired Optimization*, p. 612. Springer (2015)
13. Behera, L., Kumar, S., Pataikani, K. P., Nair, R.R., Dutta, S.: *Intelligent Control of Robotic Systems*, p. 675. CRC Press (2020)
14. Szuster, M., Henzel, Z.: *Intelligent Optimal Adaptive Control for Mechatronic Systems*, p. 387. Springer (2018)
15. Kravets, A., Bolshakov, A., Shcherbakov, M.: *Cyber-Physical Systems: Modelling and Intelligent Control*, p. 358. Springer (2021)
16. Sandizadeh, M.A., Shamszadeh, B.: *Improvement of automatic train operation using enhanced predictive fuzzy control method*, pp. 121–140. In book: *Reliability and Safety in Railway*. InTech (2012)

17. Yurenko, K.I.: Automatic control of train brakes with fuzzy logic. In: Proceedings of the First International Scientific Conference Intelligent Information Technologies for Industry (IITI' 16), vol. 2, pp. 345–351. Springer (2016)
18. Yasunobu, S., Miamoto, S., Ihara, H.A.: Fuzzy control for train automatic stop control. *Trans. Soc. Instrum. Control. Eng.* **E-2**(1), 1–9 (2002)
19. Baranov, L.A., Erofeev, E.V., Meleshin, I.S., Chin, L.M.: Optimization of train traffic management. In: Baranov, A.L.A. (ed.) *Training Manual/Edited by Doctor of Technical Sciences*, MIIT, p. 164 (2011)
20. Yurenko, K.I., Fandeev, E.I.: Structure and functions of on-board autodriver system of train. In: 2nd International Conference on Industrial Engineering. Applications and Manufacturing, ICIEAM—Proceedings, p. 7911543 (2016)
21. Yurenko, K.I., Fandeev, E.I.: Classification systems of automatic train driving with positions of the modern automatic control theory. In: 2017 International Conference on Industrial Engineering. Applications and Manufacturing, ICIEAM 2017—Proceedings, p. 8076365 (2017)
22. Yurenko, K.I., Kharchenko, P.A., Fandeev, E.I., Yurenko, I.K.: Train operation control on-based of logical-linguistic model. *EAI Endorsed Trans. Energy Web* **6**(22), e4 (2019)
23. Jang, R.: *Neuro-Fuzzy Modeling: Modeling: Architectures, Analyses and Applications*. Ph.D. University of California. Department of Electrical Engineering and Computer Science, Berkeley (1992)

# **Predictive Modelling**

# Predicting the Equipment Useful Lifetime Based on the Deep Neural Networks



Maxim Dli, Andrey Puchkov, and Ekaterina Lobaneva

**Abstract** A method for predicting the useful time of equipment based on the processing of diagnostic data using parallel deep recurrent and convolutional neural networks is proposed. In the considered method, the recurrent neural network processes diagnostic data, which are presented in the form of time series. A convolutional neural network processes images obtained on the basis of continuous wavelet transform of diagnostic data. Neural networks implement a multivalued classification of two states of technological equipment—serviceable and faulty. The forecast of the values of the output signals of neural networks, performed on the basis of the recursive least squares method, is used to estimate the useful time of equipment. The results of a model experiment carried out using a program developed in the Matlab 2020b environment that implements the proposed method are presented. To generate training datasets, a transmission system model was used, implemented in the Simulink software package, and supplied as part of the Matlab 2020b environment. The architectures of recurrent and convolutional neural networks used in the experiment are presented. The wavelet transform of training datasets for a convolutional neural network was performed based on the analytical Morse wavelet. The experimental results showed that parallel-connected deep neural networks of various architectures increase the accuracy of estimating the useful time of equipment.

**Keywords** Technical diagnostics · Data predicting · Deep neural networks

---

M. Dli (✉) · A. Puchkov · E. Lobaneva  
National Research University “Moscow Power Engineering Institute” (Branch) in Smolensk,  
Energetichesky proyezd 1, Smolensk 2014013, Russia  
e-mail: [midli@mail.ru](mailto:midli@mail.ru)

A. Puchkov  
e-mail: [putchkov63@mail.ru](mailto:putchkov63@mail.ru)

E. Lobaneva  
e-mail: [lobaneva94@mail.ru](mailto:lobaneva94@mail.ru)



## 1 Introduction

The fourth industrial revolution, which has different names: Industrie 4. in Germany, Smart Manufacturing in the USA, and Smart Factory in South Korea [1], involves the transition to cyber-physical systems in which digital services cover all production and management processes. The use of Internet of Things (IoT) technologies contributes to the emergence of new concepts in the field of asset management, for example, APM 4.0 (Asset Performance Management, APM 4.0), Lean Smart Maintenance (LSM) [2, 3]. These concepts emphasize the implementation of proactive Predictive Maintenance, which aims to find and eliminate the failure cause even before the repair is required. This extends the life of the process equipment and turns maintenance from a cost to a source of income [4, 5].

The implementation of the noted concepts in existing production means the development of an integrated approach to equipment maintenance using available data, predictive analytics, and modeling to understand what factors actually affect the performance and reliability of assets. It should be noted that the introduction of industrial IoT leads to a sharp increase in the volume of data streams generated by production equipment, only 5% of which is processed [6]. This creates great potential for improving the quality of diagnostic procedures and the accuracy of predicting equipment failures, which can directly affect the profitability of production by preventing unscheduled downtime that causes the shutdown of production important areas.

One of the tasks for Predictive Maintenance is to assess the remaining useful life of equipment (RUL), the solution of which allows stopping the undesirable consequences of its failure in time. The RUL is also desirable to be found without switching to the test mode, that is, without disconnecting the equipment from its performing functions.

Statistical models, for example, [7, 8], give only averaged RUL values, and the specific value of this time for the available equipment is desirable to be known by the maintenance personnel.

A promising area for assessing RUL is machine learning methods belonging to the class of artificial intelligence methods and widely used in technical diagnostics. For example, in [9], it is proposed to use Elman artificial neural network to predict the RUL of bearings based on monitoring data of their condition. But the use of a relatively shallow neural network does not provide a sufficient level of revealing hidden regularities in big production data and requires additional costs for preliminary features extraction and filtering. In [10], a method based on an autoencoder and similarity measurement is proposed: an autoencoder trained on normal data is used to extract degradation curves of aircraft engines and build a library of degradation model patterns. Then, pattern curves for the object are obtained and similarities are analyzed to estimate RUL based on a sliding window. The disadvantage of this method is the need to obtain patterns in advance, which is not always possible due to the variety of faults. Approaches to the definition of RUL based on fuzzy logic [11] are characterized by its conceptual shortcomings—subjectivity and lack of adaptation to new data.

Many researchers see a promising and relevant direction for improving the accuracy of RUL estimation in the application of the Deep neural network (DNN) apparatus. One of the DNN varieties, called long short-term memory (LSTM) recurrent network, is widely used for time series analysis, which is most often the data from sensors installed on the equipment. Their ability to extract hidden regularities in large data sets is significantly superior to classical statistical methods, which allows them to make better predictive models [12–16].

The aim of the study was to improve the accuracy of the RUL estimation based on the processing of diagnostic data coming from control and measuring instrumentation, which will improve the efficiency of asset management, including by reducing the time of unscheduled downtime.

The problem of the study was to develop a method for processing diagnostic information providing accuracy improvement of predicting the equipment’s remaining useful life, as well as software that implements the proposed method.

## 2 Materials and Methods

The proposed RUL prediction method is based on bagging technology, in which independent machine learning models are created and their results are combined using a certain averaging model. Such approaches based on a combination of several machine learning models can improve the quality of the final solution and are implemented in various applied areas [17, 18]. The proposed structure for processing diagnostic information is shown in Fig. 1.

The input receives discrete data samples from the sensors  $X_i = X(i\Delta t)$ , where  $\Delta t$  is the discretization interval. In general,  $X_i$  is a vector of diagnostic data and the values of the discretization intervals can be different. In the DP (data preprocessing) block, the input data are normalized and standardized—centering and bringing the standard deviation to a single value. In the DPF block, sets of  $X_i$  values are formed for training neural networks.

The structure contains two channels for input data processing. In the first channel,  $X_i$  processing is carried out by the LSTM network, the choice of which is due to its higher representative power in comparison with classical regression models and networks with a simpler architecture [12].

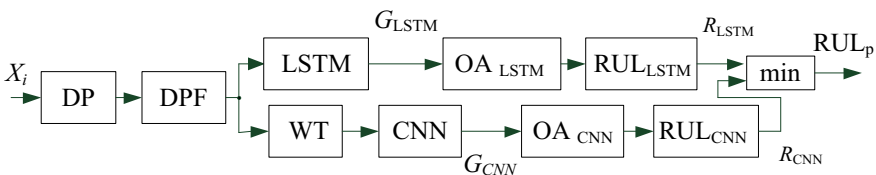


Fig. 1 Structure for processing diagnostic information

The second channel uses a convolutional neural network (CNN), which recognizes images obtained from the display of discrete samples entering the first channel. Since LSTM and CNN have different approaches to displaying the representation of input data, it is expected that this will allow revealing a larger set of hidden features and use this to obtain the final decision when predicting RUL. There are several approaches to image formation according to the time series data [19]. When choosing the type of display, it was taken into account that sensor signals often contain a high-frequency component recorded for a short period of time and a long-term low-frequency component. Windowed Fourier Transform allows analyzing either one or the other component, but not both at the same time. Therefore, to analyze both components, it is proposed to use a continuous wavelet transform that returns time–frequency representations (scalograms) which images are analyzed using CNN. This way of visualizing time series has a successful practice in machine learning problems [20, 21]. The application of the wavelet transform in the proposed structure is reflected in the WT (wavelet transform) block.

The output approximation of the LSTM and CNN neural networks is based on the recursive least squares (RLS) method and is performed in the OALSTM and OACNN blocks, respectively. In the RULLSTM and RULCNN blocks, the time before a failure occurrence is estimated based on the received RLS models in the OALSTM and OACNN blocks. The choice of the linear model is due to its simplicity and good accuracy at short time horizons.

Let  $G$  be the generalized designation for the DNN output (LSTM or CNN), which is responsible for the network confidence in the state belonging to the «operable» class. The linear model  $G$  can be represented as a superposition of orthogonal polynomials of the zero and first orders:

$$G(t) = G_0 + k(t - t_0) \quad (1)$$

where the parameters  $G_0$  and  $k$  characterize  $G$  at the time  $t = t_0$  and  $G$  rate of change, respectively. To estimate the parameters of the model (1)  $G_0$  and  $k$  and according to the data  $G_i$  arriving at discrete time  $t_i$  from the sensors installed on the equipment, the recurrent least squares method is used, which is actively applied in solving problems of estimation, control, and diagnostics [22–24]. To determine  $G_0$  and  $k$  the quadratic functional is minimized:

$$I = \sum_{i=1}^n [G_i - G_0 - k(t_i - t_0)]^2$$

from the condition of partial derivatives are equal to zero:  $\partial I / \partial \hat{G}_0 = 0$  and  $\partial I / \partial \hat{k} = 0$ . Skipping intermediate calculations, recurrent expressions for  $G_0$  and  $k$  (parameters estimate depend on the sample size at time  $t_i$ , which is reflected in square brackets) is written:

$$G_0^{[n]} = G_0^{[n-1]} + \frac{1}{n} \sum_{i=1}^n (G_i - G_0^{[n]}), \tag{2}$$

$$k^{[n]} = k^{[n-1]} + \frac{G_n - G_0^{[n-1]} - k^{[n-1]}(t_n - t_0^{[n-1]})}{\sum_{i=1}^n (t_i - t_0^{[n]})^2} (t_n - t_0^{[n]}), \tag{3}$$

where  $t_n$  is the time of data discrete under the number  $n$ .

Formulas (2) and (3) realize the recurrent correction of the parameters of the linear model as new data arrives from the sensors installed on the equipment. The position of the sampling center in (3) is calculated according to the recurrent expression:

$$t_0^{[n]} = t_0^{[n-1]} + (t_n - t_0^{[n-1]})/n.$$

The moment of failure occurrence is determined by the condition  $G_i = G_{Sh}$ , where  $G_{Sh}$  is the control boundary of the Shewhart chart (ISO 7870–2:2013 Controlcharts—Part 2: Shewhart control charts). Taking in to account (1)–(3) the recurrent formula for the RUL is obtained:

$$t_{Sh} = t_0 + \frac{G_{Sh} - G_0^{[n]}}{k^{[n]}}. \tag{4}$$

In blocks  $RUL_{LSTM}$  and  $RUL_{CNN}$  according to (4), the estimates of  $R_{LSTM}$  and  $R_{CNN}$  are calculated. In the simplest case, the final result of predicting  $RUL_p$  is realized by the operation of taking the minimum from  $R_{LSTM}$  and  $R_{CNN}$ . In more complex algorithms, if additional information about the diagnostic object is available, it is possible to use a weighted sum and elements of fuzzy inference.

### 3 Interpretation and Discussion of Research Results

Approbation of the RUL estimation method, which structure is shown in Fig. 1, was carried out in the Matlab 2020b environment. To generate training datasets, a transmission system model was used, implemented in Simulink, and supplied as part of the Matlab 2020b environment. When testing the structure shown in Fig. 1, the ShaftWear variable was used to create data to simulate shaft wear  $r(t)$ . The creation of a large number of examples for training neural networks was provided by changing  $r(t)$  with the presence of a sawtooth trend that simulates multiple shaft recovery and its subsequent wear. To obtain the time series  $X_i$ , a vibration sensor model was used that traces changes in the vibration of the transmission housing as a result of  $r(t)$  wear, therefore the analyzed diagnostic feature can be written as  $X = X(r(t))$ .

In the WT block, the time series  $X_i$  is transformed at the  $\Delta T$  interval into a scalogram image. In the Simulink model, the  $X_i$  data obtained from the vibration sensor in case a failure of a “shaft wear” type occurrence was simulated, the information indicators of which are the amplitude and frequency of the signal, so the decision was made to use an analytical Morse wavelet. It has proven itself well in analyzing signals with localized discontinuities and time-varying amplitude and frequency. The bank of continuous wavelet transforms filters CWT of the Wavelet Toolbox package was used. Determination of the decomposition levels sufficiency was carried out using the Shannon wavelet diagnostics [24]. To accelerate the calculation of scalograms in Matlab, an NVIDIA GeForce GTX 1650 4 Gb graphics processor installed on an ASUS TUF Gaming FX705DT laptop was used. Scalogram images in the form of tensors measuring  $224 \times 224 \times 3$  were fed to the CNN input.

LSTM and CNN networks in the structure in Fig. 1 are used for multivalued classification for two classes - “operational” and “nonoperational” states. The outputs of the networks, which characterize the confidence in the operational state, are approximated in the OALSTM and OACNN OALSTM and OACNN blocks. The change in this confidence is used to predict the RUL based on the approach described above. The architecture of the applied LSTM network is presented in Fig. 2, and the architecture of the CNN is shown in Fig. 3. The presented architectures are relatively simple and are created only to illustrate the efficiency of the proposed method.

The moments of intersection  $r(t)$  of the lower control boundary of the Shewhart chart were taken as the moment when a failure occurred and was considered the true RUL value, which was used to create training datasets for neural networks. In total, 30,000 pairs of data  $\{X_i, d_i\}$  were generated, where  $d_i$  is the class number corresponding to  $X(r(t_i))$ , with a breakdown into training (70% of the total number of examples), testing (20%) and validation (10%) sampling. The latter was used to select some hyperparameters of neural networks [25].

Previously, it was common to use two classes, so  $d_i$  takes two values—0 (nonoperational state) and 1 (operational state). Note that the G output also reflects the class, only the G values can continuously change in the range [0; 1]. To ensure the balance of the classes, the change in  $r(t)$  was modeled the way the number of examples of operational and nonoperational states was approximately equal.

	Name	Type	Activations	Learnables
1	sequenceinput Sequence input with 1 dimensions	Sequence Input	1	-
2	lstm LSTM with 228 hidden units	LSTM	228	InputWeights 912×1 RecurrentWe... 912×... Bias 912×1
3	fc 2 fully connected layer	Fully Connected	2	Weights 2×228 Bias 2×1
4	softmax softmax	Softmax	2	-
5	classoutput crossentropyex	Classification Output	-	-

Fig. 2 LSTM network layers

	Name	Type	Activations	Learnables
1	imageinput 224x224x3 images with 'zerocenter'...	Image Input	224x224x3	-
2	conv_1 8 3x3x3 convolutions with stride [1 1...	Convolution	224x224x8	Weights 3x3x3x8 Bias 1x1x8
3	batchnorm_1 Batch normalization with 8 channels	Batch Normalization	224x224x8	Offset 1x1x8 Scale 1x1x8
4	relu_1 ReLU	ReLU	224x224x8	-
5	maxpool_1 2x2 max pooling with stride [2 2] an...	Max Pooling	112x112x8	-
6	conv_2 16 3x3x8 convolutions with stride [1 ...	Convolution	112x112x16	Weights 3x3x8x16 Bias 1x1x16
7	batchnorm_2 Batch normalization with 16 channels	Batch Normalization	112x112x16	Offset 1x1x16 Scale 1x1x16
8	relu_2 ReLU	ReLU	112x112x16	-
9	maxpool_2 2x2 max pooling with stride [2 2] an...	Max Pooling	56x56x16	-
10	conv_3 32 3x3x16 convolutions with stride [...	Convolution	56x56x32	Weights 3x3x16x32 Bias 1x1x32
11	batchnorm_3 Batch normalization with 32 channels	Batch Normalization	56x56x32	Offset 1x1x32 Scale 1x1x32
12	relu_3 ReLU	ReLU	56x56x32	-
13	fc 2 fully connected layer	Fully Connected	1x1x2	Weights 2x100352 Bias 2x1
14	softmax softmax	Softmax	1x1x2	-
15	classoutput crossentropyx	Classification Output	-	-

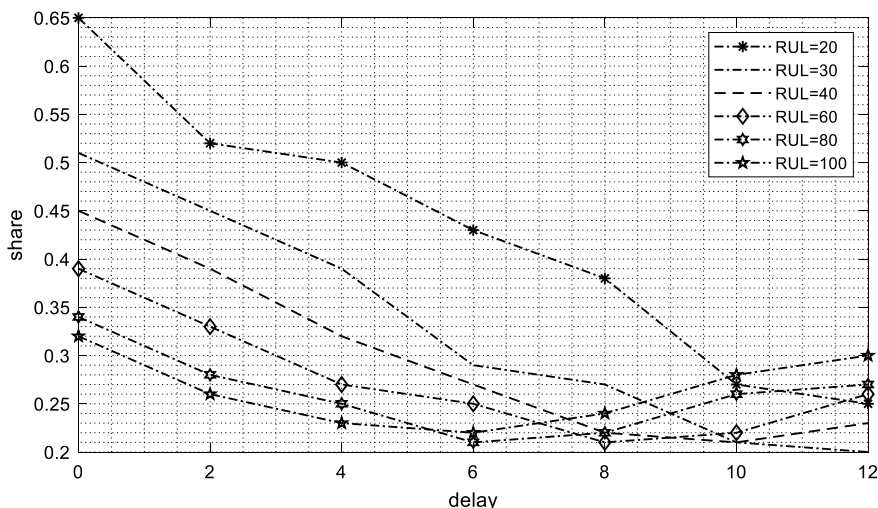
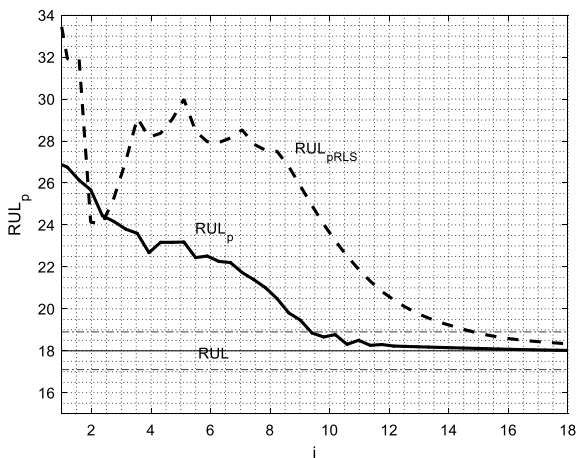
Fig. 3 CNN layers

DNN training was carried out for several values of the delay parameter, which sets the number of intervals for which the state prediction is made. As a result, several networks were created, corresponding to different delays, which will be used in a model experiment later. The networks were trained over 25 epochs, while the achieved accuracy for LSTM and CNN depended on the delay value, but the minimum value of the accuracy metric did not fall below 87%.

For comparison, Fig. 4 shows an example of applying the RLS method directly to the Xi time series (dashed line RULpRLS) and estimating RUL according to the proposed method (solid line RULpRLS) with delay = 0 (that is, when classifying the current state). The solid gray line in Fig. 4 denotes a five percent zone from the true RUL time, which composed 18 intervals from the start of the RUL estimation method application. It can be seen from the figure that the RULp line enters the zone of the indicated accuracy much earlier compared with RULpRLS, which is typical for all executed working program runs with delay = 0.

Studies of the delay parameter influence on the share of timeT5%entering the 5% zone in relation to the true value of RUL: share = T5%/ RUL have been carried out. The delay and T5%values are reflected in the number of discretization intervals. Figure 5 presents the results of experiments showing that the use of (4) for the predicted values of the state for the diagnostic object can reduce the time T5%.

**Fig. 4** Results of the RUL estimation method application



**Fig. 5** Prediction influence on the RUL estimation

This can be explained by the use of good predictive capabilities of DNN when estimating RUL in the considered method. At longer prediction intervals (delay >10 ÷ 12) for the used DNN architectures, an increase in the accuracy metric was observed; therefore, experiments on calculating RUL<sub>p</sub> were not carried out. However, the results obtained can serve as a justification for the operability of the proposed structure for processing diagnostic information to estimate the equipment's useful lifetime and the feasibility of additional research to develop recommendations for its perfection.

## 4 Conclusion

The presented method of processing diagnostic data for predicting the equipment's useful lifetime is based on the high generalization performance of DNN. Numerical experiments carried out on the model of the transmission system supplied as a part of the Matlab 2020b environment have shown the efficiency of the proposed method and its high speed of entry into the zone of a given accuracy, in comparison with the RLS method applied directly to the sensor data (without its preliminary processing in DNN) installed on the equipment.

The proposed method for estimating the equipment's useful lifetime can be used in equipment monitoring systems and decision support systems to reduce the time to achieve a given predictive accuracy.

**Acknowledgements** The reported study was funded by RFBR according to the research projects No 20-37-90062 «Postgraduates».

## References

1. Wuest, T., Weimer, D., Irgens, C., Thoben, K.D.: Machine learning in manufacturing: advantages, challenges, and applications. *Prod. Manuf. Res.* **4**(1), 23–45 (2016). <https://doi.org/10.1080/21693277.2016.1192517>
2. Biedermann, H., Kinz, A.: Lean Smart Maintenance—Value Adding, Flexible, and Intelligent Asset Management. *Berg Huetttenmaenn Monatsh* **164**, 13–18 (2019). <https://doi.org/10.1007/s00501-018-0805-x>
3. Parida, A., Stenström, C.: Dynamic asset performance management. In: Misra, K.B. (ed.) *Handbook of Advanced Performability Engineering*. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-55732-4\\_18](https://doi.org/10.1007/978-3-030-55732-4_18)
4. Adu-Amankwa, K., Attia, A.K., Janardhanan, M.N., et al.: A predictive maintenance cost model for CNC SMEs in the era of industry 4.0. *Int. J. Adv. Manuf. Technol.* **104**, 3567–3587 (2019). <https://doi.org/10.1007/s00170-019-04094-2>
5. Lee, S.M., Lee, D., Kim, Y.S.: The quality management ecosystem for predictive maintenance in the industry 4.0 era. *Int. J. Qual. Innov.* **5**, 4 (2019). <https://doi.org/10.1186/s40887-019-0029-5>
6. Sokolov, D. I., Solovyov, S.Y.: Control and monitoring of industrial equipment using the Mind-Sphere platform of Siemens. *Informatiz. Control. Syst. Ind.* **4**(76), 57–62. <https://isup.ru/articles/2/13235/> (2018)
7. Xie, G., Li, X., Peng, X., et al.: Estimating the probability density function of remaining useful life for wiener degradation process with uncertain parameters. *Int. J. Control Autom. Syst.* **17**, 2734–2745 (2019). <https://doi.org/10.1007/s12555-018-0558-z>
8. Mudunuru, V.R., Komarraju, S.: Prediction of remaining useful life of an end mill using ANSYS. In: Raju, K., Senkerik, R., Lanka, S., Rajagopal, V. (eds.) *Data Engineering and Communication Technology. Advances in Intelligent Systems and Computing*, vol. 1079. Springer, Singapore (2020). [https://doi.org/10.1007/978-981-15-1097-7\\_49](https://doi.org/10.1007/978-981-15-1097-7_49)
9. Liu, C., Zhang, C.: An elman artificial neural network for remaining useful life prediction. In: Li, Q.L., Wang, J., Yu, H.B. (eds.) *Stochastic Models in Reliability, Network Security and System Safety. JHC80 2019. Communications in Computer and Information Science*, vol. 1102. Springer, Singapore (2019). [https://doi.org/10.1007/978-981-15-0864-6\\_8](https://doi.org/10.1007/978-981-15-0864-6_8)



10. Wang, M., Li, Y., Zhao, H., Zhang, Y.: Combining autoencoder with similarity measurement for aircraft engine remaining useful life estimation. In: Jing, Z. (ed.) Proceedings of the International Conference on Aerospace System Science and Engineering 2019. ICASSE 2019. Lecture Notes in Electrical Engineering, vol. 622. Springer, Singapore (2020). [https://doi.org/10.1007/978-981-15-1773-0\\_14](https://doi.org/10.1007/978-981-15-1773-0_14)
11. Witczak, M., Lipiec, B., Mrugalski, M., Stetter, R.: A fuzzy logic approach to remaining useful life estimation of ball bearings. In: Bartoszewicz, A., Kabziński, J., Kacprzyk, J. (eds.) Advanced, Contemporary Control. Advances in Intelligent Systems and Computing, vol. 1196. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-50936-1\\_117](https://doi.org/10.1007/978-3-030-50936-1_117)
12. Dixon, M., Klabjan, D., Bang, J.H.: Implementing deep neural networks for financial market prediction on the intel xeon phi. In: Proceedings of the 8th Workshop on High-Performance Computational Finance, p. 1–6 (2015). <https://doi.org/10.1145/2830556.2830562>
13. Becerra-Rico, J., Aceves-Fernández, M.A., Esquivel-Escalante, K., et al.: Airborne particle pollution predictive model using Gated Recurrent Unit (GRU) deep neural networks. Earth Sci. Inform. **13**, 821–834 (2020). <https://doi.org/10.1007/s12145-020-00462-9>
14. Zhou, J.T., Zhao, X., Gao, J.: Tool remaining useful life prediction method based on LSTM under variable working conditions. Int. J. Adv. Manuf. Technol. **104**, 4715–4726 (2019). <https://doi.org/10.1007/s00170-019-04349-y>
15. Meshalkin, V., Puchkov, A., Dli, M., Lobaneva, Y.: Deep neural networks application in models with complex technological objects. In: Kravets, A., Bolshakov, A., Shcherbakov, M. (eds.) Cyber-Physical Systems: Advances in Design & Modelling. Studies in Systems, Decision and Control, vol. 259. Springer, Cham (2020). [https://doi.org/10.1007/978-3-030-32579-4\\_23](https://doi.org/10.1007/978-3-030-32579-4_23)
16. Schmid, P.A.E., Steinecker, A., Sun, J., Knapp, H.F.: Neural networks and advanced algorithms for intelligent monitoring in industry. In: Langheim, J. (eds.) Electronic Components and Systems for Automotive Applications. Lecture Notes in Mobility. Springer, Cham (2019). [https://doi.org/10.1007/978-3-030-14156-1\\_14](https://doi.org/10.1007/978-3-030-14156-1_14)
17. Livieris, I.E., Iliadis, L., Pintelas, P.: On ensemble techniques of weight-constrained neural networks. Evol. Syst. **12**, 155–167 (2021). <https://doi.org/10.1007/s12530-019-09324-2>
18. de Almeida, R., Goh, Y.M., Monfared, R., et al.: An ensemble based on neural networks with random weights for online data stream regression. Soft. Comput. **24**, 9835–9855 (2020). <https://doi.org/10.1007/s00500-019-04499-x>
19. Dli, M. I., Lobaneva, E.I., Puchkov, A.Y.: Algorithms for the formation of images of the states of objects for their analysis by deep neural networks. J. Appl. Inform. **V2** (80), 43–55 (2019)
20. Chen, Z., Wang, Y., Wu, J., et al.: Sensor data-driven structural damage detection based on deep convolutional neural networks and continuous wavelet transform. Appl. Intell. (2021). <https://doi.org/10.1007/s10489-020-02092-6>
21. Liang, M., Cao, P., Tang, J.: Rolling bearing fault diagnosis based on feature fusion with parallel convolutional neural network. Int. J. Adv. Manuf. Technol. **112**, 819–831 (2021). <https://doi.org/10.1007/s00170-020-06401-8>
22. Lee, C.Y., Hwang, S.H., Nam, E., et al.: Identification of mass and sliding friction parameters of machine tool feed drive using recursive least squares method. Int. J. Adv. Manuf. Technol. **109**, 2831–2844 (2020). <https://doi.org/10.1007/s00170-020-05858-x>
23. Oh, K., Seo, J.: Recursive least squares based sliding mode approach for position control of DC motors with self-tuning rule. J. Mech. Sci. Technol. **34**, 5223–5237 (2020). <https://doi.org/10.1007/s12206-020-1124-1>
24. Gangsar, P., Tiwari, R.: Diagnostics of mechanical and electrical faults in induction motors using wavelet-based features of vibration and current through support vector machine algorithms for various operating conditions. J. Braz. Soc. Mech. Sci. Eng. **41**, 71 (2019). <https://doi.org/10.1007/s40430-019-1574-5>
25. Dli, M.I., Puchkov, A.Y., Lobaneva, E.I.: Analysis of the influence of the architecture of the input layers of convolution and subsampling of a deep neural network on the quality of image recognition. J. Appl. Inform. **VI**(15), 113–122 (2020). <https://doi.org/10.24411/1993-8314-2020-10008>

# Operational Model Predictive Control on the Example of the Stabilization Process for Hydro Treatment of Oil Fractions



I. V. Gogol, N. A. Kalashnikov, O. A. Remizova, V. V. Syrokvashin, and A. L. Fokin

**Abstract** A method is proposed for solving operational control problems by predicting the behavior of a model within the framework of the MPC approach, which is an algorithm for constructing using the discrete maximum principle based on statistical mathematical models. During the study, for the parametric identification of the object, the weighted least squares method with an orthogonal decomposition of the information matrix was used, which allows you to build a regression mathematical model of the process, which is updated in real-time. A simulation model is also built, based on which the proposed optimal control strategy is compared with the strategy for which the experimental sample was obtained. As an example of the implementation of the proposed technique, a technological object was considered—the process of stabilization of the hydrogenated product, which takes place in a stripping column and is one of the most important processes of oil refining and, in particular, a hydrotreating unit of a low-temperature isomerization unit. As a result of this process, the stream is not only cleaned of unnecessary components by separating from it dissolved light hydrocarbons, water, and hydrogen sulfide, but also does not poison the installations of subsequent processing stages. This object has many limitations associated with the mode of the technological process, as well as with the parametric uncertainty of the object under study. As the goal of management, the task was set to minimize energy consumption.

**Keywords** MPC · Combined control system · Robust system · System robustness · Perturbation

## 1 Introduction

When controlling a technological process in a normal mode, the main technological variables: flow rate, temperature, pressure, viscosity, etc., vary within certain limits and take interval values. Therefore, for some processes, the search for the optimal,

---

I. V. Gogol · N. A. Kalashnikov · O. A. Remizova (✉) · V. V. Syrokvashin · A. L. Fokin  
Saint-Petersburg State Institute of Technology, Saint-Petersburg, Russia  
e-mail: [remizova-oa@technolog.edu.ru](mailto:remizova-oa@technolog.edu.ru)

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022  
A. G. Kravets et al. (eds.), *Cyber-Physical Systems: Intelligent Models and Algorithms*, Studies in Systems, Decision and Control 417,  
[https://doi.org/10.1007/978-3-030-95116-0\\_12](https://doi.org/10.1007/978-3-030-95116-0_12)

145

in a sense, values of these variables following the selected objective function at each moment and the provision of appropriate assignments of these variables can provide an additional economic effect for existing production. This provision is the basis of the proposed operational management design techniques.

The object of optimization can be a single aggregate or several related aggregates. The choice of the criterion can be associated either with a decrease in energy or other losses or with an increase in productivity. Moreover, the criterion can be selected locally on the technological scheme and represent, for example, the temperature of any flow, which can be minimized within normal values, and thus minimize costs. This can be the flow rate of the output or intermediate product, which can be maximized within the limits to increase productivity.

Such a problem can be solved at each discrete moment in time following the time interval of polling the sensors or only at some selected moments in time. For optimization according to the selected criterion, at every moment it is necessary to have a system of constraints in the form of equalities and inequalities. Inequality constraints on technological variables are obtained automatically in the form of intervals based on past statistical data that are assumed to be known.

As constraints of the equality type, equations are usually considered that link technological variables, which for each modelled block are divided into input and output. The communication model between them can be obtained in various ways. In this work, a statistical approach is chosen. Here, the main condition is the availability of experimental information obtained in modernized production.

In the chapter, the system of the lower level of the DCS hierarchy, that is, the technological process together with local stabilization systems is considered as a modelled object of operational control. For the normal mode, it is rather difficult to estimate the changing parameters of the object being stabilized, since the system can periodically be in a steady state. Then the identification problem can degenerate since for parametric identification it is necessary to excite all modes of the dynamic object. This raises the problem of choosing an identification method to obtain a statistical model.

In works [1–5] a linear multivariate model with variable parameters, which are estimated in real-time, is used. For normal operation, the linear model is justified and is traditionally used in automation. For parametric identification, we use the weighted least squares method [6], which involves forgetting outdated information that is present in the information matrix.

For the normal mode, the problem is that information about an object can have a repetitive nature, and its use leads to a degeneration of the identification procedure [7] since all information is forgotten. In this work, we used the least-squares method with an orthogonal decomposition of the information matrix [8], which ensures forgetting only information that is repeated in the incoming new data, which ensures non-degeneracy.

Based on the obtained models, equality-type constraints are formed, and together with the known inequality-type constraints and the chosen objective function, we obtain an optimal control problem for which it is required to choose a solution

method. In works [1–5] used: linear programming and the maximum principle. This work assumes the use of the well-known Model Predictive Control (MPC) strategy.

The peculiarity is that MPC control is usually used to build local control for models with complex dynamics. In this case, it is used at the level of operational management by the example of the technological process of stabilization.

The advantage of the proposed technique is that it applies to varying degrees of efficiency in almost all cases. Efficiency depends on specific conditions, such as information support, the size of the intervals of change of variables, the choice of the objective function, the fulfillment of the conditions for the existence of a solution, for example, for linear programming problems, etc.

## 2 Formulation of the Problem

The stabilization process [9, 10] takes place in a stripping column K-1. The unstable hydrogenated product through a heat exchanger, where it is heated in the shell space, is fed to the stripping column on the twelfth tray (a total of 20 valve trays). Here hydrocarbon gases, hydrogen sulfide, ammonia, and water are stripped from the unstable hydrogenation product. The upper light product of the stripping column, after cooling and condensation in air and water coolers enters the separator, where the upper product is separated into crude hydrocarbon gas and light gasoline and light gasoline is separated from the water. Light gasoline from the separator is pumped to the twentieth tray of column K-1 as reflux, and the balance quantity is output to the primary distillation unit.

The output product—stable hydrogenated product from the bottom of the K-1 column taken by pumps and returned to the column through the P-1 furnace as a “hot stream”. The P-1 furnace is designed to heat the stable hydrogenate of the K-1 column cube. Also, the bottom product of the stripping column (stable hydrogenate) is removed from the unit. Heat is supplied to the bottom of the stripping column by circulating the bottom product of the column through a heating furnace.

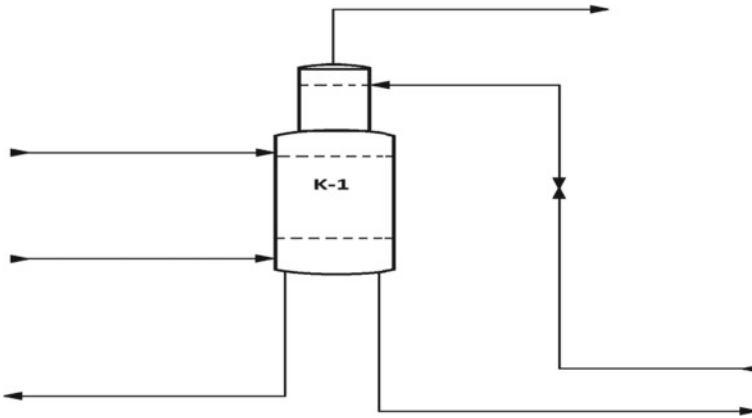
Corresponding temperatures, pressures, and flow rates of column K-1 are shown in Fig. 1.

From the experimental data, the upper and lower limits for the available measured variables were determined:

$$\begin{aligned} 394.4 \leq x_1 \leq 460.6, \quad 529.7 \leq x_2 \leq 543, \quad 97 \leq x_3 \leq 104.6, \quad 299 \leq u_1 \leq 306, \\ 59.5 \leq u_2 \leq 77.5, \quad 44 \leq u_3 \leq 59, \quad 236 \leq v \leq 243. \end{aligned} \quad (1)$$

$x_1, u_2$  flow rates are measured in  $\text{m}^3/\text{h}$ ,  $u_1, x_3$ , temperatures are measured in C,  $x_2$  pressure is measured in  $\text{kgf}/\text{cm}^2$ . The  $u_3$  level is measured in percent.

The experimental data are received with a discreteness step equal to 10 min. The number of points considered corresponding to the normal mode is 2000 (333.33 h).



**Fig. 1** Flow chart and studied variables

It is assumed that two optimal problems are solved: ensuring the highest value of the flow rate of the output product  $x_1$  to increase productivity or ensuring the lowest temperature value at the top of the column  $x_3$  to reduce costs. To solve these problems, three independent control actions are used: the temperature at the bottom of the column  $u_1$ , the flow rate of the recycling product  $u_2$ , level of the bottom product  $u_3$ .

These controls are tasks for the respective local control systems. So the  $u_1$  value is realized by stabilizing the temperature in the P-1 furnace, the  $u_2$  value using the flow stabilization regulator, the  $u_3$  value using the level regulator.

In addition to the listed variables, the following are used: one more measured variable  $x_2$ —pressure at the bottom of the column and  $f$ —the temperature of the unstable hydrogenate, which in this problem is considered a disturbance.

As a model of the dynamics of an operational control object, three standard models of the form:

$$x_i(t + 1) = \varphi_i^T(t)\theta_i, \quad i = 1, 2, 3, \quad t = 0, 1, 2, \dots \quad (2)$$

where  $\theta_i$ —vectors-columns of tunable parameters that change at each step  $t$ ,  $\varphi_i(t)$ —regressors that depend on the measured variables.

The following vectors were taken as regressors:

$$\begin{aligned} \varphi_1(t) &= [1 \ x_1(t) \ f(t) \ u_1(t) \ u_2(t) \ x_2(t) \ x_3(t) \ u_3(t)]^T \\ \varphi_2(t) &= [1 \ x_2(t) \ f(t) \ u_1(t) \ u_2(t) \ x_1(t) \ x_3(t) \ u_3(t)]^T \\ \varphi_3(t) &= [1 \ x_3(t) \ f(t) \ u_1(t) \ u_2(t) \ x_1(t) \ x_2(t) \ u_3(t)]^T \end{aligned} \quad (3)$$

Adjustable parameters  $\theta_i$   $i = 1, 2, 3$  are calculated recursively during identification. Let us introduce the following notation for them:

$$\begin{aligned}
 \theta_1 &= [d_1 \ a_{11} \ v_1 \ b_{11} \ b_{12} \ a_{12} \ a_{13} \ b_{13}]^T \\
 \theta_2 &= [d_2 \ a_{22} \ v_2 \ b_{21} \ b_{22} \ a_{21} \ a_{23} \ b_{23}]^T \\
 \theta_3 &= [d_3 \ a_{33} \ v_3 \ b_{31} \ b_{32} \ a_{31} \ a_{32} \ b_{33}]^T
 \end{aligned} \tag{4}$$

Then the system of equations of state can be written in the form:

$$x(t+1) = A(t)x(t) + B(t)u(t) + d + v(t)f(t), \tag{5}$$

where  $A(t) = \{a_{ij}(t)\}_{i,j=1}^3$ ,  $B(t) = \{b_{ij}(t)\}_{i,j=1}^3$ ,  $d(t) = [d_1(t) \ d_2(t) \ d_3(t)]^T$ ,  
 $v(t) = [v_1(t) \ v_2(t) \ v_3(t)]^T$ .

Equation (5) is used to solve optimization problems. Formulas (1)–(4) define constraints of the type of inequalities and equalities that are checked at each instant of discrete-time. The functional to be minimized is given as a linear form:

$$J(t) = c^T x(t+h), \tag{6}$$

where  $x(t) = [x_1(t) \ x_2(t) \ x_3(t)]^T$ —a state vector,  $h$ —a forecast horizon,  $c^T = [c_1 \ c_2 \ c_3]$ —a row vector of weighting factors, which is specified by the developer.

For the first task—increasing productivity  $x_1$ , the weighting factors are set in the form  $c^T = [-1 \ 0 \ 0]$ . For the second task—reducing costs by reducing  $x_3$ , we get  $c^T = [0 \ 0 \ 1]$ . In addition to these tasks, intermediate options are possible when  $c^T = [c_1 \ c_2 \ c_3]$  and  $\sum c_j = 1$ ,  $j = 1, 2, 3$ .

Thus, the procedure for obtaining an optimal solution at a working facility consists of the following stages:

1. Regressors are formed at each moment (3).
2. Obtaining estimates of the vector of parameters  $\theta$  with components (4) at each instant of discrete-time and the formation of constraints.
3. Solution of the optimal control problem with minimization of functional (6) either at each moment or at certain specified moments.

Besides, it is possible to set the task of retrospective improvement of the mode in which the experimental data were taken, due to the described optimal solution by modelling. This task can be considered either as the main one or as an auxiliary one for debugging before real work at a technological facility.

To solve such a problem, a simulation model of the technological process covered by the feedbacks of local automation is needed. For the normal mode, which occurs in a fairly narrow range of variation of variables, the simulation model can be built based on the regression model (2) in the form of an array of previously obtained parameter estimates (4) over the entire range of changes in the available experimental data. When substituting new values of controls, which are obtained when solving

the problem of optimal operational control at selected times, and disturbances into a model (2), it simulates the behavior.

But when obtaining optimal controls, this simulation model, like the object, is considered unknown to the developer, therefore, to obtain constraints of the equality type, it is required to build another regression model by the values of the variables, which are now removed from the simulation model during the control process.

### 3 Regression Model

As noted above, for statistical modelling of a process in normal mode, the main problem is sluggish changes in process variables, which leads to the degeneration of the least-squares problem due to the degeneration of its information matrix. The solution is sought in the class of recurrent algorithms with variable tunable parameters.

Then the recurrent form [6] has the form:

$$\begin{aligned} \hat{\theta}_i(t) = & \hat{\theta}_i(t-1) + R_i^{-1}(t)\phi_i(t) \left[ x_i(t) - \phi_i^T(t)\hat{\theta}_i(t-1) \right], \\ & i = 1, 2, 3, \quad t = 0, 1, 2, \dots \end{aligned} \quad (7)$$

where  $\hat{\theta}_i(t)$ —the estimate of the vector  $\theta_i$  at the moment  $t$ ,  $R_i(t)$ —information matrix.

The information matrix is calculated using the recurrent formula:

$$R_i(t) = \lambda_i R_i(t-1) + \phi_i(t)\phi_i^T(t), \quad i = 1, 2, 3 \quad (8)$$

where  $\lambda_i$ —the forgetting parameter, usually  $0.98 \leq \lambda_i \leq 0.995$ , but it can be less.

The information matrix contains all information about the identification procedure. At the current time, new information is added as a peer-to-peer matrix  $\phi_i(t)\phi_i^T(t)$ , which is formed based on the measured  $\phi_i^T(t)$  regressor.

The initial value of the information matrix is formed at time  $t_0 \geq m$  in the form

$$R_i(t_0) = \sum_{k=1}^{t_0} \lambda_i^{t_0-k} \phi_i(k)\phi_i^T(k), \quad i = 1, 2, 3 \quad (9)$$

It is known [7] that for the least-squares algorithm to function correctly, the positive definite matrix  $R_i(t)$  requires the fulfilment of the inequalities

$$\bar{c}_1 I < R_i(t) < \bar{c}_2 I, \quad (10)$$

где  $\bar{c}_2 > \bar{c}_1 > 0$ .

The main problem with identification in normal mode is that these inequalities are not met. The reason is the uniformity of the information received, which is associated with the operation of the stabilization systems of the lower level. In formula (8), the forgetting factor acts on the full rank matrix, which is the information matrix  $R_i(t)$ . Therefore, all information is equally forgotten. One that can be restored (recurring), and one for which it is impossible to do it. This is the reason for the degeneration of the algorithm.

Therefore, it is better to use the least-squares directional weighting algorithm here. Good results are obtained for directional weighting with an orthogonal decomposition of the information matrix [8]. Here, at every moment, information is divided into one that can be discarded without harm for identification, and a new one that is not contained in the information matrix.

In this case, the algorithm for recurrent calculation of the information matrix instead of (8) will have the form

$$R_i(t) = [I - M_i(t)]R_i(t - 1) + \phi_i(t)\phi_i^T(t), \quad i = 1, 2, 3 \quad (11)$$

where

$$M_i(t) = (1 - \lambda_i)[\phi_i^T(t)R_i(t - 1)\phi_i(t)]^{-1}R_i(t - 1)\phi_i(t)\phi_i^T(t), \quad |\phi_i(t)| > \varepsilon \quad (12)$$

$$M_i(t) = 0, \quad |\phi_i(t)| \leq \varepsilon \quad (13)$$

where  $\varepsilon$ —the error, taking into account which makes the algorithm robust to round-off errors and to measurement noise if any.

The application of this algorithm guarantees the fulfillment of inequalities (10). To assess the adequacy of the resulting model, the Fisher  $F$ —criterion can be used [11, 12]. For multivariate regression, it has the form:

$$F_i = \frac{(n - m) \sum_{t=1}^n (\hat{x}_i(t) - \bar{x}_i)^2}{(m - 1) \sum_{t=1}^n (x_i(t) - \hat{x}_i(t))^2}, \quad i = 1, 2, 3 \quad (14)$$

where  $n$ —the number of points in the experiment,  $m$ —the number of coefficients in the regression equation,  $\hat{x}_i(t)$ —the estimate of the output quantity, which is calculated by the formula

$$\hat{x}_i(t + 1) = \phi_i^T(x_i(t))\hat{\theta}_i(t) \quad (15)$$

In this case, the inequality



$$F > F_{1-\alpha, m-1, n-m} \quad (16)$$

where  $F_{1-\alpha, m-1, n-m}$ —Fisher distribution (according to the table).

## 4 Creating a Simulation Model

As mentioned in Sect. 3, a simulation model is necessary to compare the proposed optimal operational control strategy with control, in which a sample of experimental values of technological variables was obtained, based on the selected function.

The simulation model is built based on estimates of the parameters of the regression model (2) or (15), which is obtained from experimental data. The purpose of a simulation is to generate output variables  $x_i(t)$ ,  $i = 1, 2, 3$  for arbitrary values of control variables  $u_i(t)$  and disturbance  $v(t)$ .

It is possible to memorize the values of the obtained estimates (7) in the form of data arrays  $\left\{ \hat{\theta}_i(t) \right\}_{t=1}^N$ , on a time interval of  $N$  points, in this case,  $N = 2000$ . Then, for arbitrary values of controls  $u_i(t)$  and perturbation  $v(t)$ , which is present in the experimental data, it is possible to obtain the values of the output variables for all times using formula (15).

The condition for the consistency of the simulation model with the used experimental data is that when the vector control  $u(t)$  is fed to the input of the simulation model, which coincides with the control in the original sequence of experimental data, a vector random process  $x(t)$  should appear at the output, close to the captured sequence of experimental data. Proximity is understood in terms of root mean square error. The simulation shows that they practically coincide.

## 5 Solution of the Optimal MPC Problem

Here the control is used, which is obtained based on the maximum principle for linear discrete systems with a terminal criterion in the form of a linear form (6). It is known that a solution to this problem always exists [13]. The solution is sought in a finite time interval.

Under the approach adopted in the MPC [14–16] strategy, the optimal problem is solved for each time point  $t$  each time anew. In this case, the value of the final time interval at which the optimal problem is solved is determined by the value of the forecast horizon  $h$ , which was introduced when determining the optimality criterion (6).

The system of equations of state (5) is used as a model, the matrices of which are obtained based on estimates of the parameters of the regression equations obtained at the time  $t$ . But since the solution is sought in the time interval  $[t, t + h]$ , it is necessary

to obtain a forecast for the following points in time:  $t + 1, t + 2, \dots, t + h$ . These predictions can be made in different ways. Consider this algorithm for the moment  $t$ , obtained based on the maximum principle.

The Hamiltonian following (5), (6) for the moments:  $t + 1, t + 2, \dots, t + h$  is calculated by the formula

$$H(t+k) = \lambda^T(t+k) \left[ \hat{A}(t+k)\hat{x}(t+k) + \hat{B}(t+k)u(t+k) + d + v(t+k)f(t+k) \right] \quad (17)$$

where  $k = 0, \dots, h$ ,  $\hat{A}(t+k)$ ,  $\hat{B}(t+k)$ —the predicted values of the matrices at  $k$  steps,  $\hat{x}(t+k)$ —the predicted values of the state vector at  $k$  steps.

The conjugate system has the form

$$\lambda(t+k-1) = \frac{\partial H(t+k)}{\partial x} = \hat{A}^T(t+k)\lambda(t+k), \quad \lambda(t+h) = c. \quad (18)$$

This system is solved for each value of time  $k$  in reverse time, therefore the sign on the right side is replaced with the opposite one. In this case, the control vector is obtained from the condition for the maximum of the Hamiltonian, therefore, at each step  $t+k$ , it is required to find the maximum of the linear form while observing constraints (1). A linear programming problem has the form

$$\max \lambda^T(t+k)\hat{B}(t+k)u(t+k) \quad (19)$$

To solve problems (18), (19), it is necessary to know the predictions of the matrices  $\hat{A}(t+k)$ ,  $\hat{B}(t+k)$  [17–20]. This problem can be solved in different ways, for example, using formulas (7), (11)–(13), but with a forecast not by one step, but by steps. Then for each one can obtain estimates of the matrices  $\hat{A}(t+k)$ ,  $\hat{B}(t+k)$  using the least-squares method.

But, unfortunately, this approach is very difficult to implement, since the algorithm must work from any time value  $t$ , and, therefore, not  $h$  least-squares filters are needed, but much more. One for a one-step forecast ( $t+1$ ), two for a two-step forecast ( $t+2$ ), three for a three-step forecast ( $t+3$ ), etc.,  $k$  for a ( $t+k$ )-step forecast. The initial values for each filter must be shifted one step.

As a result, we get an arithmetic progression with the sum  $h(1+h)/2$ . This is the total number of filters. In this case, at any time of the forecast  $t+k$ , the values of the forecasts for 1, 2, etc. will be available simultaneously  $k$  steps. In this case, the matrices  $\hat{A}(t+k)$ ,  $\hat{B}(t+k)$  are obtained as a result of identification by formula (7) with a further transition to formula (5).

In this case, simpler formulas (7), (11)–(13) were used with a forecast one step from the moment  $t$  at  $k=0$ . Then, solving the adjoint system (18), we obtain

$$\hat{A}(t+k) = A^k(t), \quad k = 1, \dots, h \quad (20)$$

The average value obtained at the stage of building the regression model was taken as the matrix  $\hat{B}(t+k)$ .

As a result of solving (18), we obtain the values of  $\lambda(t+h), \dots, \lambda(t)$ . Control  $u(t+k)$  can be calculated for any  $k$  from condition (19), but following the MPC control strategy, we are only interested in the control at the time  $t$ . Therefore, the problem is solved

$$\max \lambda^T(t)B(t)u(t) = \max[\lambda_1(t) \lambda_2(t) \lambda_3(t)] \begin{bmatrix} b_{11}(t) & b_{12}(t) & b_{13}(t) \\ b_{21}(t) & b_{22}(t) & b_{23}(t) \\ b_{31}(t) & b_{32}(t) & b_{33}(t) \end{bmatrix} \begin{bmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{bmatrix} \quad (21)$$

The controls are calculated by the formulas:

$$u_i(t) = \begin{cases} u_{1\max}, & b_{1i}(t)\lambda_1(t) + b_{2i}(t)\lambda_2(t) + b_{3i}(t)\lambda_3(t) > 0, \\ u_{1\min}, & b_{1i}(t)\lambda_1(t) + b_{2i}(t)\lambda_2(t) + b_{3i}(t)\lambda_3(t) < 0, \end{cases} \quad (22)$$

where  $i = 1, 2, 3$ ,  $u_{1\max}$ ,  $u_{1\min}$  – derived from normal mode statistics.

## 6 Modelling

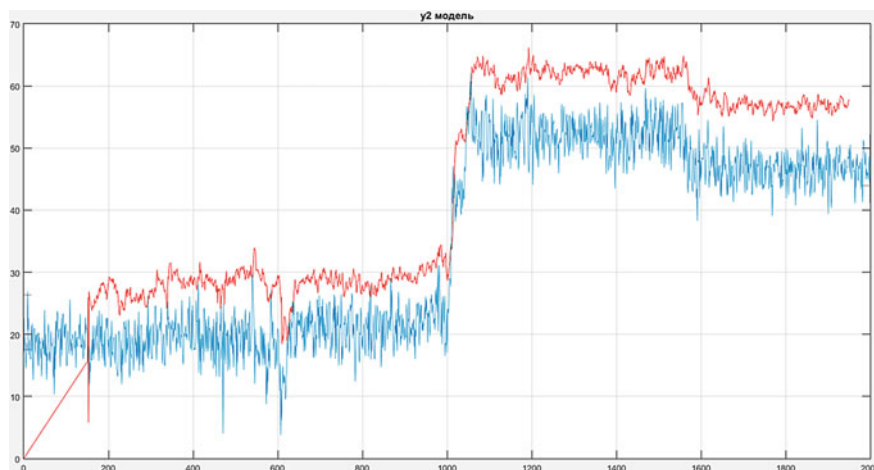
When simulating the work, firstly, the problem of increasing productivity was solved for the forecast horizon  $h = 3$  in comparison with experimental data. Moreover, in (18)  $c^T = [-1 \ 0 \ 0]$ . With optimal control, a performance increase of 1.43% was obtained. Time characteristics for the flow rate of the still product before and after optimization are shown in Fig. 2.

The increase in performance is visible here (red curve).

Besides, the problem of minimizing the temperature at the top of the column was solved within the normal mode. Moreover, in (18)  $c^T = [0 \ 0 \ 1]$ . With optimal control, a temperature decrease of 2.59% was obtained. This helps to reduce energy costs.

## 7 Conclusion

The chapter continues the study of the previously considered method of operational optimal control based on statistical modelling. In previous studies, linear programming and the maximum principle were used to obtain optimal controls. In this case, the MPC control strategy is considered, for the implementation of which the discrete



**Fig. 2** Time dependencies for the consumption of the still product

maximum principle was used for linear systems with limited time and terminal functional at the right end, for which a solution exists, which is important for the system's performance when the optimal problem is solved at each discrete time instant.

The method of operational control is shown by the example of the control of column K-1 of the stabilization department during the hydrotreating of oil fractions. A model with three inputs and three outputs is considered. The tasks of increasing the productivity of the column and reducing energy costs for heating have been solved.

## References

1. Gogol, I.V., Remizova, O.A., Syrokvashin, V.V., Fokin, A.L.: Adaptive-robust control of technological processes with delay on control. In: *Cyber-Physical Systems: Advances in Design & Modelling*, vol. 259. Springer (2019)
2. Gogol, I.V., Remizova, O.A., Syrokvashin, V.V., Fokin, A.L.: Robust control of technological processes with delay on control. In: *Cyber-Physical Systems: Advances in Design & Modelling*, vol. 259. Springer (2019)
3. Gogol, I.V., Kadyrov, J.D., Remizova, O.A., Syrokvashin, V.V., Fokin, A.L.: Operativnoe upravlenie po statisticheskim modeljam processom plavki v pechi Vanjukova [Operational management by statistical models of the melting process in the Vanyukov furnace]. *Izv. Spbsti(TU)* **41**, 128–132 (2017). (in Russia)
4. Islam, M., Okasha, M., Sulaeman, E.: A model predictive control (MPC) approach on unit quaternion orientation based quadrotor for trajectory tracking. *Int. J. Control Autom. Syst.* **17**, 2819–2832 (2019)
5. Lee, J., Kim, J., Song, H., Shim, H.: A constrained consensus problem using MPC. *Int. J. Control. Autom. Syst.* **9**(952) (2011)
6. Singh, A.K.: An extended linear quadratic regulator and its application for control of power system dynamics/AK. Singh, BC. Pal. In: *Proceedings of IEEE First International Conference on Control, Measurement and Instrumentation (CMI)*, pp. 110–114 (2016)

7. Lee, J., Babadagli, T. Comprehensive review on heavy-oil emulsions: Colloid science and practical applications *Chemical Engineering Science* **228**, 115962 (2020)
8. Pyrkin, A., Smyshlyaev, A., Bekiaris-Liberis, N., Krstic, M.: Rejection of sinusoidal disturbance of unknown frequency for linear system with input delay. American Control Conference, Baltimore, USA (2010)
9. Grüne, L., Pannek, J.: *Nonlinear Model Predictive Control*, pp. 43–66. Springer, Berlin (2011)
10. Xu, Z., Zhao, J., Qian, J., Zhu, Y.: Nonlinear MPC using an identified LPV model. *Ind. Eng. Chem. Res.* **48**, 3043–3051 (2009)
11. Pyrkin, A.A., Bobtsov, A.A., Nikiforov, V.O.: Output adaptive controller for linear system with input delay and multisinusoidal disturbance. In: *International Conference on Control Applications*, vol. 23, pp. 1777–1782. IEEE, Antibes, France (2014)
12. Schnelle, F., Eberhard, P.: Adaptive nonlinear model predictive control design of a flexible-link manipulator with uncertain parameters. *Acta. Mech. Sin.* **33**, 529–542 (2017)
13. Boscariol, P., Gasparetto, A., Zanotto, V.: Model predictive control of a flexible links mechanism. *J. Intell. Rob. Syst.* **58**, 125–147 (2010)
14. Darby, M.L., Nikolaou, M.: MPC: current practice and challenges. *Control. Eng. Pract.* **20**(4), 328–342 (2012)
15. Eqtami, A., Dimarogonas, D., Kyriakopoulos, K.: Aperiodic model predictive control via perturbation analysis. In: *IEEE Conference on Decision and Control (CDC)*, pp. 7193–7198 (2012)
16. Zou, Y., Wang, Q., Jia, T., Niu, Y.: Multirate event-triggered MPC for NCSs with transmission delays. *Circuits Syst. Signal Process.* **35**, 4249–4270 (2016)
17. Zykina, A., Kaneva, O., Savkin, V., Fink, T.: Problem Statement for preparing a single batch of end product under uncertainty. In: Sergeyev, Y., Kvasov, D. (eds.) *Numerical Computations: Theory and Algorithms. NUMTA 2019. Lecture Notes in Computer Science*, vol. 11974, pp. 519–527. Springer, Cham (2020)
18. Mayne, D.Q., Rawlings, J.B., Rao, C.V., Scokaert, P.O.M.: Constrained model predictive control: Stability and optimality. *Automatica* **36**(6), 789–814 (2000)
19. Long, Y., Liu, S., Xie, L., Johansson, K.H.: Distributed non-cooperative robust MPC based on reduced-order models. *Control Theory Technol.* **14**, 11–20 (2016)
20. Basturk, H., Krstic, M.: Adaptive sinusoidal disturbance cancellation for unknown LTI systems despite input delay. *Automatica* **58**, 131–138 (2015)

# The Models for Predictive Maintenance of Robotic Agricultural Vehicles



Sergey Susarev , Sergey Orlov , Elizaveta Bizyukova ,  
and Roman Uchaikin 

**Abstract** A cyber-physical system is considered, including a robotic agricultural vehicle system and a complex simulation model for predictive maintenance based on continuous monitoring of the technical condition. Equipping a robotic vehicle with a measuring subsystem, including many sensors and implementing wireless access, provides the transfer of the necessary data for assessing the technical condition in real-time. Continuous monitoring makes it possible to move from preventive maintenance to predictive maintenance, a component of the Industry 4.0 concept, and uses the Internet of Things (IoT) technology. When building an intelligent diagnostic system, the authors used a model-based approach. Together with digital twins of vehicle units, it is possible to identify deterioration processes and predict equipment defects and failures. The general structure of the diagnostic system for robotic agricultural vehicles is presented. The hierarchical structure of simulation models for a robotic vehicle is described. The analysis of existing works in the field of modeling the maintenance for various objects is carried out. A timed colored Petri net model is proposed. The model belongs to the class of stochastic Petri nets and making it possible to evaluate predictive maintenance's effectiveness for a given failure rate of vehicle units and aggregates. A formal description of a robotic vehicle has been developed in the form of the primary parameter multisets. The logical conditions of events in the simulation model are determined. The experimental results have confirmed the model adequacy. The results were applied to study a robotic vehicle group's work in agricultural fields under difficult operating conditions.

**Keywords** Robotic vehicles · Cyber-Physical system · Predictive maintenance · Simulation · Petri nets

---

S. Susarev (✉) · S. Orlov · E. Bizyukova · R. Uchaikin  
Samara State Technical University, 244 Molodogvardeyskaya str, Samara 443100, Russia  
e-mail: [susarev\\_sergey@mail.ru](mailto:susarev_sergey@mail.ru)

E. Bizyukova  
e-mail: [lizaveta5.6@mail.ru](mailto:lizaveta5.6@mail.ru)

R. Uchaikin  
e-mail: [uchaykinra@yandex.ru](mailto:uchaykinra@yandex.ru)

## 1 Introduction

Nowadays, the development of robotics in the agrotechnical industry leads to increased management tasks for the maintenance and repair of robotic vehicles (RV) [1]. Unmanned robotic trucks are complex cyber-physical system that combines components and mechanical units and autonomous subsystems for control, monitoring and information transmission [2, 3]. Many sensors and embedded data processing systems in a robotic vehicle allow obtaining in real-time a sufficient amount of measuring information about the robotic vehicle's technical condition [4]. This chapter discusses the research results carried out by SamSTU jointly with KAMAZ PTC in the framework of a project to create an agricultural robotic vehicle system (ARVS) [5, 6].

Equipping autonomous vehicles with means of measuring parameters and communicating with remote information processing systems made it possible to create "Digital Twins" of the vehicle's main components and assemblies. Continuous monitoring makes it possible to move from preventive maintenance to predictive maintenance, a component of the Industry 4.0 concept, and uses the Internet of Things (IoT) technology [7, 8]. When creating the ARVS, the task was to develop simulation models to optimize an individual RV's maintenance and repair processes, depending on its operation conditions. This problem is being solved based on the developed general system for autonomous vehicle diagnostics and technical condition prediction (DTCP). DTCP uses a model-based approach and tools of the networked concept of vehicle fleet monitoring [5].

## 2 Related Work

The existing works in this research field address several issues, including modeling, analysis, and efficient organization of maintenance and repair of industrial equipment. The work [9] is devoted to constructing decision-making systems to maintain and repair cyber-physical systems, developing methods for analyzing the Key Performance Indicator. It sets out a systemological approach to creating cognitive and multi-agent models of the equipment life cycle, including maintenance and repair.

The article [10] proposes developing an automated system for monitoring and diagnostics of a controlled robotic vehicle. An algorithmic approach is used based on the application of models in the form of fuzzy behavior charts. The authors built models of malfunctions of the electric motor, the battery, and the infrared proximity sensor, making it possible to diagnose malfunctions and abnormal situations.

Many researchers are developing discrete event dynamic models to describe the maintenance and repair of equipment formally. Petri nets stand out among such tools, in which the mathematical and graphic foundations are convenient for simulating and theoretical analysis of parallel processes with discrete events [11–13]. The organization of predictive maintenance and repair requires identifying the properties of

physical processes of deterioration, degradation, and failures of a robotic vehicle's elements. Such processes are random, and therefore it is advisable to use stochastic Petri nets [14]. A maintenance and repair model on a stochastic temporary Petri net-based is presented in [14]. The model structure is given, the resource places corresponding to the consumable resources, and reusable maintenance resources are introduced. The model makes it possible to estimate both the time-related parameters and the economic parameters.

Works [15, 16] are devoted to the offshore wind turbine maintenance organization, usually in coastal waters. Problems of availability and remote control of maintenance are considered. It is proposed to use stochastic Petri nets to simulate weather conditions and three types of service: periodic, condition-based, and corrective maintenance. In [17], simulation models were developed on generalized stochastic Petri nets with predicates and the implementation of the Monte Carlo method. Times to failure of degraded components follow a Weibull distribution with increasing failure rate over time. The maintenance strategies are optimized to minimize the total maintenance costs of the system while maximizing availability. The issues of organizing vehicles' maintenance (aircraft fleet, a fleet of electric car-sharing vehicles) are considered in [18, 19]. They also proposed to use stochastic Petri nets to optimize maintenance while monitoring the object's functioning.

Thus, the study of world experience shows the prospects and advisability of developing simulation models on stochastic Petri nets. We used this mathematical apparatus to develop a diagnostic system for agrotechnical robotic vehicles. The models were based on timed colored Petri nets [20, 21]. Such models are easily structured in a hierarchy, both in top-down and bottom-up development technology.

In the development of the RV diagnostic system, it adopted the concept of using a single device for building simulation models. The main requirement for models is to simulate random time processes and events in heterogeneous system objects. For this purpose, we have used the timed colored Petri nets (TCPN) [20, 21]. Such models implement hierarchical structure in both top-down and bottom-up development technology.

### 3 The Diagnostic and Technical Condition Prediction System

Generalized structure DTCP for robotic vehicles "KAMAZ" is shown in Fig. 1. The system includes units for measuring and processing information on the functioning of a robotic vehicle. The system's second component contains a set of unit's digital twins and simulation models of operating processes, maintenance, and repair [22, 23]. During the operation of a robotic agricultural vehicle, the system solves the following tasks:

- data processing of the RV on-board system and information storage in the cloud storage;



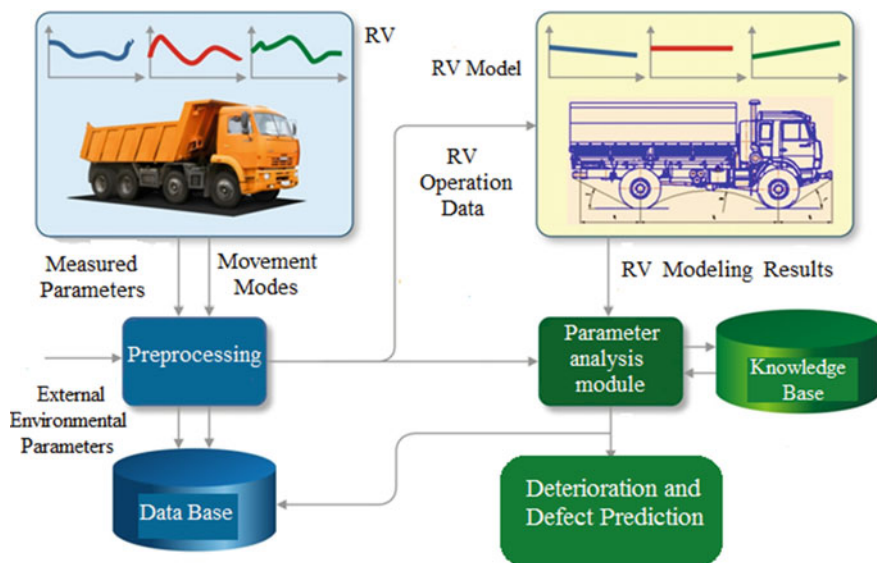


Fig. 1 The DTCP structure

- analysis and statistical processing of diagnostic data to determine their reliability;
- identification of the functioning indicators of RV units, assemblies, and systems;
- comparison of the obtained data with the corresponding parameters of the dynamic system models, adaptive deterioration and degradation models of the systems (considering interference patterns of components and systems);
- analysis of the discrepancy between the parameters of the on-board system and model parameters and the identification of pre-failure states of systems, assemblies, and units;
- statistical analysis of the history of unit and assembly failures, taking into account the RV operation history, forecasting the residual life of units, assemblies, and systems of the vehicle;
- phenomenological analysis of the mutual influence of systems, units, pre-failure states, taking into account the residual life of the equipment, units, and assemblies;
- decision making on conducting RV predictive maintenance.

The units' digital twins are used to analyze the actual vehicle's current parametric characteristics obtained by continuous monitoring. In this case, the system performs the following functions:

1. Accumulates and stores the vehicle functioning parameters (driving modes, load, dynamic parameters, and information from the diagnostic system sensors).
2. It processes RV's parameters by cutting off signal fluctuations caused by non-critical environmental disturbances: disturbances of the road surface, weather

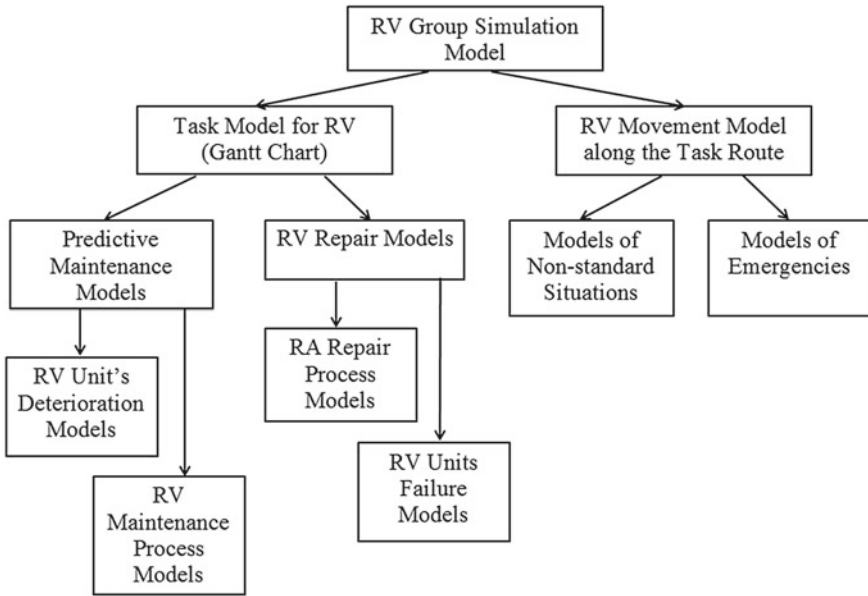


Fig. 2 Hierarchy of RV simulation models for predictive maintenance

conditions, climatic conditions, technogenic disturbances to generate adequate car functioning statistics.

3. Transfer of information to a stationary database to accumulate the “history” of the car’s functioning, or data transfer using the GLONASS system in case of emergency: road accidents and accidental defects.
4. It processes statistical and parametric data in the analytical module to determine the most probable pre-defective and pre-emergency conditions to timely perform predictive maintenance or repair operations.

In the present work, we focus on creating simulation models that operate in a single complex with “digital twins” of car units and aggregates. Figure 2 shows the hierarchical structure of RA simulation models associated with predictive maintenance technology.

#### 4 Simulation Model on the Timed Colored Petri Nets

A timed colored Petri net SM is defined as [24]:

$$SM = (C; B, P; T; A, V, F, w),$$

where  $C$  is a finite set of colors;  $B$  is a finite set of tokens with colors  $c \in C$ ;  $P$  is a finite set of places  $\{p_1, \dots, p_m\} \in P$ , each place  $p$  has a color  $c(p) \in C$  and a set of tokens  $b(p) \subseteq B$  of a given color;  $T$  is a finite set of transitions;  $V$  is a finite set of variables  $v \in V$  corresponding to colors  $c \in C$ ;  $F$  is a finite set of functions that are expressions to describe actions on arcs connecting places and transitions;  $A$  is a matrix of incidence of positions and transitions,  $A \subseteq (P \times T) \cup (T \times P)$ ;  $w: T \rightarrow Z^+$  is a time firing function that assigns a positive integer to each transition in the network.

The following expression represents the description of the robotic vehicle parameters in terms of service and repair:

$$RA_j = (n_j, \text{mod}_j, \text{mdf}_j, \text{mil}_j, RM_j, RL_j), \quad j = 1, \dots, J, \quad (1)$$

where  $J$  is the quantity of RV;  $n_j$  is an individual car number;  $\text{mod}_j$  is a model of RV;  $\text{mdf}_j$  is the number of model modification;  $\text{mil}_j$  is mileage, km;  $RM_j = \{t_{Mjk}\}$  is a set of time before RV unit maintenance, hour;  $RL_j = \{t_{Ljk}\}$  is a set of unit residual life, hour,  $\forall k$ ;  $k = 1, \dots, K$ ;  $K$  is the aggregate quantity.

Colored tokens describe the objects under study in the form of multisets in colored Petri nets. At the initial stage, for simplicity, we consider the description of the robotic vehicle as a multiset  $RV = \{n, \text{"Model"}, \text{mdf}\}$ . The operation of the timed multisets addition is denoted as " + + + ." For example, the presence of two robotic cars in the same position of the Petri net is interpreted as the addition of multisets:

$$RV_1 + + + RV_2 = \{5, \text{"KAMAZ"}, 2\} + + + \{10, \text{"KAMAZ"}, 3\},$$

where the RA models of KAMAZ of the second and third modifications with individual numbers 5 and 10, respectively, are presented.

A simulation model on a timed colored Petri net for investigating the processes of maintenance and repair of RA when performing a specific set of Z production tasks is presented in Fig. 3. The model is constructed using the CPN Tools software system [25]. Models of a similar structure are denoted by SM $n$ m, where  $n$  is the number of task sets,  $m$  is the number of robotic vehicles in the group.

In Fig. 3, model SM13 describes the functioning of a group of three RVs, with one RV performing tasks  $z_1, \dots, z_5$ , and two RVs are in reserve (Reserve place). The set of tasks includes  $z_1$ —the RV movement along the route from the base point BP to the target point TP,  $z_2$  and  $z_4$ —maneuvering near TP,  $z_3$ —performing a production operation,  $z_5$ —returning RV to the base point. For this model, the real and model time ratio is as follows: one cycle MTU is equivalent to 0.5 h, MTU is the model time unit.

Tables 1 and 2 describe transitions and places for the SM13 simulation model.

Two groups of transitions and positions in the network define random processes of failure and requests for maintenance by the predicted estimates of the state of RV units and aggregates located on the route. The Vehicle place describes the technical condition of the robotic vehicle, allowing the execution of tasks. In the event of a

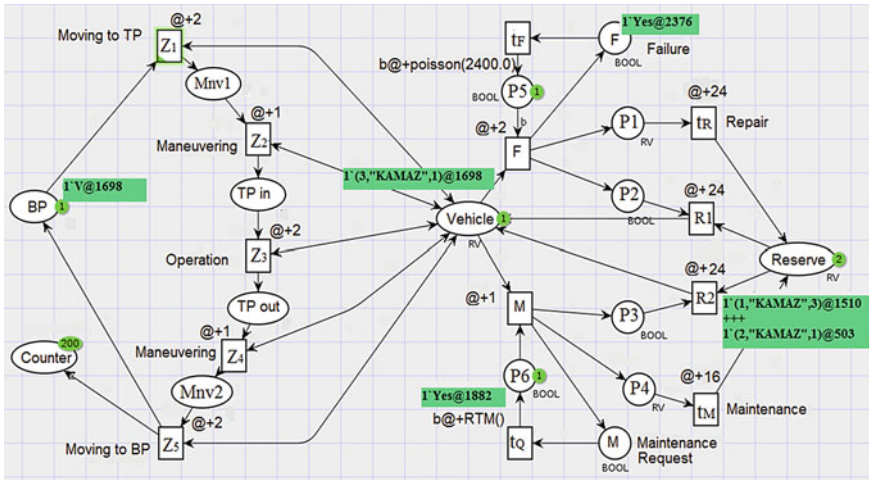


Fig. 3 Simulation model  $SM_{13}$  of the robotic vehicles group

Table 1 Interpretation of transitions of the Petri net model  $SM_{13}$

Transitions	Meaning and events
$z_1$	RV moving to Target Point (TP)
$z_2$	RV maneuvering near TP upon arrival
$z_3$	RV performs a technological operation
$z_4$	RV maneuvering near TP on departure
$z_5$	RV moving to Base Point (BP)
$t_F$	A failure of RV equipment appears, and RV decommissioning initiates
F	RV picked up for repair by service from the current operation
$t_Q$	A maintenance request appears, and RV decommissioning initiates
M	RV picked up for maintenance by service from the current operation
$t_R$	RV repair
$t_M$	RV maintenance
R1	Return of RV from reserve to operation after repair
R2	Return of RV from reserve to operation after maintenance

maintenance or failure request, the RV is extracted and replaced with a serviceable reserve car. In the model in Fig. 3, failures appear according to Poisson’s law. Maintenance requests appear by a generator of random events with a uniform distribution in a given time range.

**Table 2** Interpretation of places of Petri net model  $SM_{13}$ 

Places	Meaning
BP	The base point of RV location and the beginning of the route
Mnv1	It's marking the end of the movement route and readiness for maneuvering
TP in	It marks the arrival of RV at the target point TP and readiness to perform operations
TP out	It marks the departure of RV from the target point TP and readiness to maneuver
Mnv2	It's marking the end of the maneuvering and readiness for the movement route
Counter	Counter of task execution cycles
Vehicle	It contains the number of robotic vehicles at a given moment. The vehicle is available and ready
F	It's marking the appearance of a failure
M	It's marking the appearance of a maintenance request
Reserve	Its marking represents cars ready and available for operations
P1, P2, P3, P4	Auxiliary positions
P5	Decommissioning permit for repair
P6	Decommissioning permit for maintenance

## 5 Experiments and Discussion

Figure 3 shows the model's state after 200 cycles of performing a set of tasks Z. The TSM execution time on the model is 1698 clock cycles, while in the ideal case (without failures and maintenance), the TSI time should be 1600 clock cycles. The decommissioning and replacement of the RV took 98 cycles or 49 h. At the same time, we see that two cars entered the Reserve place after maintenance at cycle 503 and repair at cycle 1510. The model also shows when the dispatcher will take the robotic vehicle out of service: (a) a maintenance request will come at 1882 cycle; (b) repair will come at 2376 cycle.

The proposed model has the advantage that individual places and transitions can, in turn, be represented by a complex Petri net. For example, the Reserve position can be expected for several Petri subnets, each of which describes a particular robotic vehicle's operation.

Table 3 shows the simulation experiment results on the SM13 model (Fig. 3) for options with different backup RV numbers. The simulation was performed by the method of statistical tests (Monte-Carlo method).

The quality of the RV group functioning was assessed by the following indicators:

- the completion task shift factor of the production cycle  $K_Z = T_{SM}/T_{SI}$ ;
- delay  $\Delta$  of task execution;
- the number of stand-by RV necessary to minimize task delays associated with repair or maintenance work.

**Table 3** Results of simulation modeling at a failure rate  $\lambda = 0.8 * 10^{-3}$  and an average time before maintenance equal to 360 h

Cycles	Cycle time, h	$T_{SI}$ , h	$T_{SM}$ , h	$K_Z$	Exceeded assignment deadline $\Delta$ , h	RV reserve, units
200	4	800	849	1.06	49	2
200	4	800	874	1.09	74	1
200	4	800	897	1.12	97	0
400	8	1600	1645	1.03	22.5	2
400	8	1600	1675	1.05	37.5	1
400	8	1600	1707	1.07	53.5	0

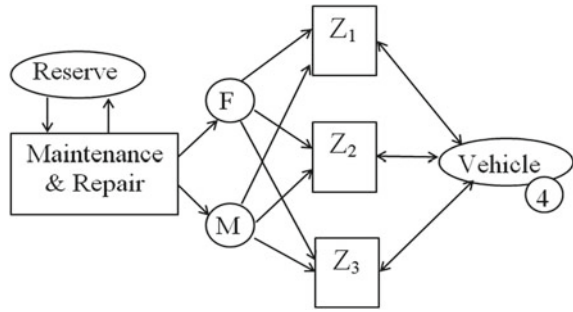
In the experiment, two variants of the cycles of agrotechnical production operations (800 and 1600 h) were set. The shift values in the tasks' timing are obtained for a different number of reserve vehicles. The delay increases with decreasing reserves. The cost of creating a vehicle's reserve and the number of losses from the suspension of production operations consider at the next stage of the experiments.

## 6 Development of the Hierarchy of the Maintenance and Repair Simulation Model

Further development of the proposed simulation model is carried out in the following directions.

1. Detailing the RV model by representing the colored Petri net places in units and aggregate multisets built based on expression (1). In this case, the completion task shift factor is the function  $KZ = f(G, \Lambda, RM, RL, NR)$ , where  $G$  is the Gantt diagram of the task schedule  $Z$ ,  $\Lambda = (\lambda_1, \dots, \lambda_K)$  is are the failure rates at the nodes, and aggregates,  $RM$  is the set of time before maintenance for the aggregates,  $RL$  is the set of values of the residual life of the aggregates,  $NR$  is the number of reserve RV. All of the above parameters are set in the corresponding subnets of the general Petri net.
2. Different distribution laws for random events will provide modeling of deterioration, degradation, and defects for vehicles. The CPN Tools system can specify such laws of distribution of random variables as Weibull, uniform, Poisson, Erlang, Student's, Bernoulli.
3. Detailing the process of performing production tasks depends on the complexity of the RV routes, road conditions, and operations' characteristics at the route's target point. Such an approach allows you to consider stress effects (temperature, a vibration of units, shock loads).
4. Simulate the organization of maintenance and repairs in the field without removing robotic vehicles from the production operation place.

**Fig. 4** Simulation model  $SM_{34}$



Model  $SM_{34}$ , to analyze the actions of an RV group of four vehicles serving three tasks, is shown in Fig. 4. Here, each substitution transition,  $Z_1$ ,  $Z_2$ , and  $Z_3$ , are represented by subnets similar to the  $SM_{13}$  model. The substitution transition "Maintenance and repair" refers to a subnet, which simulates the process of repair and maintenance and implements the probabilistic distribution of the random variable:

- failure time;
- maintenance request time;
- recovery time after repair;
- recovery time after maintenance.

Positions  $F$  and  $M$  correspond to the conditions of failure and maintenance request occurrence.

## 7 Conclusion

The obtained results show the practicality of using temporary colored Petri nets as simulation models. Constant monitoring of robotic vehicles and correction of simulation models based on it are the basis of the RV's predictive maintenance system. Timely determination of the times for maintenance can reduce the losses on vehicle extracting and replacement. The developed models make it possible to determine the optimal composition of reserve vehicles, consistent with the schedule of tasks performed and the procedure for carrying out maintenance, which reduces operating costs.

**Acknowledgements** The work was funded by Russian Foundation for Basic Research according to the research projects No. 20-08-00797 and No. 20-37-90014.

## References

1. Bell, B.: Farm Machinery. 6th edn. Old Pond Publishing (2015)
2. Platzer, A.: Logical Foundations of Cyber-Physical Systems. Springer International Publishing (2018)
3. Khaitan, S.K, McCalley, J.D.: Design techniques and applications of cyber-physical systems: a survey. *IEEE Syst. J.* (2015)
4. Lee, E.A., Seshia, S.A.: Introduction to Embedded Systems. A Cyber-Physical Systems Approach. 2nd edn. The MIT Press (2017)
5. Gubanov, N., Miheev, Yu., Morev, A., Odintsov, V., Akhtyamov, R.: Architecture of a system for diagnosing and predicting the technical condition of a robotic vehicle. In: Proceedings of XXI International Conference Complex Systems: Control and Modeling Problems (CSCMP). *IEEE Xplore*, pp. 671–674 (2019)
6. Orlov, S.P., Susarev, S.V., Kravets, O. Ya., Morev, A.S.: Information system of agricultural robotic KAMAZ cars. *J. Phys. Conf. Ser.* **1399** (033020) 1–5 (2019)
7. Werbińska-Wojciechowska, S.: Technical System Maintenance. Springer International Publishing, Delay-Time-Based Modelling (2019)
8. Chachada, A. et al.: Maintenance 4.0: intelligent and predictive maintenance system architecture. In: Proceedings of IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA) (Turin, Italy). *IEEE Xplore* (2018)
9. Kizim, A.V., Kravets, A.G.: On Systemological approach to intelligent decision-making support in industrial cyber-physical systems. In: Kravets A., Bolshakov A., Shcherbakov M. (eds) *Cyber-Physical Systems: Industry 4.0 Challenges. Studies in Systems, Decision, and Control*, vol. 260, pp. 167–183. Springer Nature Switzerland (2020)
10. Bazhanov, A., Vashchenko, R., Rubanov, V., Bazhanova, O.: Development of an automated system for monitoring and diagnostics a guided robotic vehicle. In: Kravets, A., Bolshakov, A., Shcherbakov, M. (eds.) *Cyber-Physical Systems: Advances in Design & Modelling. Studies in Systems, Decision, and Control*, vol. 259, pp. 93–107. Springer International Publishing, Cham (2020)
11. Wang, J.: Time Petri Nets. In: *Timed Petri Nets. The Kluwer International Series on Discrete Event Dynamic Systems*, vol. 9. Springer, Boston, MA (1998)
12. Del Foyo, P.M.G., Silva, J.R.: Improving the verification of real-time systems using time Petri nets. *J. Control Autom. Electric. Syst.* **28**(6), 774–784 (2018)
13. Pla, A., Gay, P., Melendez, J., Lopez, B.: Petri net-based process monitoring. A Workflow management system for process modeling and monitoring. *J. Intell. Manuf.* **25**, 539–554 (2014)
14. Lu, Z., Liu, J., Dong, L., Liang, X.: Maintenance process simulation-based maintainability evaluation by using stochastic colored Petri net. *Appl. Sci.* **9**(16), 3262 (2019)
15. Leigh, J.M., Dunnett, S.J.: Use of Petri Nets to model the maintenance of wind turbines. *Qual. Reliab. Eng.* **32**(1), 1–30 (2014)
16. Yan, R., Dunnett, S.J.: Improving the strategy of maintaining offshore wind turbines through Petri net modeling. *Appl. Sci.* **11**(2), 574, 1–20 (2021)
17. Santos, F.P., Teixeira, A.P., Soares, C.G.: Modeling, simulation, and optimization of maintenance cost aspects on multi-unit systems by stochastic Petri nets with predicates. *SIMULATION* **95**(5), 461–478 (2018)
18. Sheng, J., Prescott, D.: A colored Petri net framework for modeling aircraft fleet maintenance with cannibalisation. *Reliab. Eng. Syst. Saf.* **189**, 67–88 (2019)
19. Hamroun, A., Labadi, K., Lazri, M.: Modelling and performance analysis of electric car-sharing systems using Petri nets. In: Proceedings of E3S Web of Conferences, vol. 170 (03001) (2020)
20. Jensen, K.: *Coloured Petri Nets. Basic Concepts, Analysis Methods, and Practical Use.* Springer, Berlin Heidelberg (1997)
21. Jensen, K., Kristensen, L.M.: *Coloured Petri Nets: Modelling and Validation of Concurrent Systems.* Springer, Berlin (2014)



22. Orlov, S.P., Susarev, S.V., Morev, A.S., Kravets, O.Ya.: Digital tests of the robotic chassis cyber-physical system for an agricultural unmanned vehicle. *J. Phys. Conf. Ser.* **1399**(044032) (2019)
23. Orlov, S., Susarev, S., Pugachev, A.: Intelligent diagnostic system of robotic KAMAZ vehicle's units. In: *Proceedings of XXI International Conference Complex Systems: Control and Modeling Problems (CSCMP)*. IEEE Xplore, pp. 307–310 (2019)
24. Popova-Zeugmann, L.: *Time and Petri Nets*. Springer, Berlin Heidelberg (2013)
25. CPN Tools. Modeling with Coloured Petri Nets. <http://cpntools.org/2018/01/16/getting-started>. Accessed 20 Sept 2020

# Development of a Cyber-Physical System for Neurofuzzy Prediction of the Concentration of the Contained Prime During Transportation of Oil Wells Emulsion



Artur Sagdatullin  and Gennady Degtyarev

**Abstract** Over time, the classical methods of production and development of deposits become insufficient for the formation of a stable flow of emulsion from wells. This is due, for the most part, to the increasing water cut, the presence of impurities in the oil, affecting its physicochemical structure. All these factors lead to significant uncertainties reflected in the mathematical description of objects and processes. The problem of the development of mathematical models for these objects is the continuous change in the parameters of the pumping equipment, in turn, affects the operation of the built-in control and automation systems of the objects under consideration. The purpose of the research work is to develop a system that allows predicting the concentration of contained impurities during the transportation of oil emulsion from wells. The criterion for achieving this goal is the ability of the developed system to predict the nonlinear characteristics of an object with an error of less than 22%. To solve this goal, a neuro-fuzzy system has been developed for predicting the concentration of impurities during the transportation of oil emulsion from wells. The description of the fuzzy set of the system under consideration is considered. The proposed system showed an error of 21.34%; the training of the system includes 55 epochs.

**Keywords** Production · Transportation and treatment of oil · Presence of impurities · Neural network fuzzy model · Architecture · Information system · Profitability

## 1 Introduction

Technological processes of oil production and transportation are the most energy-intensive in the fields. Many objects are characterized by nonlinearities and parametric uncertainties, which affect the nature of the equipment operation modes. The

---

A. Sagdatullin (✉) · G. Degtyarev  
Kazan National Research Technical University Named After A. N. Tupolev - KAI, K. Marx st.,  
10, Kazan 420111, Russia  
e-mail: [saturn-s5@mail.ru](mailto:saturn-s5@mail.ru)

regulation of pumping units essentially depends on the load on the rotor of centrifugal pumps transporting oil. Consistency can change depending on the mining horizon and over time. It is quite difficult to plan such a change in load in advance. Therefore, it is relevant to develop a method for predicting the concentration of impurities during the transportation of oil emulsion from wells.

Description of the scientific problem. Over time, the classical methods of production and development of fields become insufficient for the formation of a stable flow of emulsion from wells. This is mainly due to the increasing water cut, the presence of impurities in the oil, affecting its physicochemical structure. All these factors lead to significant uncertainties reflected in the mathematical description of objects and processes. The continuous change in the parameters of the pumping equipment, in turn, affects the operation of the built-in control and automation systems of the objects under consideration. This influence is expressed in the characteristics of controlled quantities and errors in the regulation process. Deviation from the nominal characteristics leads to energy losses and an increase in the cost of the technological process itself [1, 2].

The purpose of the research work is to develop a system for neuro-fuzzy prediction of the concentration of impurities contained during the transportation of oil emulsion from wells. The criterion for achieving this goal is the ability of the developed neuro-fuzzy system to predict the nonlinear characteristics of an object with an error of less than 22%.

To achieve the goal set in scientific work, the following tasks are solved:

1. Analysis of the existing method of working with information about the parameters of the incoming emulsion.
2. Investigation of the construction of automatic control systems for multidimensional technological processes.
3. Development of a method for neuro-fuzzy prediction of the concentration of impurities during transportation of oil emulsion from wells.

## **2 Statement of the Problem of Obtaining Information Method About an Object and Technological Process**

The most accurate method for obtaining information about an object, technological process, and the emulsion is the creation of enlarged models of the field. These models make it possible to accurately simulate most of the processes occurring in the field and, as a rule, serve to support and make decisions. The main disadvantage of this method is the rather complicated process of developing this model and the inertia inherent in large models of objects.

In this regard, to improve the quality of existing models, it is promising to develop and apply methods based on fuzzy logic and neural networks. These systems include neural algorithms.

Among the known methods for solving the problem, artificial neural networks (ANN) have been selected, which will be used in the refinement and correction of

hydrodynamic models. The idea of neural networks and their models have shown efficiency for the analysis of unstructured data, for which linear models are not very effective [3-5].

The basis of this system is fuzzy sets, representing a mapping of one of the input parameters  $\times 1, \times 2, \dots, \times n$  in the conditional range  $[0, 1]$ , where  $\mu(x)$  is a membership function.

The set is formed based on a set of possible values or a range in which a given number exists.

$$A = \frac{\mu_A(x_1)}{x_1} + \frac{\mu_A(x_2)}{x_2} + \dots + \frac{\mu_A(x_3)}{x_3}; \tag{1}$$

For example, consider a fuzzy set describing the number 5:

$$A = \frac{0,7}{4} + \frac{1}{5} + \frac{0,8}{6};$$

The membership function can be described by the following system (2):

$$\mu_A(x) = \begin{cases} 0, & \text{if } x < a_1, \\ \frac{x - a_1}{a_2 - a_1}, & \text{if } a_1 < x < a_2, \\ 1, & \text{if } a_2 < x < a_3, \\ \frac{x - a_3}{a_4 - a_3}, & \text{if } a_3 < x < a_4, \\ 0, & \text{if } x > a_4, \end{cases} \tag{2}$$

where  $a_1, a_2, a_3, a_4$ —coefficients of the set are formed.

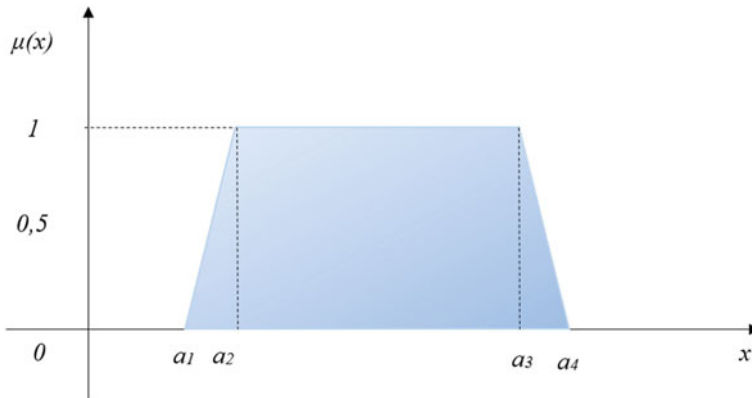
Figure 1 shows an example of a fuzzy set of the system (2).

The description of the fuzzy set of the system under consideration can be represented as follows:

$$\mu_A(x) = \begin{cases} 0, & \text{if } x < a_1, \\ 1, & \text{if } a_2 < x < a_3, \\ \frac{x - a_3}{a_4 - a_3}, & \text{if } a_3 < x < a_4, \\ 0, & \text{if } x > a_4, \end{cases}$$

where  $a_1 = a_2, a_3 \approx a_4$ , what is possible to implement at the software level.

The features of the system under consideration include an estimated, conjectural judgment on the value of the parameter. For example, it can be said that the presence of impurities is about 0.15%, but definitely not more than 0.25% and not less than 0.07%. Based on these data, you can describe the triangular or rectangular membership functions:



**Fig. 1** Fuzzy set described by the system (2)

$$A = (a_l, m, b_r);$$

$$A = (a_l, m_1, m_2, b_r),$$

where  $a_l = 0,07$ ,  $m = 0,15$ ,  $b_r = 0,25$ .

The sum expressions for these sets can be represented as:

$$\begin{aligned} A_1 + A_2 &= (a_{l1}, m_1, b_{r1}) + (a_{l2}, m_2, b_{r2}) = \\ &= (a_{l1} + a_{l2}, m_1 + m_2, b_{r1} + b_{r2}) \end{aligned}$$

$$\begin{aligned} A_1 + A_2 &= (a_{l1}, m_{11}, m_{12}, b_{r1}) + (a_{l2}, m_{21}, m_{22}, b_{r2}) = \\ &= (a_{l1} + a_{l2}, m_{11} + m_{21}, m_{12} + m_{22}, b_{r1} + b_{r2}) \end{aligned}$$

### 3 The Concept of Development and Integration of New Neural Fuzzy System Models for Parameters Prediction and Identification

The neural fuzzy system includes the basis of fuzzy logic for constructing fuzzy rules and forming the corresponding membership functions that condition the coordination of input and output parameters. As a rule, a generalized form of a condition is formed by a set of antecedents. They are the link between linguistic variables and output terms. Fuzzy rules in the form “IF–THEN” and fuzzy conditional operators are expressions of the form: IF event A THEN event B, where A and B are consequents of fuzzy sets characterized by the corresponding membership functions:

$$IF x_k = A_k^i, THEN K^i,$$

where  $x_k$ —input value ( $k \in [0..n]$ ,  $n \in \mathbb{R}$ ),  $A_k^i$ —membership function of the  $k$ -th input value and the  $i$ -th rule  $i \in [0..n]$ , the confidence factor or weight of the given rule.

The neural fuzzy prediction system contains six levels of feedforward, and no connection is particularly important. As a result, all logic is located exclusively in the links representing neurons and neural network connection topologies. A neural network can have the functionality of a controller with a Sugeno output, where for the first level an activation function is usually used with a maximum equal to 1 and a minimum-0 for the membership function  $\mu$ . The Sugeno-type rule base can be formulated, for example, as follows [1–9]:

$$\begin{aligned}
 &PREDICATE \quad x_1 \text{ IS } A_1^1 \quad \text{AND} \quad x_k \text{ IS } A_k^1 \\
 &RULES : \\
 &IF \quad x_1 = A_1^1 \quad \text{AND} \quad x_k \text{ IS } A_k^1, \quad \text{THEN} \quad y \text{ IS } B^1, \\
 &\dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots, \\
 &IF \quad x_1 = A_1^i \quad \text{AND} \quad x_k \text{ IS } A_k^i, \quad \text{THEN} \quad y \text{ IS } B^i, \\
 &\dots \quad \dots \quad \dots \quad \dots \quad \dots \quad \dots, \\
 &IF \quad x_1 = A_1^{i+n} \quad \text{AND} \quad x_k \text{ IS } A_k^{i+n}, \quad \text{THEN} \quad y \text{ IS } B^{i+n}, \\
 \hline
 &OUTPUT : \\
 &y = B^{i+n}
 \end{aligned}$$

A fuzzy inference system represents the ratio of multiple input–output signals for a neuro-fuzzy system. Thanks to the understandable verbal form of presentation of fuzzy rules in the “if–then” format, it becomes possible to simulate imprecise ways of reasoning [10–15]. This method plays an important role in a person’s ability to make decisions in conditions of uncertainty and imprecision.

An example describing the presented system can be generalized and presented for a system with multiple inputs and rules that describe the structure (Fig. 2).

The neural fuzzy system includes several layers: input parameters ( $x, y$ ), membership functions, rule base ( $x * y$ ), normalization, linearization and forecasting. Input values entering the system pass through a fuzzifier with specified membership functions and rules of fuzzy inference. The output of the consequents in this case will be a classic linear equation. Then the product of the levels of membership functions is carried out to calculate the chain weight of each rule included in the universal set. The weights obtained at this stage are supplied to the linearization block [16, 17].

Figure 3 shows a graphical form of transformation of verbal rules and premises to rules in a given system, consisting of  $n$  rules for the  $t$ -norm [18–25].

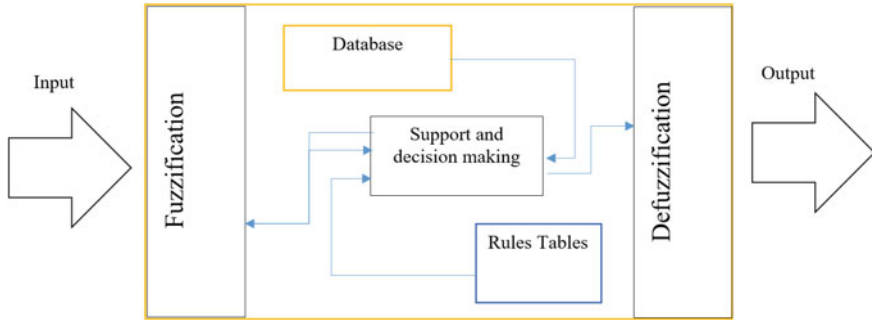


Fig. 2 The system with multiple inputs and rules

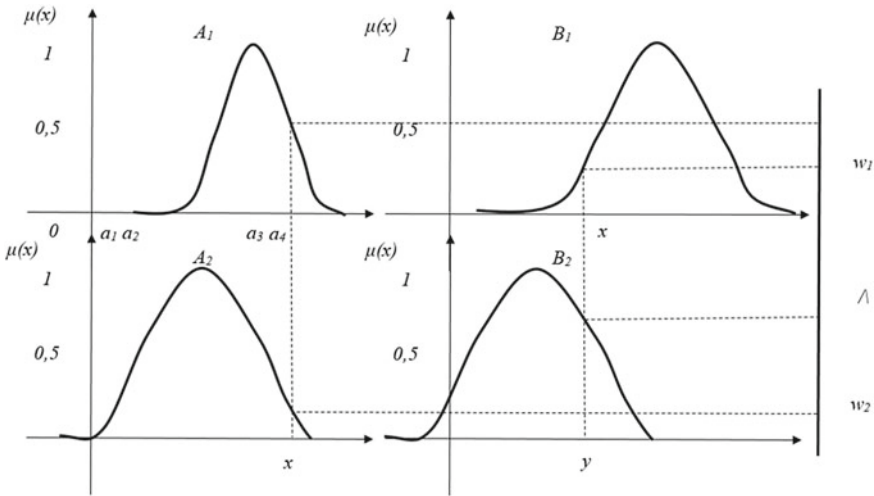


Fig. 3 Graphical form of transformation of verbal rules and premises to rules in a given system

The algorithm of the program can be represented by a block diagram (Fig. 4). The operation of the algorithm begins with the connection of libraries and the formation of a test set. Then the introduction of constants and the initialization of variables are carried out. At this step, the membership functions are entered. In this case, the Gauss membership function and the discrete membership function are considered.

In case of errors during initialization, a message is displayed—Error message. Then, if no errors appear, the first layer is formed, the membership functions are applied to the formed network and then learned according to the given algorithm.

Figure 5 shows the results of training a model based on real data based on the Gauss membership function.

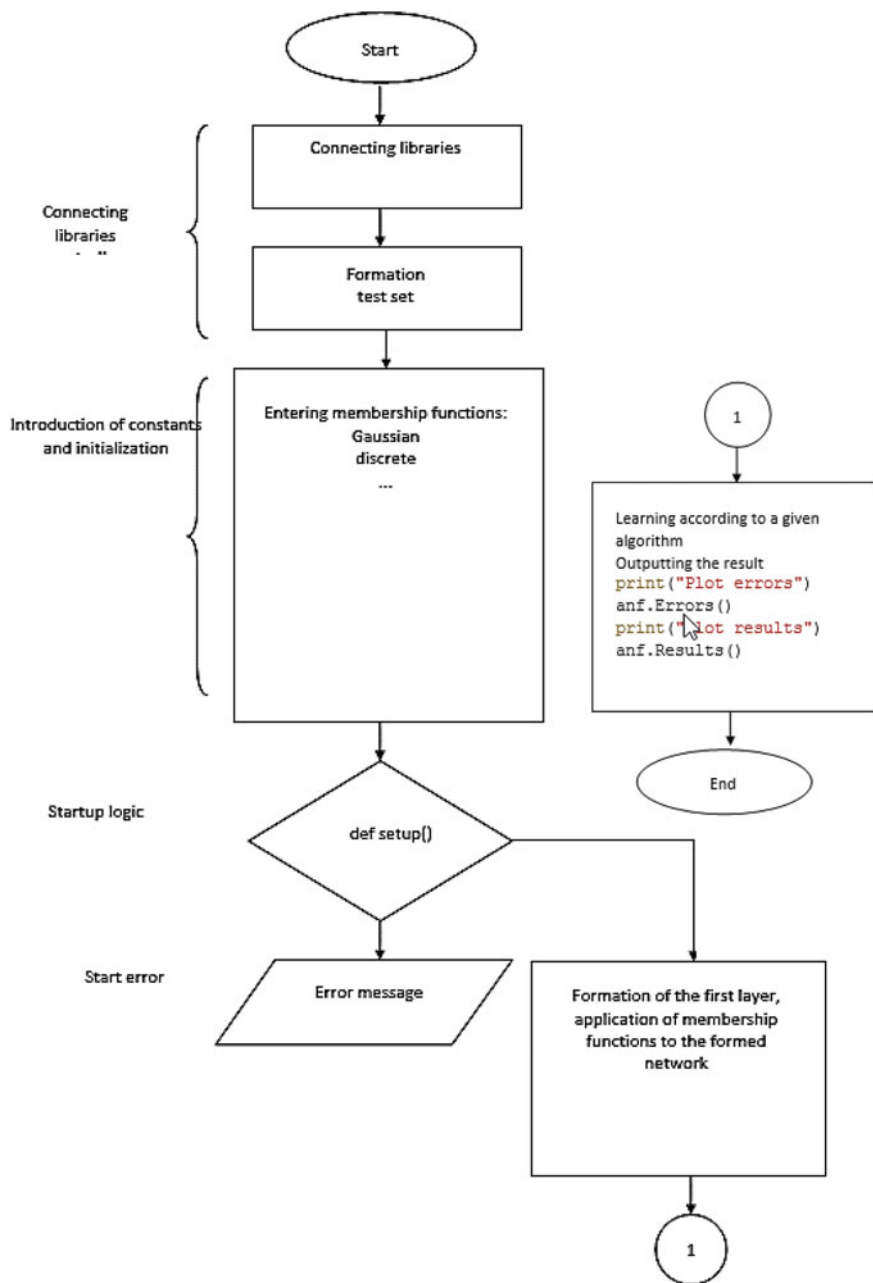


Fig. 4 Algorithm of the program to control blocks of the prediction system



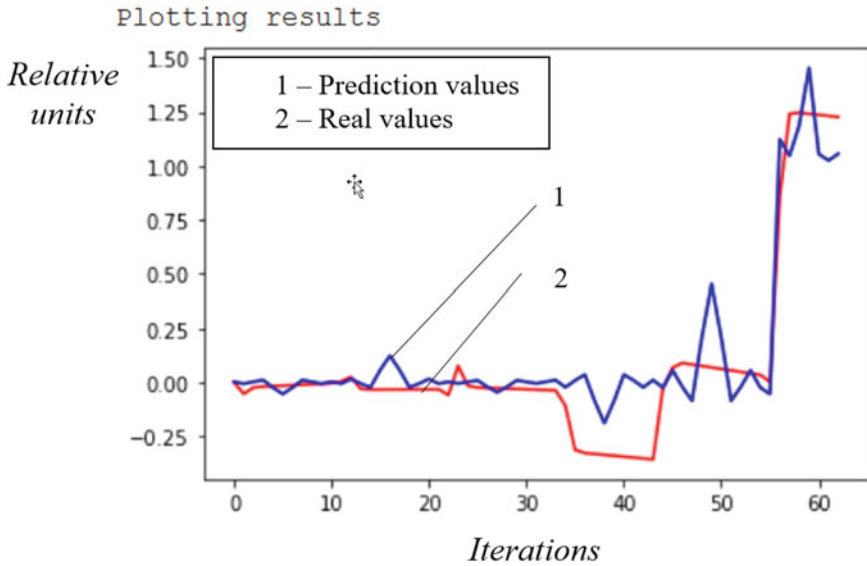


Fig. 5 Results of training a model based on real data based on the Gauss membership function

Figure 6 shows the iterations or learning epochs and the error representing the model's ability to predict actual outcomes. In this case, the error exceeds 37%, which does not meet the initial conditions.

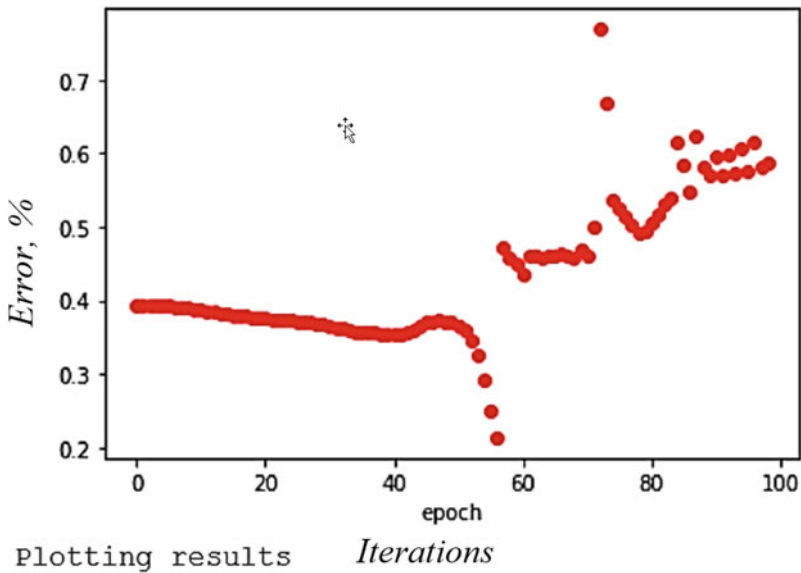
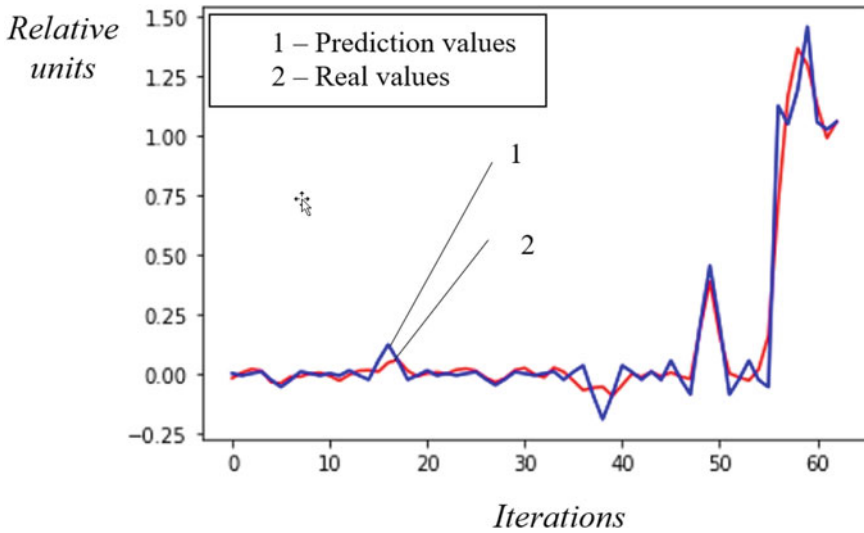


Fig. 6 Results of error change during training of a model



**Fig. 7** Graph of the dependence of the concentration of impurities (%) on the number of values (discrete set)

Figure 7 shows a graph of the dependence of the concentration of impurities (p.u.) on the number of values (discrete set) based on a discrete membership function.

The training of the system includes 55 epochs, while the error of the forecasted values does not exceed 21.34%.

Based on the resulting graph, the following conclusions can be drawn. The developed system of neuro-fuzzy prediction of the concentration of contained impurities during the transportation of oil emulsion from wells makes it possible to effectively predict the values of the non-linear function for a small number of epochs—55. The prediction error was less than 21.34%.

## 4 Conclusion

Fuzzy logic and its applications represent a powerful tool for modeling real processes. In this work, the problem of predicting the concentration of impurities contained during the transportation of oil emulsion from wells was considered. This task is essentially non-trivial due to the presence of gaps in the data, their unstructuredness, and distribution over time. Therefore, the main objectives of this scientific research were: analysis of the existing method of working with information about the parameters of the incoming emulsion; research of construction of systems of automatic control of multidimensional technological processes; development of a method for neuro-fuzzy prediction of the concentration of contained impurities during transportation of oil emulsion from wells. It was revealed that to obtain data on the

concentration of various impurities, enlarged models of the field are created, since they allow to accurately simulate most of the processes occurring in the field and, as a rule, serve to support and make decisions. However, the rather complicated process of developing this model and the inertia inherent in large models of objects does not allow them to be corrected as often as the technological process changes [26–37].

To improve the quality of existing models, methods based on fuzzy logic and neural networks, including neural algorithms, are used. The neural fuzzy system includes the basis of fuzzy logic for the construction of fuzzy rules and the formation of the corresponding membership functions causing the matching of input and output parameters. As a rule, the generalized form of the condition is formed by a set of antecedents. They are the link between linguistic variables and output terms. Fuzzy rules in the form "IF–THEN" and fuzzy conditional operators are expressions of the form: IF event A THEN event B, where A and B are consequents of fuzzy sets characterized by the corresponding membership functions. The graphical form of transformation of verbal rules and premises to rules in a given system consisting of  $n$  rules for  $t$ -norm is presented. The algorithm of the program is presented in the form of a block diagram. The results of training the model based on real data based on the Gaussian membership function are presented. A graph of the dependence of the concentration of impurities (p.u.) on the number of values (discrete set) based on a discrete membership function is presented. The system training includes 55 epochs, while the error of the predicted values does not exceed 21.34%.

## References

1. Harris, J.: *Fuzzy Logic Applications in Engineering Science*. Springer (2006). <https://doi.org/10.1007/1-4020-4078-4>
2. Sagdatullin, A.: Improving automation control systems and advantages of the new fuzzy logic approach to object real-time process operation. In: 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA), pp. 256–260 (2019). <https://doi.org/10.1109/SUMMA48161.2019.8947538>
3. Tabbussum, R., Dar, A.Q.: Performance evaluation of artificial intelligence paradigms—artificial neural networks, fuzzy logic, and adaptive neuro-fuzzy inference system for flood prediction. In: *Environmental Science and Pollution Re-search* (2021). <https://doi.org/10.1007/s11356-021-12410-1>
4. Wang, P.P., Ruan, D., Kerre, E.E.: *Fuzzy Logic*. In: *A Spectrum of Theoretical and Practical Issues*. Springer Berlin Heidelberg (2007). <https://doi.org/10.1007/978-3-540-71258-9>
5. Sagdatullin, A.M.: Development of a start-up model of a submersible electric motor when the electric centrifugal pump is installed and set to operating mode. In: 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA), pp. 456–460 (2019). <https://doi.org/10.1109/SUMMA48161.2019.8947544>
6. Khan, M.S.: Fuzzy time control modeling of discrete event systems. In: *Proceedings of the World Congress on Engineering and Computer Science 2008 WCECS 2008*, October 22–24, San Francisco, USA (2008)
7. Abdullah J. H. et al.: Fuzzy control system review. *Int. J. Sci. Eng. Res.* **4**(1) (2013)
8. Jang J.: Anfis: adaptive-network-based fuzzy inference system. *IEEE Trans. Syst. Man Cybern.* **23**(3), 665685 (1993). <https://doi.org/10.1109/21.256541>

9. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning, Parts 1–3, *Inf. Sci.*, no. 8, 199–249; no. 8, 301–357; no. 9, 43–80 (1975)
10. Soltanali, H., Rohani, A., Abbaspour-Fard, M.H., Farinha, J.T.: A comparative study of statistical and soft computing techniques for reliability prediction of automotive manufacturing. *Appl. Soft Comput.* 106738 (2021)
11. Yager, R.R., Filev, D.P.: *Essentials of Fuzzy Modeling and Control*. Wiley, New York (1994)
12. Zhao, J., Bose B.K.: Evaluation of membership functions for fuzzy logic controlled induction motor drive. In: *IECON 02 [Industrial Electronics Society, IEEE 2002 28th Annual Conference of the]*, vol. 1, pp. 229–234. IEEE (2002)
13. Sagdatullin, A., Degtyarev, G.: Development of a cyber-physical subsystem for support and decision making of managing oil production and transportation processes under uncertainty conditions. In: *Studies in Systems, Decision and Control*, vol. 342, pp. 145–154 (2021). [https://doi.org/10.1007/978-3-030-66081-9\\_11](https://doi.org/10.1007/978-3-030-66081-9_11)
14. Babanezhad, M., Behroyan, I., Nakhjiri, A.T., Marjani, A., Shirazian, S.: Performance and application analysis of ANFIS artificial intelligence for pressure prediction of nanofluid convective flow in a heated pipe. *Sci. Rep.* **11**(1), 902 (2021)
15. Mamdani, E.H., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. *Int. J. Man Mach. Stud.* **7**(1), 1–13 (1975). [https://doi.org/10.1016/S0020-7373\(75\)80002-2](https://doi.org/10.1016/S0020-7373(75)80002-2)
16. Sugeno, M. (ed.): *Industrial applications of fuzzy control*. Amsterdam; New York: New York, N.Y., U.S.A.: North-Holland ; Sole distributors for the U.S.A. and Canada, Elsevier Science Pub. Co. (1985)
17. Sagdatullin, A.: Functioning and development of a real-time information system for the oil treatment technological process control. In: *2020 2nd International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency (SUMMA)*, pp. 847–852 (2020). <https://doi.org/10.1109/SUMMA50634.2020.9280577>
18. Li, W.: Design of a hybrid fuzzy logic proportional plus conventional integral-derivative controller. In: *IEEE Trans. Fuzzy Syst.* **6**(4), 449–463 (1998). <https://doi.org/10.1109/91.728430>
19. Taylan, E.D., Terzi, Ö., Baykal, T.: Hybrid wavelet–artificial intelligence models in meteorological drought estimation. *J. Earth Syst. Sci.* **130**(1), 38 (2021)
20. Allawi, M.F., Ahmed, M.L., Aidan, I.A., Deo, R.C., El-Shafie, A.: Developing reservoir evaporation predictive model for successful dam management. *Stoch. Env. Res. Risk Assess.* **35**(2), 499–514 (2021)
21. Sagdatullin, A.: Analysis of a discrete object control of oil and gas pumping system under uncertainty and unfull information conditions. In: *2020 International Russian Automation Conference (RusAutoCon)*, pp. 405–409 (2020). <https://doi.org/10.1109/RusAutoCon49822.2020.9208112>
22. Mousazadeh, F., Naeem, M.H.T., Daneshfar, R., Soulgani, B.S., Naseri, M.: Predicting the condensate viscosity near the wellbore by ELM and ANFIS-PSO strategies. *J. Pet Sci Eng* **204**, 108708 (2021)
23. Raharja, M.A., Darmawan, I.D.M.B.A., Nilakusumawati, D.P.E., Supriana, I.W.: Analysis of membership function in implementation of adaptive neuro fuzzy inference system (ANFIS) method for inflation prediction. *J. Phys. Conf. Ser.* **1722**(1), 012005 (2021)
24. Sagdatullin, M.: Development of mathematical model and technological process flow diagram of delivery of reagent in initial treatment of oil production of wells in separator tanks. *Chem. Pet. Eng.* **55**(11–12), 884–887. <https://doi.org/10.1007/s10556-020-00708-9>
25. Wang, Q., Liang, D., Du, J.: Design of mining flameproof voltage stabilizing transformer adopting adaptive fuzzy PID controller. In: *Proceedings of 17th International Conference on Electrical Machines and Systems (ICEMS)*, pp. 678–681 (2014). <https://doi.org/10.1109/ICEMS.2014.7013554>
26. Li, C., Ding, Z., Zhang, G., Xu, L.: Prediction of building energy consumption: a comparative study. In: *Proceedings - 2017 Chinese Automation Congress, CAC 2017*, pp. 1691–1697 (2017)
27. Adyanti, D.A., Asyhar, A.H., Novitasari, D.C.R., Lubab, A., Hafiyusholeh, M.: Forecasts marine weather on java sea using hybrid methods: TS-ANFIS. In: *International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)* (2017)

28. Sagdatullin, A.: Development of an intelligent control system based on a fuzzy logic controller for multidimensional control of a pumping station. In: *Advances in Intelligent Systems and Computing*, 1127 AISC, pp. 76–85 (2020). [https://doi.org/10.1007/978-3-030-39216-1\\_8](https://doi.org/10.1007/978-3-030-39216-1_8)
29. Pujianto, A., Kusri, Sunyoto, A.: Designing decision support system for scholarship prediction using adaptive neuro fuzzy inference system algorithm. *J. Phys. Conf. Ser.* **1140**(1), 012049 (2018)
30. Şahin, M.A.: Comparative analysis of dropout prediction in massive open online courses. *Arab. J. Sci. Eng.* **46**(2), 1845–1861 (2021)
31. Naderkhani, R., Behzad, M.H., Razzaghnia, T., Farnoosh, R.: Fuzzy regression analysis based on fuzzy neural networks using trapezoidal data. *Int. J. Fuzzy Syst.* (2021). <https://doi.org/10.1007/s40815-020-01033-2>
32. Mucha, W.: Comparison of machine learning algorithms for structure state prediction in operational load monitoring. *Sensors* **20**(24), 7087 (2020). <https://doi.org/10.3390/s20247087>
33. Tamir, D.E., Rishé, N.D., Kandel A.: Fifty years of fuzzy logic and its applications. In: *Studies in Fuzziness and Soft Computing*, p. 679. Springer (2015)
34. Rios, Y. Y., García-Rodríguez, J.A., Sanchez, E.N., Alanis, A.Y., Ruiz-Velázquez, Garcia, E.: A.P., Treatment for T1DM patients by a neuro-fuzzy inverse optimal controller including multi-step prediction. In: *ISA Trans.* (2021). <https://doi.org/10.1016/j.isatra.2021.07.045>
35. Lin, Yu-Ch. et al.: Air quality prediction by neuro-fuzzy modelling approach. *Appl. Soft Comput.* **86**, 105898 (2020). <https://doi.org/10.1016/j.asoc.2019.105898>
36. Qi, C., Ly, H.-B., Le, L.M., Yang, X., Guo L., Thai B.Ph.: Improved strength prediction of cemented paste backfill using a novel model based on adaptive neuro fuzzy inference system and artificial bee colony. In: *Construction and Building Materials*, vol. 284, pp. 122857 (2021). <https://doi.org/10.1016/j.conbuildmat.2021.122857>
37. Bensaber, B.A., Pereira, D.C.G., Lahrouni Y.: Design and modelling an Adaptive Neuro-Fuzzy Inference System (ANFIS) for the prediction of a security index in VANET. *J. Comput. Sci.* **47**, 101234 (2020). <https://doi.org/10.1016/j.jocs.2020.101234>

# Non-life Insurance Reserve Prediction Using LightGBM Classification and Regression Models Ensemble



Vladimir Soloviev  and Vadim Feklin 

**Abstract** To predict insurance reserves at the micro-level without data aggregation, a two-stage machine learning model based on enhanced LightGBM decision trees is proposed. The first stage is a classification task: whether there are claims that have arisen but have not been submitted under the contract (IBNR). In the second stage, the insurance reserve is predicted using a regression model in the case of IBNR, or it is assumed to be equal to the sum insured otherwise. It is shown that the proposed method is more effective than traditional methods of insurance reserves forecasting.

**Keywords** Insurance · Reserve prediction · Individual claims · LightGBM

## 1 Introduction

Forecasting insurance reserves is a vital actuarial task since it is important to be sure that reserves are sufficient to cover losses. The most challenging problem is assessing claims that are incurred but not reported (IBNR) [1].

Traditionally, for forecasting insurance reserves, the chain ladder method, the Bornhuetter-Ferguson method, and various regression models. In recent years copula models have become popular.

However, machine learning methods, in particular, boosted gradient trees, are not used at all to assess reserves.

This chapter proposes an approach to forecasting insurance reserves based on LightGBM boosted decision trees [2].

The remainder of the chapter is structured as follows.

Section 2 reviews related works on forecasting insurance reserves.

---

V. Soloviev (✉) · V. Feklin

Financial University Under the Government of the Russian Federation, 4 Fourth Veshnyakovsky, Moscow 109456, Russia

e-mail: [VSoloviev@fa.ru](mailto:VSoloviev@fa.ru)

V. Feklin

e-mail: [VFeklin@fa.ru](mailto:VFeklin@fa.ru)

Section 3 describes the initial data and the research methodology, which is the use of a two-step machine learning model, which in the first step, using the LightGBM classification model, determines if there is an IBNR claim under this contract, and in the second step, using the LightGBM regression model, predicts the reserve value for IBNR cases.

Section 4 discusses the outcome of the work. Using our technique for reserves forecasting, results are compared with using AdaBoost trees and ridge regression.

## 2 Literature Review

The actuarial practice of insurance reserving is traditionally based on a body of claims data structured in triangles.

There are several statistical methods for estimating insurance reserves. Classical techniques of using the chain ladder method [3–7] and the Bornhutter–Ferguson method [7, 8] have proven themselves in practice.

However, these techniques are effective only when the analyzed claims have a high probability and a low impact on the reserve volume.

This approach was used in assessing the reserve volumes in conditions of limited incoming information.

Currently, limited information is not a limitation. Therefore, more and more researchers are inclined to conduct studies at the micro-level based on individual pay-offs [9–11].

Recently, works have begun to appear related to applying new methods for forecasting insurance payments and reserves, including machine learning methods.

One of the approaches to assessing insurance payments and reserves related to health insurance is an approach based on forecasting the costs of citizens for health care.

In [12], Takeshima, Keino, Aoki, Matsui, and Iwasaki used variations of lasso regression to predict the costs of Japanese citizens for medical care. In [13], Frees, Gao, and Rosenberg, a two-component model is used to estimate health care costs: one part of the model estimated the frequency of certain events (visits to clinics, the number of hospitals, etc.), and the other predicted the costs associated with a particular of these events.

Various methods are used to assess insurance benefits directly. So, Taylor et al. [14] used a combination of models (paid and incurred models), which made it possible to reduce the prediction error compared to traditional approaches significantly.

Gschlöbl and Czado [15] found a significant relationship between the number of requirements and their size. To assess the frequency of requirements and their size, they used the Bayesian approach, and the parameters were estimated using the Monte-Carlo method based on Markov chains.

Erhardt and Czado [16] studied the distribution of annual amounts of claims with zero requirements. To do this, they used bundles of discrete and continuous copulas,

which allowed them to approximate continuous copulas that describe the annual sums of requirements.

Pettere and Kollo [17] used a two-dimensional Clayton copula model for calculating IBNR requirements and investigating the relationship between the requirement value and the time from when this requirement arose until the time the payment was made.

To forecast reserves for IBNR requirements, Zhao et al. [18] developed a semi-parametric model of aggregated requirements using the maximum likelihood method.

To solve many problems, generalized linear regression models are successfully used. So Frees and Wang [19] used them to estimate the limit of claims distribution, and Garrido et al. [20] weakened the traditional insurance condition for the independence of the number of claims and their size, including rating factors in the model, which are determined using generalized linear regression models.

Krämer et al. [21] also used a regression approach to estimating the magnitude of requirements and their quantity, provided that they are dependent. To account for this dependence, two-dimensional copulas are used. The authors showed that the explicit inclusion of this dependence in the model profoundly affects these estimates. They offered an algorithm for finding the optimal family of copulas.

Shi et al. [22] offered two approaches for describing the dependence of the number of requirements and their size. The first is based on the decomposition of conditional probability and considers the number of requirements as covariates in the regression model for the average size of requirements. The second uses the copula model to describe the joint distribution of the number and size of requirements. To compare these approaches, a simulation experiment was conducted that showed the advantage of the second approach. In particular, the Gini index for the copula model turned out to be 42.14 against 38.64 for the model based on the first approach and 38.23 for the traditional Tweedie compound Poisson model.

Hua [23] proposed the use of a mixed copula regression based on the GGS copula for modeling aggregate loss under conditions where there is a negative relationship between the loss frequency and its value. This allowed to significantly reduce the forecast error of the total losses. Thus, the total loss estimate based on the GGS copula for the panel data for 2010 turned out to be 4 362 626 with the actual value of 4 159 322, while the total loss estimate, which was based on the Tweedy independence model, turned out to be 6 147 354.

Lee and Shi [24] proposed an approach for periodic insurance claims modeling in a longitudinal installation using copulas to determine the dependence of the claims frequency on time, the dependence of the claims volume on time, and the relationship between the claims frequency and volume.

Heberle and Thomas [25] used fuzzy methods to estimate the requirements values. They built a fuzzy chain ladder (FCL) model, uses the TFN uncertainty to make predictions, and proposes a new approach to forecast prediction error.

de Andrés Sánchez [26] proposes a reserve estimation method based on a combination of Tanaki Ishibuchi and Nii fuzzy regression, with Sherman's reservation scheme.



Lally and Hartman [27] uses the Gaussian process regression with input warping and several covariance functions to estimate future claims to predict reserves. The authors showed that the Gaussian process regression models dominate the chain ladder and growth curve models in terms of accurate predictive accuracy as measured by RMSE. Several variants of Gaussian process models were also proposed, and it was shown that a model with a quadratic exponential covariance function works stably well for all data sets considered.

In contemporary literature, various variations of lasso and ridge regression, copula models, and fuzzy regression are mainly used to model individual reserves.

Machine learning tools that use all the available non-aggregated information can ultimately overcome the problems that arise with more traditional approaches.

But for various reasons, boosted decision tree models are not used to predict insurance reserves.

It seems that the use of such algorithms can significantly improve the quality of reserves estimation.

In this chapter, we consider the possibility of using LightGBM boosted decision trees algorithms for insurance reserves forecasting.

## 3 Research Methods

### 3.1 Data

The following features are used:

- AgreementNo—insurance contract number;
- AccidentDate—date of the insured event occurrence;
- ReportingDate—date of application of the client of the insurance company as a result of the insured event;
- PaymentDate—date of insurance compensation paid to the client by the insurance company;
- LoB—the line of business;
- Status—loss status (“Paid”—payout, “Reserve”—pending);
- SumInsured—the insurance amount equal to the maximum possible compensation amount specified in the insurance contract;
- ReserveAmount—the final amount of payment/reserve for an insured event revalued as a result of an internal investigation by the insurance company;
- ReportTime—time elapsed from the moment of the insured event occurrence until the moment the client contacted the insurance company, years;
- FinTime—time elapsed from the moment the client submitted an insured event statement to the insurance company until the final settlement by the insurance company, that is, the refusal of payment, years;

- Label—synthetic feature equal to 1 if the maximum possible compensation amount is not equal to the final calculated pay-off amount for an insured event, and zero otherwise;
- Target—the difference between the maximum possible compensation amount and the final calculated pay-off amount for an insured event. To estimate the insurance reserves, we use data that includes the period from 2004 to 2019 provided to the authors by a large insurance company from Central Europe.

### 3.2 Algorithm

To implement the solution of the reserves forecasting problem, two subtasks were formulated:

- Determination of whether there is an IBNR situation under the contract;
- Forecasting the reserve volume for the case when the contract is in an IBNR situation or determining a reserve equal to the insurance amount if the contract is not an IBNR contract.

To do this, synthetic features are created:

- Label—is the maximum possible insurance amount equal to the actual estimated pay-off amount;
- Target—the difference between the maximum possible insurance amount and the actually estimated pay-off amount.

The model assumes the following assumptions:

1. Since the valuation date is the date of the report on the insured event, the model assumes that all information about the claim is known at that date, and therefore it can be used to predict future pay-offs;
2. The insurance company pays the policyholder the amount as soon as the final pay-off is established, thus generating a series of cash flows. The model takes into account one single aggregate payment for each claim payable at the closing date;
3. Settlement time is considered a discrete value expressed in years. Analysis of the available statistical data shows that the period for claim resolving does not exceed 6 years;
4. Forecasting is carried out for 2019. The model values obtained from the classification, indicating payment for already declared cases in the next 2020 and later, will relate to the reserve of declared but unresolved losses, not to the reserve of losses.

All amounts are calculated in Euros.

The training data set from 2004 to 2018 has 66,414 records, and the test dataset (2019) was 20 507 records.

## 4 Results and Discussion

As a result of the hyperparameters tuning for the LightGBM classification model, whether there is an IBNR situation under the contract, the following parameters were the best:

- Number of leaves: 256;
- Minimum leaf instances: 50;
- Learning rate: 0.025;
- Number of trees: 500.

As a result of the hyperparameters tuning for the LightGBM regression model for reserve volume prediction in the case when the contract is in the IBNR situation, the following parameters were the best:

- Number of leaves: 256;
- Minimum leaf instances: 50;
- Learning rate: 0.025;
- Number of trees: 500.

The final model produces, as the result,  $MAE = 30.37$ ,  $MAPE = 0.16$ .

For comparison, the LightGBM model without preliminary IBNR/Non-IBMR classification shows  $MAE = 31.32$ ,  $MAPE = 0.18$ , regression on AdaBoost trees— $MAE = 39.59$ ,  $MAPE = 0.26$ , ridge regression— $MAE = 135.64$ ,  $MAPE = 0.87$ .

We propose the use of LightGBM boosted decision trees in the algorithm, which first determines whether the contract has an IBNR situation, and then for the case when the contract is in the IBNR situation, another LightGBM regression model predicts the reserve volume. For this case, if the contract is not an IBNR contract, the reserve is determined to be equal to the insurance amount.

The technique proposed is more effective than traditional methods.

It also can be used as a method for assessing the size of individual insurance reserves.

## References

1. Jewell, W.: Predicting IBNYR events and delays: I Continuous time. *ASTIN Bull.* **19**, 25–56 (1989). 0.2143/AST.19.1.2014914
2. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.-Y.: LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **30**, 1–9 (2017). <https://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree>, Accessed 19 Apr 2020
3. Kaas, R., Goovaerts, M.J., Dhaene, J., Denuit, M.: *Modern Actuarial Risk Theory Using R*. Springer, New York (2008)
4. Wuthrich, M.V., Merz, M.: *Stochastic Claims Reserving Methods in Insurance*. Wiley, New York (2008)

5. Zhang, Y.: A general multivariate chain ladder model. *Insur. Math. Econ.* **46**(3), 588–599 (2010). <https://doi.org/10.1016/j.insmatheco.2010.03.002>
6. Denuit, M., Trufin, J.: Collective loss reserving with two types of claims in motor third party liability insurance. *J. Comput. Appl. Math.* **335**, 168–184 (2018). <https://doi.org/10.1016/j.cam.2017.11.044>
7. Martínez-Miranda, M.D., Nielsen, J.P., Verrall, R.: Double Chain Ladder and Bornhuetter-Ferguson. *North Am. Actuar. J.* **17**(2), 101–113 (2013). <https://doi.org/10.1080/10920277.2013.793158>
8. Hiabu, M., Margraf, C., Martínez-Miranda, M.D., Nielsen, J.P.: Cash flow generalisations of non-life insurance expert systems estimating outstanding liabilities. *Expert Syst. Appl.* **451**, 400–409 (2016). <https://doi.org/10.1016/j.eswa.2015.09.021>
9. Antonio, K., Denuit, M., Pigeon, M.: Individual loss reserving with the multivariate skew normal framework. *ASTIN Bull.* **43**(3), 398–428 (2013). <https://doi.org/10.1017/asb.2013.20>
10. Jessen, A.H., Samorodnitskiy, G., Mikosch, T.: Prediction of outstanding payments in a Poisson cluster model. *Scand. Actuar. J.* **2011**(3), 210–335 (2014). <https://doi.org/10.1080/03461238.2010.481080>
11. Plat, R., Antonio, K.: Micro-level stochastic loss reserving for general insurance. *Scand. Actuar. J.* **2014**(4), 648–670 (2014). <https://doi.org/10.1080/03461238.2012.755938>
12. Takeshima, T., Keino, S., Aoki, R., Matsui, T., Iwasaki, K.: PRM23: development of medical cost prediction model based on statistical machine learning using health insurance claims data. *Value Health* **21**(Supplement 2), s97 (2018). <https://doi.org/10.1016/j.jval.2018.07.738>
13. Frees, E.W., Gao, J., Rosenberg, M.A.: Predicting the frequency and amount of health care expenditures. *North Am. Actuar. J.* **15**(3), 377–392 (2011). <https://doi.org/10.1080/10920277.2011.10597626>
14. Taylor, G., McGuire, G., Sullivan, J.: Individual claim loss reserving conditioned by case estimates. *Ann. Actuar. Sci.* **3**(1–2), 215–256 (2008). <https://doi.org/10.1017/S1748499500000518>
15. Gschlöbl, S., Czado, C.: Spatial modelling of claim frequency and claim size in non-life insurance. *Scand. Actuar. J.* **2007**(3), 202–225 (2007). <https://doi.org/10.1080/03461230701414764>
16. Erhardt, V., Czado, C.: Modeling dependent yearly claim totals including zero claims in private health insurance. *Scand. Actuar. J.* **2012**(2), 106–129 (2012). <https://doi.org/10.1080/03461238.2010.489762>
17. Pettere, G., Kollo, T.: Modelling claim size in time via copulas. In: Transactions of 28th International Congress of Actuaries, Paris, France, 28 May–2 June 2006, 1–10 (2006). [https://www.researchgate.net/publication/228883603\\_Modelling\\_claim\\_size\\_time\\_via\\_copulas](https://www.researchgate.net/publication/228883603_Modelling_claim_size_time_via_copulas). Accessed 19 Apr 2020
18. Zhao, X.B., Zhou, X., Wang, J.L.: Semiparametric model for prediction of individual claim loss reserving. *Insur. Math. Econ.* **45**(1), 1–8 (2009). <https://doi.org/10.1016/j.insmatheco.2009.02.009>
19. Frees, E.W., Wang, P.: Copula credibility for aggregate loss models. *Insur. Math. Econ.* **38**(2), 360–373 (2006). <https://doi.org/10.1016/j.insmatheco.2005.10.004>
20. Garrido, J., Genest, C., Schulz, J.: Generalized linear models for dependent frequency and severity of insurance claims. *Insur. Math. Econ.* **70**, 205–215 (2016). <https://doi.org/10.1016/j.insmatheco.2016.06.006>
21. Krämer, N., Brechmann, E.C., Silvestrini, D., Czado, C.: Total loss estimation using copula-based regression models. *Insur. Math. Econ.* **53**(3), 829–839 (2013). <https://doi.org/10.1016/j.insmatheco.2013.09.003>
22. Shi, P., Feng, X., Ivantsova, A.: Dependent frequency–severity modeling of insurance claims. *Insur. Math. Econ.* **64**, 417–428 (2015). <https://doi.org/10.1016/j.insmatheco.2015.07.006>
23. Hua, L.: Tail negative dependence and its applications for aggregate loss modeling. *Insur. Math. Econ.* **61**, 135–145 (2015). <https://doi.org/10.1016/j.insmatheco.2015.01.001>
24. Lee, G.Y., Shi, P.: A dependent frequency–severity approach to modeling longitudinal insurance claims. *Insur. Math. Econ.* **87**, 115–129 (2019). <https://doi.org/10.1016/j.insmatheco.2019.04.004>

25. Heberle, J., Thomas, A.: Combining chain-ladder claims reserving with fuzzy numbers. *Insur. Math. Econ.* **55**, 96–104 (2014). <https://doi.org/10.1016/j.insmatheco.2014.01.002>
26. de Andrés Sánchez, J.: Calculating insurance claim reserves with fuzzy regression. *Fuzzy Sets Syst.* **157**(23), 3091–3108 (2006). <https://doi.org/10.1016/j.fss.2006.07.003>
27. Lally, N., Hartman, B.: Estimating loss reserves using hierarchical Bayesian Gaussian process regression with input warping // insurance. *Math. Econ.* **82**, 124–140 (2018). <https://doi.org/10.1016/j.insmatheco.2018.06.008>

# Machine Learning-Based Cyber-Physical Systems for Forecasting Short-Term State of Unstable Systems



Alexander Musaev and Dmitry Grigoriev 

**Abstract** We consider the task of developing algorithms for cyber-physical systems (CPS) for proactively managing the state of unstable systems with a chaotically evolving state vector. Examples of such processes are changes in the state of gas- and hydrodynamic environments, stock price evolution, thermal phenomena, and so on. The main problem of this type of CPS is creating forecasts that would allow us to compare the efficiency of different feasible control actions. The presence of a chaotic element in the state dynamics of unstable systems does not allow to build of control CPS based on conventional statistical extrapolation algorithms. Hence, in the current chapter, we consider forecasting algorithms built upon machine learning and instance-based data analysis. In the conditions of chaotic influences, which are common in unstable immersion environments, obtaining an accurate forecast is highly complicated. Within the conducted computational experiment that employed direct averaging by three after-effects of analog windows, the average forecast accuracy oscillates between 15 and 20%. Effective forecasting of a chaotic process of a complicated inertia-less nature based on the considered computational schemes has not been achieved yet. This means that additional research, based on multidimensional statistical measures, is required.

**Keywords** Precedent forecasting · Controlling of a chaotic system · Matrix similarity measures

---

A. Musaev (✉)

Saint-Petersburg State Institute of Technology (Technical University), 90013 St. Petersburg, Russia

e-mail: [amusaev@technolog.edu.ru](mailto:amusaev@technolog.edu.ru)

D. Grigoriev

Saint Petersburg State University (SPBU), 199034 St. Petersburg, Russia

e-mail: [d.a.grigoriev@spbu.ru](mailto:d.a.grigoriev@spbu.ru)

## 1 Introduction

We consider the problem of developing cyber-physical systems (CPS) intended for controlling unstable processes [1, 2]. Examples of such processes are changes in the state of gas- and hydrodynamic environments [3, 4], stock price evolution [5], thermal phenomena [6], and so on. It is quite obvious that most of the complexity of proactively controlling such environments lies in the low efficiency of traditional forecasting algorithms. Any change in the experimental conditions violates the applicability of the entire probabilistic-statistical paradigm and requires new approaches to forecasting and proactive management [7].

Machine learning can be employed as an alternative solution to this problem, in particular, precedent-based data analysis [8–10]. This technique is similar to the *nearest neighbor* and *k-nearest neighbors* [11, 12], and it supposes a database that contains large series of retrospective observational data. First, segments of observational data similar to the current situation are identified using specified similarity measures. Next, a hypothesis is made that such areas of observations entail similar after-effects. Therefore, the segment of retrospective observations that follows the identified precedent segment most similar to the current situation can be used as a forecast.

It is obvious that this approach does not use conventional statistic extrapolation, and it allows to detect and predict sudden and jump-like changes in the state of the observed unstable system. The efficiency of the obtained forecast depends on many factors: the degree of chaos in the observed processes, the selected similarity measure, the size of the segment used to describe the current situation, etc. The chapter closely investigates the efficiency of this approach for creating proactive management CPS.

An important particularity of this task is the focus on the control of unstable systems. In this case, the conventional two-parameter additive observation model:

$$Y(k) = X(k) + v(k), \quad k = 1, \dots, n, \quad (1)$$

where  $X(k)$  is a systemic component that implements dynamic chaos [1, 2],  $v(k)$  is a random process imitating system noise or measurement error. As noted previously, extrapolation is not efficient for this type of processes [7, 8], which actualizes the current study.

## 2 Problem Formalization and Precedent Forecasting Algorithm

Let  $Y(k)$ ,  $k = 1, \dots, N$  be a series of observations a real unstable system described by a model that contains a chaotic systemic component. At the same time, the systemic component of  $X(t)$  is often generated by unstable system state evolution dynamics: it is a non-periodic oscillatory process. Consequently, the conventional

techniques of extrapolating probabilistic processes are unapplicable: it is impossible to identify the dynamics of the process whose parameters and structure continuously change in accordance with unknown laws. Nevertheless, data analysis techniques based on classification theory and machine learning can be suitable for many inert physical processes.

For this purpose, we will look up analog segments similar to the current situation in a database of retrospective time series, denoting the most similar segment as a precedent. The current situation refers to an interval of the observation series that immediately precedes the current moment of time, i.e.  $Y_W(k) = [Y(k), \dots, Y(k - L)]$ ,  $k = L + 1, \dots, n$ . Window size  $L$  is selected according to the dynamic properties of the observed process. In adaptive CPS,  $L$  can be selected during the parametric identification of the model.

Similar situations are found by sequentially scanning the database of retrospective data of size  $N_0$  with a sliding observation window  $Y_S(i)$ ,  $\forall i = 1, \dots, N_0 - L$  of the same size  $L$ . At each step of the scan, the similarity measure of the current situation  $Y_W(k)$  and the data interval from the sliding window  $Y_S(i)$  is sequentially evaluated. The most similar situation, i.e. the one with the smallest difference between the current state window and the scan window, is a precedent.

As noted previously, this method is based on the assumption that the observed system evolves in a similar fashion when following identical situations. Therefore, an important prerequisite of this technique is the hypothesis that the after-effect of the precedent can forecast changes in the state that corresponds to the current situation. Note that this approach resembles the cognitive approach to forecasting in the human brain [13], and therefore can be applied in artificial intelligence systems.

It is natural to use the square of the distance between the current state window  $Y_W(k) = [Y(k), \dots, Y(k - L)]$  and the database scan windows  $Y_w(i) = Y(i, i + L)$ ,  $i = 1, \dots, N_0 - L$  as the simplest similarity measure  $\rho(Y_W(k), Y_W(k - i))$ . For unstable systems, the values of  $L$  and the feasible prediction interval  $\tau$  can be refined during the correction of the observation model. In the general case, the length of the prediction interval  $\tau$  is defined and fixed by the requirements of the control system for which the CPS was created.

This approach has a lot in common with the nearest neighbor method [11, 12]. As the precedent, they select the scan window number  $i^*$  that has the smallest difference with the current state window  $Y_W(k)$ , i.e.

$$Y_W(i^*): \rho[Y_W(k), Y_W(i^*)] = \min \forall i = 1, \dots, N_0 - L. \quad (2)$$

The most widely used similarity measure [14]  $\rho[Y_{W1}, Y_{W2}]$  is the standard deviation

$$\rho[Y_{W1}, Y_{W2}] = \left( \frac{1}{n} \sum_{i=1}^L (Y_{W1}(i) - Y_{W2}(i))^2 \right)^{\frac{1}{2}}. \quad (3)$$



In some cases that arise due to abnormal observations, a similarity measure based on the mean value of absolute differences can be more efficient:

$$\rho[Y_{W1}, Y_{W2}] = \frac{1}{n} \sum_{i=1}^L |Y_{W1}(i) - Y_{W2}(i)|. \quad (4)$$

In this case, a forecast  $\tilde{Y}(k + \tau)$  is selected to be the after-effect of the precedent window:

$$\tilde{Y}(k + \tau) = Y(i^* + L + \tau) \quad (5)$$

For chaotic processes, forming a forecast based on a single window can lead to unstable estimates. Hence, it is reasonable to average out the after-effects by  $m_a$  analog windows with numbers  $(i_1^*, \dots, i_{m_a}^*)$ , for which the measures  $\rho[Y_W(k), Y_W(i)] \forall i = 1, \dots, N_0 - L$  is the smallest:

$$\tilde{Y}(k + \tau) = \frac{1}{m_a} \sum_{j=1}^{m_a} Y(i_j^* + L + \tau)$$

Further development of this approach is focused on using a weighted sum of after effects with coefficients proportional to the degree of similarity as a forecast. If the precedent is weighted with  $w_1 = 1$ , then all similar analogs are weighted with  $w_j = \rho_1/\rho_j = 1/\rho_j, j = 2, \dots, m_a$ . In this case, the forecast is defined as follows:

$$\tilde{Y}(k + \tau) = \frac{1}{S(m_a)} \sum_{j=1}^{m_a} \omega_j Y(i_j^* + L), \quad (6)$$

where  $S(m_a) = \sum_j^{m_a} w_j$ .

### 3 Specifics of Algorithm and Precedent Forecasting Implementation

The sequential search of analog windows that correspond to the similarity criteria leads to the selection of analog windows consisting of adjacent scan windows. Consequently, using groups of analog windows virtually do not increase the stability of the forecast. In order to eliminate this effect, the algorithm of the analog window selection needs to be modified. In particular, it is reasonable to use a computational scheme that considers the prohibited selection areas  $\Delta_j$  for counts  $i_j, j = 1, \dots, m_a$  corresponding to the selected analogs. In case the next in order of similarity measure value scan window ends up in  $\Delta_j$ , it will be discarded.

Another significant particularity of searching for analogs is the need to dynamically smooth the observed process. In this case, the search for analog situations will be focused on the analogs of the systemic component of the observations, which will allow increasing the certainty of the solutions. In our study, we have used an exponential filter of the form  $\hat{Y}(t) = \alpha Y(t) + (1 - \alpha)\hat{Y}(t - 1)$  with a transfer coefficient  $\alpha \in [0.01, 0.04]$ . For an adaptive version of the precedent forecast, it is reasonable to refine the value of the coefficient during sequential parametric optimization [15, 16].

## 4 Numerical Studies and an Example Implementation of Precedent Forecasting

As an example of a forecasted process, let us consider variations of normalized density of non-stationary gas-dynamic flow. A plot of this process, centered by its average, can be seen in Fig. 1.

An example selection of analog situations can be seen in Fig. 1. The area is split in two in a sufficiently random fashion: the zone of retrospective data used for creating a precedent forecast, and the work zone used for testing the precedent forecasting algorithm.

The current observation window  $Y_W(k)$  immediately adjoins the moment of time  $k$  and has a length of 300 counts. Next, the retrospective data zone is scanned with a

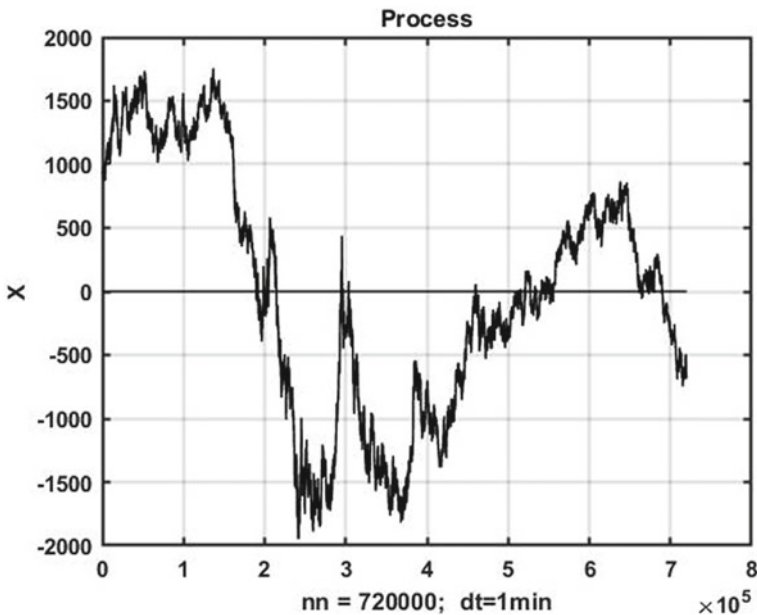


Fig. 1 Observation time series of the forecast process with a chaotic systemic component

scan window  $Y_w(i) = Y(i, i + L), i = 1, \dots, N_0 - L$  and the value of the similarity measure is calculated on each step. The window of retrospective data that corresponds to the minimum value of the measure is selected as the forecast, and its after-effect is considered to be the forecast of how the current situation will develop. In the case of several analog windows, the forecast is selected as the result of averaging (4) or weighted averaging of their after-effects. Figure 2 shows an observation window that corresponds to the current state of the observed system, as well as three analog windows with the highest similarity to the current observation window, which have been detected during the retrospective database scan. The second top window is the precedent according to the nearest neighbor technique since it is the most similar to the current situation depicted on the top plot.

It is easy to see the high degree of geometric similarity of the detected analogs in the current state window, which proves the capability of the search algorithm. Further, Fig. 3 shows plots of prehistoric processes augmented with real process evolution plots and plots of after-effects in analog windows used as forecasts.

Note the specifics of chaotic dynamics, for which the identity of the initial processes allows us to infer the similarity of their after-effects with some limited probability. In particular, it can be seen in Fig. 3 that the after-effect of the third analog cannot be a correct forecast. However, the degree of similarity of the analog window to the precedent window is only 0.49. This means that prior fixing of the analog number for chaotic processes is unreasonable. Additional analog pruning by the threshold level of admissible similarity should be applied.

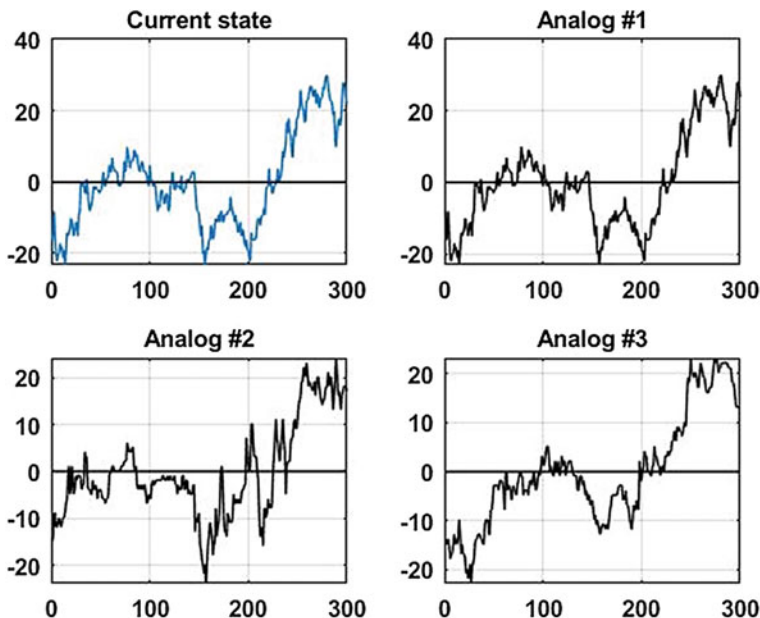
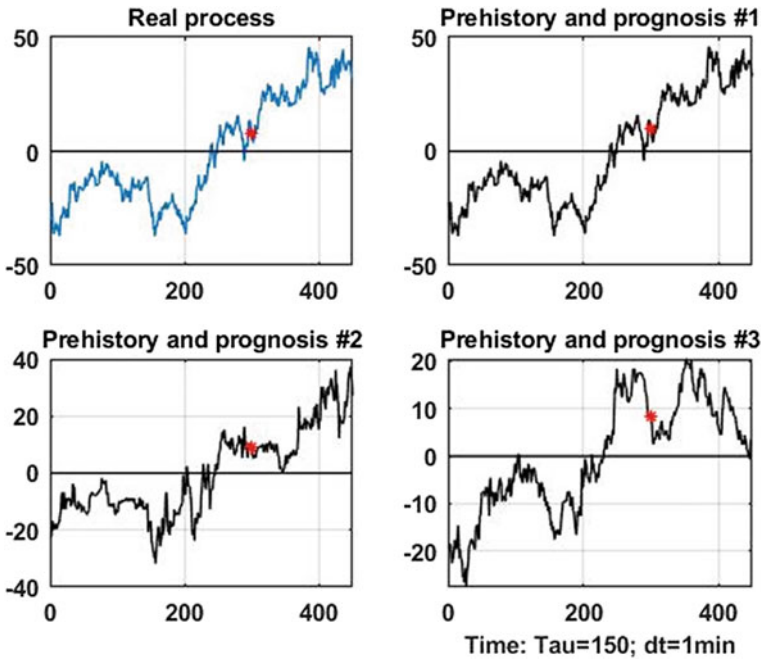


Fig. 2 Current state window and three analog windows



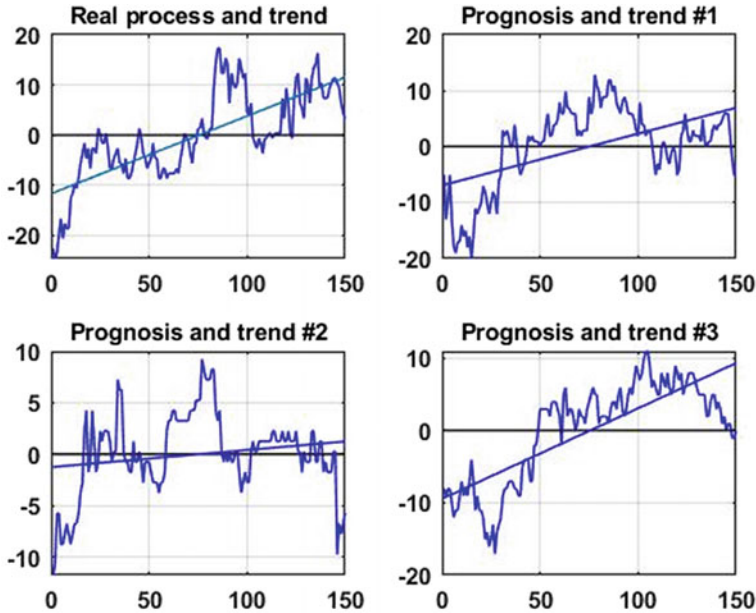
**Fig. 3** The real process (after the asterisk) and three versions of after-effect that correspond to the found analog windows, i.e. three versions of a forecast

Note that for processes with a systemic chaotic component, common quality measures can be unfeasible for control tasks solved by the CPS. For example, the excessive similarity of the forecast to the systemic component can lead to type II statistical errors (“false alarms”) [17, 18]. This is why, for many practical tasks, simplified measures such as the frequency of correct trend estimation (i.e. the direction of process evolution) can be employed.

Figure 4 shows the evolution of the real process, three versions of its forecast estimates, which were obtained as after-effects of three analogs with the highest degree of similarity, and their trends that are linear approximations of forecasts. It is easy to see that the trend directions correspond to the dynamics of the process in all three forecast versions. At the same time, the second version’s forecast significantly differs in shape from the real process.

Further improvement in the quality of forecasting can be achieved by using a system of distributed forecasters based on weak classifiers followed by joint processing.

A particular type of classifiers that could be employed is independent classifiers based on relations (3)–(5) that select the solution from the forecasting scheme with the most accurate prediction on the previous sliding observation segment. This scheme will allow us to obtain satisfactory results for a specific class of applied problems. For example, forecasting the current temperature in a chemical reactor used in the



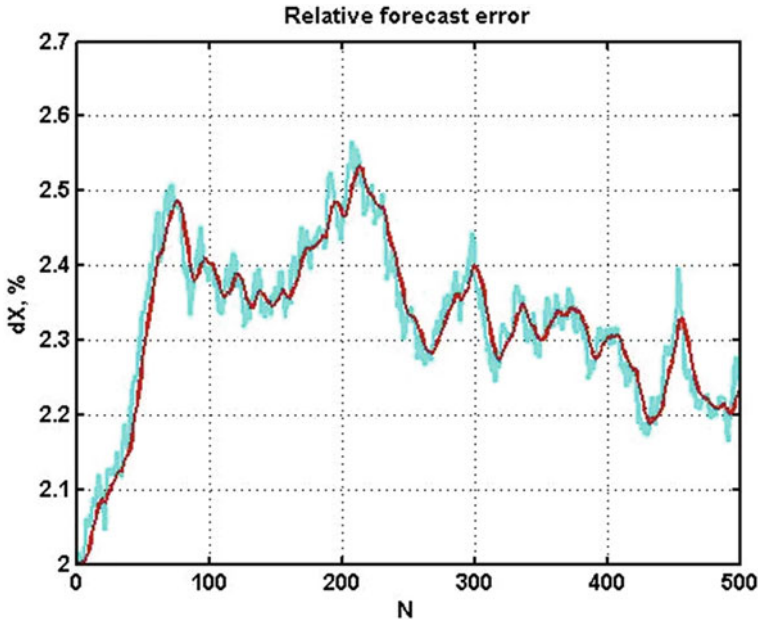
**Fig. 4** The process and three versions of the forecast that correspond to the analog windows augmented with their trends

production of aromatic hydrocarbons can have an error, as can be seen in Fig. 5, of just several percent. The scheme that was preferred most often is the one with three nearest neighbors, analog pruning by the level of similarity  $<50\%$ , and exponential smoothing with first-order astaticism and transfer coefficient  $\alpha = 0.02$ .

Note that the accuracy of the forecast will significantly depend on the size of the prediction interval. It is obvious that decreasing its size will allow increasing its accuracy. For the considered example, shrinking down the interval from 300 to 150 counts decreases the relative prediction error by approximately 15%. However, this is not always feasible due to the specifics of the considered task. For example, in a proactive management system for technological forecasting, the number of parameters of the output product cannot be smaller than the time needed to convert the raw material in its current state into the product.

## 5 Comparison of Precedent Forecast Implementations

For the considered observation model (1) with a chaotic systemic component, the quality of various implementations can only be compared via computational experiments. Consider the following forecast implementations:



**Fig. 5** An example of change in relative accuracy in forecasting a parameter of a technological process with non-stationary dynamics

1. Precedent prediction using the nearest neighbor technique (3);
2. Three-analog prediction based on the *k*-nearest neighbors (*kNN*) technique (4);
3. Three-analog prediction based on weighted *kNN* (5);
4. Three-analog prediction based on *kNN* with a restriction on the degree of similarity to the precedent.

To evaluate the efficiency of the forecast, we use standard deviation  $K_1$  :  $Err = (\frac{1}{N} \sum_{i=1}^N (\tilde{Y}_{t, t+L} - Y_{t+L})^2)^{1/2}$  and mean absolute deviation  $K_2$  :  $Err = \frac{1}{N} \sum_{i=1}^N abs(\tilde{Y}_{t, t+L} - Y_{t+L})$ . The values of these indicators for the aforementioned four precedent forecasts on an interval of 2888 counts are presented in Table 1.

It can be seen that a single precedent forecast has a lower accuracy than the average three-analog forecast. Weighting the analogs proportionally to their degree of similarity to the precedent barely affects the result in comparison to simple averaging. Additional analog pruning by the admissible similarity level  $\gamma_0 = 0.85$  did not increase the accuracy of the forecast.

The best result was obtained by the non-weighted averaging implementation. However, this conclusion appears to not be adequately substantiated in the case of chaotic influences. It is possible that in some situations limiting the degree of similarity of the analogs or weighting their after-effects could be significant. The degree to which a forecast is viable is determined by the specific control task. In general, an

**Table 1** The efficiency of the precedent forecast implementations on four observation intervals

Interval	Criterion	Forecast accuracy values			
		1	2	3	4
1	K <sub>1</sub>	34.09	23.89	24.57	27.70
	K <sub>2</sub>	46.38	32.46	31.24	35.14
2	K <sub>1</sub>	35.00	19.52	19.54	29.60
	K <sub>2</sub>	49.15	24.26	24.97	42.16
3	K <sub>1</sub>	34.88	23.23	22.77	25.09
	K <sub>2</sub>	45.66	31.38	31.97	35.86
4	K <sub>1</sub>	43.94	35.17	35.85	39.52
	K <sub>2</sub>	55.23	45.68	47.61	49.32

increase in forecast accuracy can be achieved by adapting the algorithms structure-wise and parameter-wise. The structure-wise adaptation consists of selecting the number of analog windows extracted from the retrospective observation database.

In the conditions of chaotic influences, which are common in unstable immersion environments, obtaining an accurate forecast is highly complicated. Within the conducted computational experiment that employed direct averaging by three after-effects of analog windows, the average forecast accuracy oscillates between 15 and 20%. However, as we have mentioned before, for many tasks, the excessive similarity of the forecast to the real process can not only be redundant but also lead to additional “false alarms” errors. Due to this, the accuracy requirements must be strictly coherent with the requirements of the hierarchically superior proactive management system.

The main merit of precedent forecasting is its ability to function in unstable immersion environments and obtain satisfactory results in situations when common extrapolation techniques lead to a divergent result. However, this merit can only be fully realized only when the database of retrospective observations is large enough, containing observation intervals that are similar to the current state vector of the observed system.

Effective forecasting of a chaotic process of a complicated inertia-less nature based on the considered computational schemes has not been achieved yet. This means that additional research, based on multidimensional statistical measures, is required.

## 6 Conclusion

The results obtained in the current chapter allow us to conclude that using precedent forecast algorithms is feasible in tasks of proactive management in complex unstable environments described by statistical and chaotic dynamic models. This conclusion is significant since most traditional algorithms of statistical forecasting produce divergent estimates. Nevertheless, it can be ill-advised to make generalized

conclusions on the viability of any algorithm for given conditions. The selection of the precedent forecasting technique should be tied to the specifics of the given applied control task. In particular, the applicability of the proposed techniques in inertia-less environments such as financial markets [19], requires additional research.

An important research direction for precedent forecasting is employing the metric approach for multidimensional environments. Using vector measures such as Euclidian distance, Mahalanobis distance, or measures based on Hotelling's  $t$ -squared statistic will allow increasing the quality of the forecast by taking into account relatively inert correlations. However, this approach loses spatial similarity used during analog window detection. This gives rise to a new research direction based on matrix similarity measures for multidimensional time series with a chaotic systemic component. We plan to focus our further research in the area of machine learning and forecast theory on this question.

**Acknowledgements** The research of Alexander Musaev described in this paper is partially supported is partially supported by the Russian Foundation for Basic Research (grant 20-08-01046), state research FFZF-2022-0004. Dmitry Grigoriev research for this paper was funded by a Support from The Endowment Fund of St Petersburg University. The authors are grateful to participants at the Center for Econometrics and Business Analytics (CEBA, St. Petersburg University) seminar series for helpful comments and suggestions.

## References

1. Broer, H., Takens, F.: *Dynamical Systems and Chaos*, T. 172. Springer Science & Business Media (2010)
2. Smith, L.: *Chaos: A Very Short Introduction*. OUP Oxford (2007)
3. Sivakumar, B.: *Chaos in Hydrology: Bridging Determinism and Stochasticity*. Springer (2016)
4. Li, M., et al.: Prediction of gas solubility in polymers by back propagation artificial neural network based on self-adaptive particle swarm optimization algorithm and chaos theory. *Fluid Phase Equilib.* **356**, 11–17 (2013)
5. Peters, E.: *Chaos and order in the capital markets: a new view of cycles, prices, and market volatility*, 2nd edn. Wiley, NY (1996)
6. Beck, C., Schögl, F.: *Thermodynamics of Chaotic Systems: An Introduction*, no. 4. Cambridge University Press (1995)
7. McCauley, J.L.: *Chaos, Dynamics, and Fractals: An Algorithmic Approach to Deterministic Chaos*, vol. 2. Cambridge University Press (1994)
8. Musaev, A.A., Borovinskaya, E.S.: Evolutionary optimization of case-based forecasting algorithms in chaotic environments. *Symmetry*. **13**(2), 301–317 (2021)
9. Kovshov, N.V., Moiseev, V.L., Ryazanov, V.V.: Algorithms for finding logical regularities in pattern recognition. *Comput. Math. Math. Phys.* **48**, 314–328 (2008)
10. Kalyuzhnaya, A., et al.: Precedent-based approach for the identification of deviant behavior in social media. In: *International Conference on Computational Science*. Springer, Cham (2018)
11. Dash, R., Swain, S.C.: A review on nearest-neighbor and support vector machine algorithms and its applications. *AI Manuf. Green Technol.: Methods Appl.* **83** (2020)
12. Abbasifard, M.R., Ghahremani, B., Naderi, B.H.: A survey on nearest neighbor search methods. *Int. J. Comput. Appl.* **95**(25), 39–52 (2014)



13. Buchanan, T.M., Buchanan, J., Kadey, K.R.: Predicting with your head, not your heart: forecasting errors and the impact of anticipated versus experienced elements of regret on well-being. *Motiv Emot* **43**, 971–984 (2019)
14. Abu Alfeilat, H.A., et al.: Effects of distance measure choice on k-nearest neighbor classifier performance: a review. *Big data* **7**(4), 221–248 (2019)
15. Yu, T., Zhu, H.: Hyper-parameter optimization: a review of algorithms and applications (2020). arXiv preprint [arXiv:2003.05689](https://arxiv.org/abs/2003.05689)
16. Feurer, M., Hutter, F.: Hyperparameter optimization. In: *Automated Machine Learning*, pp. 3–33. Springer, Cham (2019)
17. Naik, S.M., Jagannath, R.P.K., Kuppili, V.: Estimation of the smoothing parameter in probabilistic neural network using evolutionary algorithms. *Arab. J. Sci. Eng.* **45**, 2945–2955 (2020)
18. Pataky, T.C., et al.: Smoothing can systematically bias small samples of one-dimensional biomechanical continua. *J. Biomech.* **82**, 330–336 (2019)
19. Guegan, D.: Chaos in economics and finance. *Ann. Rev. Control.* **33**(1), 89–93 (2009)
20. Musaev, A., Grigoriev, D.: Numerical studies of statistical management decisions in conditions of stochastic chaos. *Math.* **10** (2) (2022)

# **Computer Vision and Image Processing**

# Application of Computer Vision Tools to Create a System for Monitoring the Work of Ground Equipment in Open Pits of Gold Mining Enterprises



Anna Alekhina , Alexandr Gurenko , and Mikhail Dorrer 

**Abstract** The main task of the proposed research is to organize automatic monitoring of equipment operation in an open pit of a gold mining enterprise. The solution to this problem allowed the management of the enterprise to reduce the risks of errors in accounting for the volume of work and improve the accuracy of planning the company's economic indicators. The problem is solved based on a computer vision system built based on a B2710RVZ street surveillance camera and an artificial neural network of YOLO v3 architecture, operating based on a Jetson Nano single-board computer. The neural network was trained on a personal computer with an Nvidia GeForce 1060 video card. The implemented system showed very high accuracy, correctly recognizing from 99 to 100% of technological loading operations. The result allows us to assert the possibility of automating the accounting and control of the operation of surface mining equipment using simple and reasonably cheap means based on computer vision technologies.

**Keywords** Production monitoring · Computer vision · Neural networks

## 1 Introduction

Since the beginning of work on implementing lean manufacturing in the global economy, there have been tremendous changes. Still, it is also true that for most operating organizations, the potential of Lean was practically not used [1]. Production systems in the world are intensively transformed. In the twenty-first century, enterprises and manufacturing companies are faced with entirely new generations of technologies, services, and products based on computer technology. To compete in global markets and ensure long-term success, companies need to adapt to shorter delivery

---

A. Alekhina (✉) · A. Gurenko  
Solution Factory, Krasnoy Armii, 10c3, Krasnoyarsk 660001, Russian Federation  
e-mail: [mdorrer@mail.ru](mailto:mdorrer@mail.ru)

M. Dorrer  
Reshetnev Siberian State University of Science and Technology, Krasnoyarsky Rabochy Av. 31,  
Krasnoyarsk 660037, Russian Federation

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022  
A. G. Kravets et al. (eds.), *Cyber-Physical Systems: Intelligent Models and Algorithms*, Studies in Systems, Decision and Control 417,  
[https://doi.org/10.1007/978-3-030-95116-0\\_17](https://doi.org/10.1007/978-3-030-95116-0_17)

203

times, increased product diversity, and high market volatility so that businesses can respond quickly and responsively to constant and unexpected changes. One of the main “cornerstones” for solving these problems is the introduction of digital information and communication technologies in production systems, processes, and technologies that allow introducing innovative developments by combining the physical world and fast access to data and their processing via the Internet (Industry 4.0) [2].

Since the beginning of work on the introduction of lean manufacturing in the world economy, there have been tremendous changes. At the same time, artificial intelligence technologies are actively used in various sectors of the economy to support the tasks of lean production. For example, [3] proposes a solution to improve the procurement and dispatch of drugs in a hospital using machine learning methods combined with traditional lean manufacturing methods. In a series of articles by the authors [4, 5], computer vision to improve the efficiency of retail trade is considered. In [6], a description of a system for automatic quality control of an object using computer vision and methods for processing digital images is presented. In [7], the authors proposed using artificial neural networks to implement the computer vision of an agricultural robot. The study [8] investigated the relationship between lean manufacturing, the digitalization of enterprises, and production efficiency. The study was based on data from a cross-sectional survey of manufacturing companies and was conducted using hierarchical multiple regression analysis.

The task of applying digital technologies in the mining industry is also relevant. Thus, in [9], it was suggested that industry 4.0 technologies could give a new impetus to productivity in the mining industry.

Thus, using artificial intelligence methods to improve the efficiency of mining is an urgent problem.

The sand beneficiation technology should ensure the maximum recovery of both coarse and fine gold [10].

The problem of monitoring the quality of work of people in the workplace is a hot topic. At first, special people were engaged in this; the current trend is to replace manual accounting with automated monitoring systems. For example, in [11], monitoring systems for ensuring safety in the construction industry are considered, and a five-stage safety system is being developed based on the identified key safety factors. A similar real-time security system is being developed to capture and identify unsafe behavior and hazards based on 2D images and video. The system records risk events, tries to identify the causes of a risk event, and proposes ways to prevent it [12]. Also, based on the analysis of the material of previous emergencies by images, videos, and biometric indicators, the solution proposed in [13] allows optimizing the monitoring systems for human health and safety. Such schemes allow for guaranteed real-time tracking of events affecting human safety and health in places where it is impossible to observe by the human observer. Such systems can detect objects, track and classify actions [14].

The purpose of this work was to automate the process of manual calculation by the foreman of the number of full buckets and the formation of a system for timely informing the state of landfills.

This chapter proposes an approach to monitoring the effectiveness of striping operations in mining using computer vision methods. At the present stage of development of the monitoring system, the authors set the following tasks:

1. To classify the condition of the front loader bucket on the video streams received from the cameras of the landfill.
2. Perform the timing of loading the soil onto the manipulator platform.
3. Calculate the number of soil loads on the manipulator platform per day.
4. Track the facts of downtime of the front loader at the landfill.

According to the test results, an utterly complete system within two weeks of continuous work at the test site showed 98% of correct answers. During testing, 288 h of video were processed. The configuration of the system allows the foreman to analyze the efficiency of a shift at several sites.

## 2 Methods and Materials

### 2.1 *Research Technology of Mining*

The design and layout of the flushing device for flushing gold-bearing sands considered in the chapter are based on the PGSh-50 flushing device. A bulldozer fed sands from the bottom of the face to the receiving table of the hydrocradle 1. After disintegration, the liquefied fraction of <60 mm entered the pressure chest. The oversize fraction (boulders, crushed stone, pebbles) was washed and dumped through the “gander” of the device into the dump. The pulp was fed to a deep filling lock 4 with a slope of 7–9°, and then passed through a screen with 20 mm inlet holes. The under-size fraction passed through a fine filling lock 5 with a slope of 7–9°. The airlock was reinforced with rubber mats, and the fine filling airlock was covered with metal stencils of the appropriate size and design.

The indicators of water consumption for the washing device were established due to numerous studies of the sand enrichment process, taking into account the experience and operation without hydraulic elevator washing plants, providing a rational technological mode of operation of the primary concentrators. The performance of the GMN-640 water monitor is 640 m<sup>3</sup>/h. The productivity of the washing device for sands is 40 m<sup>3</sup>/h, the flow rate of process water per one cubic meter of the original rock mass when processing gold-bearing sands in the washing device is 16 m<sup>3</sup>/h. A characteristic design feature of the device is the presence of a hydraulic screen of an original design. The material to be separated is accelerated with the formation of a pulsating pulp flow along the sluice chute with a bottom in the form of stencil components. In contrast, each part of the composite stencil is closed from above by a screen sheet with holes drilled in it. Each part of the stencil has an entrance slot between the screen sheet and the docking groove, smaller in clearance than the exit slot. The pulp under the screening sheet is accelerated in the stencil part and is brought to

the top of the screening sheet of the following part of the stencil, installed in series. The material to be separated and the resulting pulp flow at the sluice act as elements of a device for separating the pulp into a lower, heavier part, and also contribute to the release of particles of a valuable, heavier component into the holes of the screen sheet for their subsequent capture. As a result of washing with pulp intake through the inlet slot, particles of the valuable component accumulate in the cells of the rubber mat located under each part of the stencil. The slurry with a high content of a fine and fine valuable piece is periodically removed. In this case, the following technical result is achieved: a decrease in the total flow rate of the slurry during flushing with a simultaneous increase in the productivity of the flushing device.

Figure 1 shows a diagram of the processing of enriched soil on a flushing device.

This technology assumes optimal processing of about 200 m<sup>3</sup> of enriched soil per day in two shifts of workers.

The quality of the work performed by the loader and the operator in the shift is monitored by the head of the landfill (foreman). Currently, the number of full buckets brought to the washer is manually counted in the logbook. Such a control system is ineffective since there is a risk of providing inaccurate data by an unscrupulous supervisor at the quarry. For high-quality planning of the company's economic indicators, it was necessary to exclude the human factor in accounting for productivity by introducing an automated accounting system that does not depend on the performers. This approach provides operational control over the personnel of the mine and an increase in the accuracy of economic planning at a gold mining enterprise.

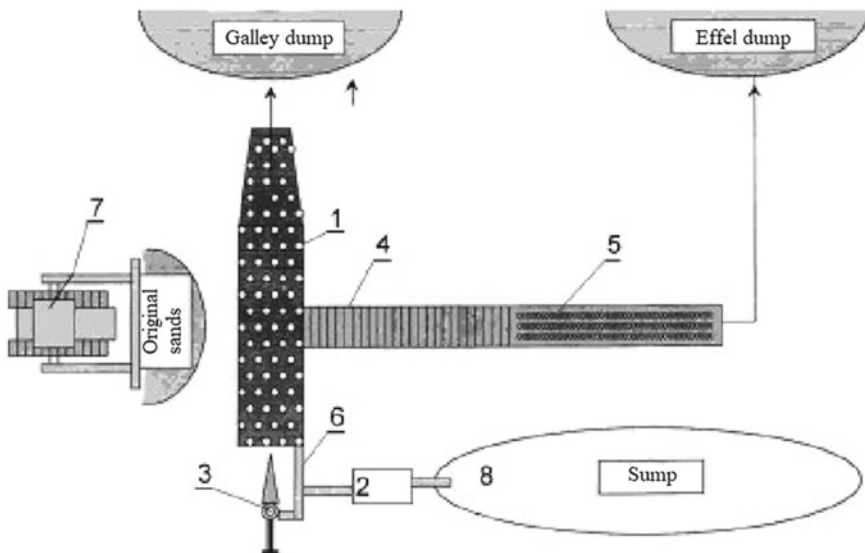


Fig. 1 Scheme of the work of the soil washing station at the landfill

## 2.2 Computer Vision Tasks

In the tasks of computer vision, the following sections are distinguished [15]:

1. Classification—classification of an image by the type of object it contains.
2. Object detection—detection of all objects of the specified classes and the definition of an encompassing frame for each of them.
3. Semantic segmentation—definition of all pixels of objects of a specific class or background in the image. If several things of the same class overlap, their pixels are not separated from each other in any way.
4. Instance segmentation—definition of pixels belonging to each object of each class separately.

Figure 2 shows an example of the results of solving computer vision problems.

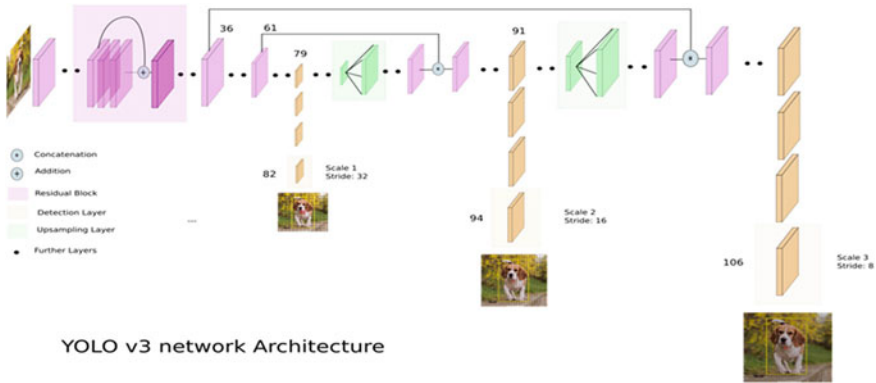
### YOLOv3

YOLOv3 was used as the basic architecture of the artificial intelligence of the artificial vision system. YOLO or You Only Look Once [16] is a currently top-rated convolutional network (CNN) architecture that is used to recognize multiple objects in an image. Figure 3 shows the YOLOv3 architecture.

The advantage of this approach is that the network analyzes the entire image at once and considers the context when identifying and recognizing an object. Also,



**Fig. 2** An example of performing computer vision tasks, where **a**—the results of the classification task, **b**—the results of object detection, **c**—the results of the semantic segmentation task, **d**—the results of the instance segmentation task



YOLO v3 network Architecture

Fig. 3 Visual graph of YOLOv3 architecture

YOLO is 1000 times faster than R-CNN architecture and about  $100 \times$  faster than Fast R-CNN.

YOLOv3 is an enhanced version of the YOLO architecture. It consists of 106 convolutional layers and is better at detecting small objects than its predecessor YOLOv2. The main feature of YOLOv3 is that at the output, there are three layers, each of which is designed to detect objects of different sizes.

In YOLO v3, detection is carried out by applying  $1 \times 1$  detection kernels on maps of objects of three different sizes at three different points on the network. Detection core shape  $-1 \times 1 \times (B \times (5 + C))$ . The formula determines the number of bounding rectangles that a cell in a feature map can parse. A score of “5” corresponds to 4 bounding box attributes and the validity of a single object, and a score C defines the number of classes. In YOLO v3, trained on the Microsoft COCO dataset [17],  $B = 3$  and  $C = 80$ , so the kernel size is  $1 \times 1 \times 255$ . The function map generated by this kernel has the same height and width as the previous function map and has depth attributes as described above. Detection at different levels helps to identify small objects, a common problem on YOLO v2. Up-sampling layers, combined with previous layers, help preserve fine-grained objects, that is, detect small objects. The  $13 \times 13$  layer detects large objects, the  $52 \times 52$  layer detects smaller objects, and the  $26 \times 26$  layer detects medium-sized objects.

For an input image of the same size, YOLOv3 predicts more bounding boxes than YOLOv2. For example, with an original resolution of  $416 \times 416$ , YOLOv2 predicted  $13 \times 13 \times 5 = 845$  image blocks. 5 blocks were found in each cell of the grid. On the other hand, YOLOv3 indicates blocks at 3 different scales. For the same  $416 \times 416$  image, the number of predicted blocks is 10,647. This means that YOLOv3 expects 10 times more blocks than YOLOv2. At each scale, each grid can predict 3 blocks using 3 anchor zones. Since there are three scales, the total number of anchoring zones used is 9.3 for each plate.

Algorithm optimization solutions helped improve the quality in YOLOv3 versus YOLOv2 without losing definition quality.



**Region of Interest (ROI)**

The Region of interest (ROI) metric [18, 19] represents patterns in a dataset for a specific purpose. The ROI concept is commonly used in many application areas. In geographic information systems (GIS), ROI can be thought of as a polygon sample on a 2D map. In computer vision and optical character recognition, ROI defines the boundaries of the object in question. In many applications, character (text) labels are added to the ROI to describe its contents concisely. Individual points of interest (POIs) can be located within an area of interest.

ROI tasks use the following data markup principles:

1. 1D dataset: time or frequency interval on a function graph.
2. 2D dataset: boundaries of an object in the image.
3. 3D dataset: Contours or surfaces that outline an object (sometimes known as a volume of interest (VOI)) in a volume.
4. 4D dataset: the contour of an object at a specific time interval or during a specific time interval in a spatial volume.

**Assessment Metrics**

The MAP Ranking Quality Score (Average Accuracy) is used to measure the accuracy of object detection. This is a critical parameter: average AP (average accuracy) [20] for each class from the training sample.

$$AP = \frac{1}{11} \sum_{r \in (0,01...1)}^0 P(r) \tag{1}$$

The main indicators of the assessment are accuracy and sensitivity, which are based on the adjacency matrix (Table 1).

- TP—truly positive, the classifier correctly attributed the object to the class under consideration;
- TN—truly negative, the classifier correctly states that the object does not belong to the class in question;
- FP—false positive, the classifier incorrectly attributed the object to the class in question;
- FN—false negative, the classifier incorrectly states that the object does not belong to the class in question.

Thus, based on the adjacency matrix, you can calculate the accuracy, which shows how accurate the predictions are: the percentage of correctly predicted classes (2)

**Table 1** Adjacency matrix as an example of parsing mAP

	Class A predicted: Yes	Class A predicted: No
Class A: Yes	True positive	False-positive
Class A: No	False-negative	True negative

and completeness, which is the fraction of objects belonging to the found class. Over the network concerning all objects of this class in example (3):

$$Precision = \frac{TP}{TP + FP}; \quad (2)$$

$$Recall = \frac{TP}{TP + FN}. \quad (3)$$

Another indicator is IoU (Intersection over Union), which allows you to measure the coverage of two boundaries: predicted by the classifier and the real one that frames the object in the image (4):

$$IoU = \frac{overlap}{union}. \quad (4)$$

The *F1* score can be interpreted as a weighted average of precision and recall, where the *F1* score reaches its best value at 1 and the worst value at 0. The relative contribution of accuracy and responsiveness to the *F1* score is. *F1* evaluation formula:

$$F1 = 2 * \frac{(precision * recall)}{(precision + recall)} \quad (5)$$

In the case of multiple grades and multiple marks, this is the average *F1* score of each grade, weighted according to the mean.

## 2.3 Materials

### Test Zone

Let us consider the description of the conditions of the test zone of the experiment. The climate of the Krasnoyarsk Territory in the Priisk region is sharply continental; most of the atmospheric precipitation occurs in July–August. The air temperature in summer reaches +25° C, and in winter, it drops to −35 ÷ −45 °C. Snow cover falls at the end of October and melts in mid-May. The flushing season is six months, and the 2020 season starts in May and ends in mid-November. The rivers rise in the first half of November and open up at the end of April.

The sands are low-clayey, characterized by good leaching, but they contain many boulders, the size of some reaches 2.5–3.0 m. Occasionally, lenses of permafrost are found.

Sands belong to the III and IV categories (80%—III, 20%—IV). Peat is represented by rocks of II–III strength category (70 and 30%, respectively). With a thickness of about 0.2 m, the soil-vegetation layer does not affect the sort of peat. The

hydrogeological conditions for the preparatory, stripping, and operational work are expected. In general, the placer belongs to the shallow open type. The underlying sands of the rock are attributed to the bed of the formation—the raft. The gold-bearing stratum is confined to the lower part of the alluvial deposits and captures the upper destroyed part of the bedrock. The productive layer is distinguished only by the gold content, having no other distinctive features. The gold-bearing layer is 70% located within alluvial and technogenic formations, 30%—in bedrocks (chlorite silicified shales). The mineral composition of the schlich complex contains scheelite, which is a constant companion of gold. In the upper reaches, there are galena, cinnabar, and basic bismuthite.

Based on the material composition of the sands (their washing capacity) of the deposit, the size of the gold contained in them, the planned productivity and profitability of the Arctic Technology Service LLC, as well as from the experience of working with the existing equipment, a washing device with a sluice sand processing technology was used to develop the placer.

### **Technical Characteristics of the Monitoring Device**

Next, we will consider the technical characteristics of the monitoring device. The NVIDIA Jetson Nano Developer Toolkit is a compact and powerful computer ideal for learning AI. The Jetson Nano paves the way for robotics and the deployment of deep learning algorithms to edge devices for real-time image classification, object detection, segmentation, speech processing, and more. It is an excellent tool for students, inventors, and developers to create their first AI project and learn how to work with popular machine learning frameworks such as PyTorch and TensorFlow [5, 16] (Fig. 4 and Table 2).

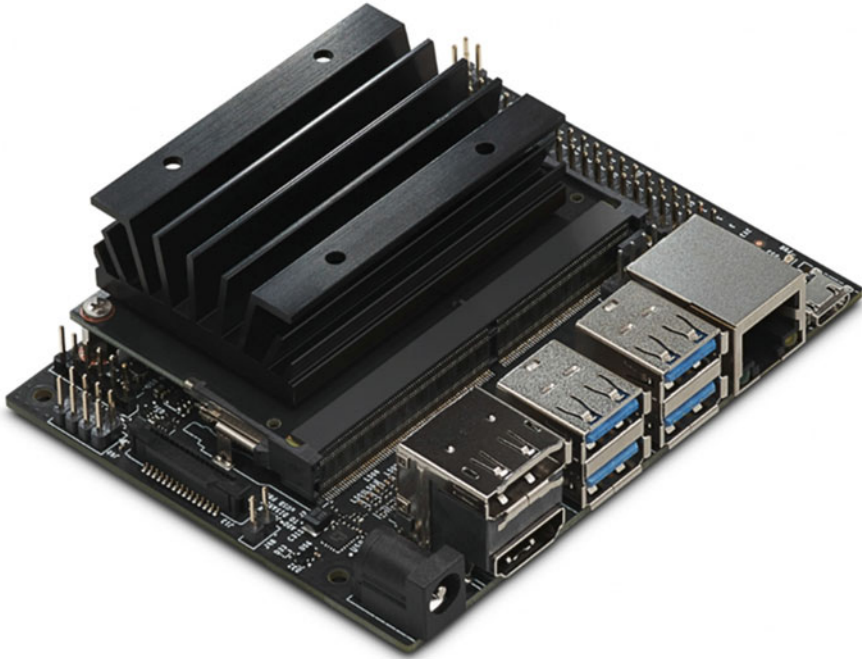
The B2710RVZ outdoor professional IP camera with a motorized lens is convenient for installation and operation; the 2MP SONY Exmor CMOS sensor with high sensitivity provides excellent clarity and correct color reproduction. The camera supports the “day/night” mode, has a built-in IR illumination and an electromechanical IR filter. Signal processing systems such as DWDR and 2D/3DNR have successfully improved observation performance in contrasting or low light conditions. The B2710RVZ supports various power connections, including PoE, to simplify the IP camera installation process and increase economic benefits.

In Fig. 5 shows the appearance of the camera B2710RVZ (Table 3).

Since the electricity is unstable in the working area and the device, voltage converters, and stabilizers were added.

### **Test Sampling of Images**

The training material was created from a field site in natural conditions in August. The complete recording of the educational video material took two full-fledged work shifts (day of work). The camera’s night mode was not helpful since, at night, two bright illumination sources worked at the landfill: a spotlight near the flushing device near the manipulator and spotlights of the front loader itself.



**Fig. 4** Jetson Nano single-board computer

**Table 2** NVIDIA Jetson Nano Developer Kit specifications

CPU	Quad-core ARM A57 @ 1.43 GHz
GPU	128-core Maxwell
RAM	4 GB 64-bit LPDDR4 25.6 GB/s

The original image resolution used during training is  $1920 \times 1080$ . No image pre-processing was used. The training sample for the first test model was 4000 hand-labeled images.

There are two classes in the model:

- Class “empty”—the front loader bucket is not complete or empty.
- Class “full”—the front loader bucket occupies approximately  $3.5 \text{ m}^3$  of volume.

Figure 6 shows an example of material markup for a model.

**Fig. 5** Outdoor surveillance camera B2710RVZ



**Table 3** Technical characteristics of the B2710RVZ camera

Parameter	Value
<i>General characteristics</i>	
Sensor	2 Mp, CMOS 1/2.8", Day/Night mode
Sensitivity	0.01 lx (day)/0.005 lx (night)/0 lx (backlit)
Lens	Motorized, 2.8–11 mm, f 1.4
Noise reduction	2d/3dnr/smartNR
<i>Video</i>	
Compression format	H.264 MP/BP
Video stream	H.264/MJPEG
Resolution	1920 × 1080, 1280 × 720, 960 × 528, 640 × 362, 480 × 256

### 3 Results

#### 3.1 Learning Outcomes of the Neural Network

The neural network was trained on a specially prepared darknet-53 weight map on an Nvidia GeForce 1060 video card. The training took about 8 h of computer time. A total of 3600 iterations were performed, making it possible to achieve an mAP value of 99.7% (see Fig. 7).



**Fig. 6** Example of marking material for counting front loader buckets, where the top row is daytime and the bottom row is nighttime

### 3.2 Accuracy Metric Results

The Table 4 shows the AP results of each class of the trained model and the number of false positives and true positives of the YOLOv3 model on the validation set.

The Table 5 shows the results of the calculated parameters of the metrics of the trained model on the validation set.

Based on the results obtained, it can be understood that the trained network has trained very well, showing the precision and recall values of 0.99 and 1.0, respectively. Such a result on a similar sample allows the model to be used as an expert for evaluating and counting objects in the monitoring system being developed.

### 3.3 Results of Model Operation on the Test Site

Figure 8 shows the analyzed data from a working gold mine.

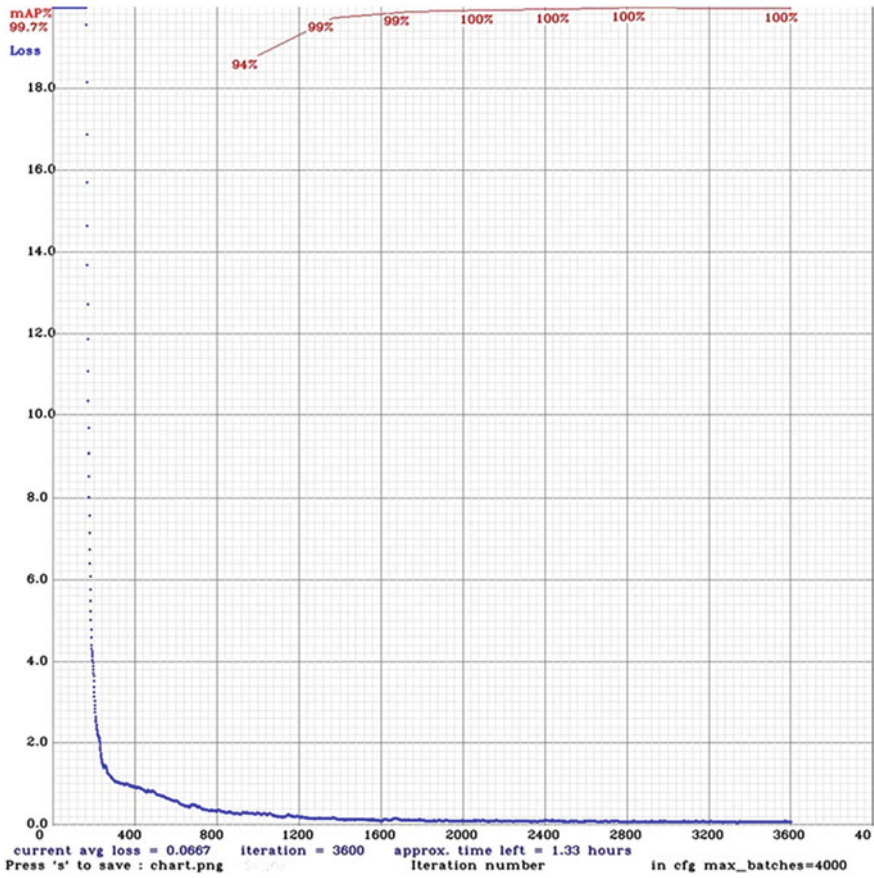


Fig. 7 Training graph of the YOLOv3 model

Table 4 AP of each class in the validation set

Name of class	AP%	TP	FP
Empty	99.65	1934	17
Full	99.94	3810	19

Table 5 Results of the obtained model metrics on the validation set

Precision	Recall	F1-score	TP	FP	FN	IoU%	mAP
0.99	1.0	0.99	5744	36	22	83.97	99.80



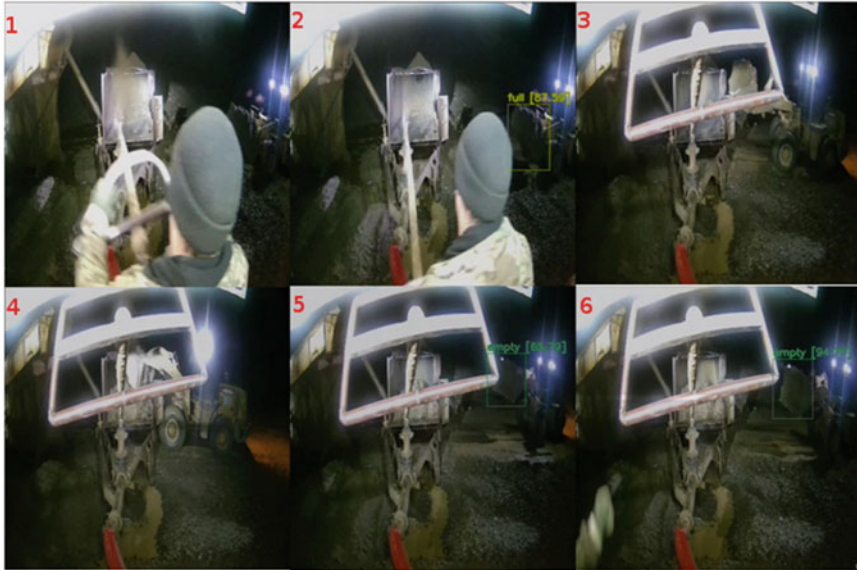
Fig. 8 An example of the work of Yolov3 at the working site of a gold mining company

After a two-week round-the-clock operation of the collection system for monitoring the loading of full buckets, no failures were noticed in the electronics of the computer vision system; the system also successfully survived emergency stops of the electric motor without losing the accumulated data at the time of the emergency stop. During the entire operation of the model, situations were recorded in which the system lost the loader or found it out of time.

Figure 9 shows a series of frames of one typical model lost load.

As you can see, because the camera is installed very close to the manipulator, there is a physical overlap of the view of the bucket and the moment of class change is missed by the monitoring system. It was decided to process several such omissions and supplement the training sample to reduce the error of the monitoring system in such situations. To optimize the network, the foreman manually marks the missed loads and thus generates additional training material, which will further improve the network results.





**Fig. 9** A sequential row of frames by a lost model by loading, where 1–6 is a sequence of frames

## 4 Discussion

Based on the results of the preparation of a test version of monitoring the counting and quality control of ore loading based on deep convolutional neural networks on the YOLOv3 architecture, it allows you to deploy the system quickly and, within one iteration, adjust the model to the abnormal position of the camera and thereby automate the control task.

In the future, it is planned to optimize the model by adding utility classes for identifying and alerting unstable situations, such as physical obstruction of the camera's view by people or individual parts of equipment.

The implemented solution allows the timing of surface mining equipment in open pits with minimal effort of employees and high accuracy. This result will enable you to analyze the efficiency of work in the open pit and identify the reasons for the loss of productivity to develop measures to improve efficiency.

## References

1. Netland, T.H., Powell D.J.: The Routledge Companion to Lean Management. Routledge (2016)
2. Helmold, M.: Lean Management and Artificial Intelligence (AI), pp. 131–137 (2020)
3. Jordon, K., Dossou, P.-E., Junior, J.C.: Using lean manufacturing and machine learning for

- improving medicines procurement and dispatching in a hospital. *Procedia Manuf.* **38**, 1034–1041 (2019). <https://doi.org/10.1016/j.promfg.2020.01.189>
4. Dorrer M.G., Tolmacheva A.E.: Application of the Canny's operation and the template matching method to improve the training set in computer vision training's issue system. *J. Phys. Conf. Ser.* **1679**, 032089 (2020) <https://doi.org/10.1088/1742-6596/1679/3/032089>
  5. Tolmacheva A.E., Ogurtsov, D.A., Dorrer, M.G.: Justification for choosing a single-board hardware computing platform for a neural network performing image processing. In: *IOP Conference Series: Materials Science and Engineering* (2020)
  6. Vergara-Villegas, O.O., Cruz-Sánchez, V.G., de Jesús Ochoa-Domínguez H., et al: Automatic Product Quality Inspection Using Computer Vision Systems. In: *Lean Manufacturing in the Developing World*, pp. 135–156. Springer International Publishing, Cham (2014)
  7. Dorrer, M.G., Popov, A.A., Tolmacheva, A.E.: Building an artificial vision system of an agricultural robot based on the DarkNet system. *IOP Conf. Ser. Earth Environ. Sci.* **548**, 032032 (2020). <https://doi.org/10.1088/1755-1315/548/3/032032>
  8. Buer, S.-V., Semini, M., Strandhagen, J.O., Sgarbossa, F.: The complementary effect of lean manufacturing and digitalisation on operational performance. *Int. J. Prod. Res.* **59**, 1976–1992 (2021). <https://doi.org/10.1080/00207543.2020.1790684>
  9. Humphreys, D.: Mining productivity and the fourth industrial revolution. *Miner. Econ.* **33**, 115–125 (2020). <https://doi.org/10.1007/s13563-019-00172-9>
  10. Council, N.R.: *Evolutionary and Revolutionary Technologies for Mining*. The National Academies Press, Washington, DC (2002)
  11. Guo, B.H.W., Zou, Y., Fang, Y., et al.: Computer vision technologies for safety science and management in construction: a critical review and future research directions. *Saf. Sci.* **135**, 105130 (2021). <https://doi.org/10.1016/j.ssci.2020.105130>
  12. Fang, W., Love, P.E.D., Luo, H., Ding, L.: Computer vision for behaviour-based safety in construction: a review and future directions. *Adv. Eng. Inform.* **43**, 100980 (2020). <https://doi.org/10.1016/j.aei.2019.100980>
  13. Zhang, M., Shi, R., Yang, Z.: A critical review of vision-based occupational health and safety monitoring of construction site workers. *Saf. Sci.* **126**, 104658 (2020). <https://doi.org/10.1016/j.ssci.2020.104658>
  14. Seo, J., Han, S., Lee, S., Kim, H.: Computer vision techniques for construction safety and health monitoring. *Adv. Eng. Inform.* **29**, 239–251 (2015). <https://doi.org/10.1016/j.aei.2015.02.001>
  15. Sonka, M., Hlavac, V., Boyle, R.: *Image Processing, Analysis and Machine Vision*. Springer US, Boston, MA (1993)
  16. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection (2015)
  17. COCO dataset. <https://cocodataset.org/#home>
  18. Kaare, K.K., Otto, T.: Smart health care monitoring technologies to improve employee performance in manufacturing. *Procedia Eng* **100**, 826–833 (2015). <https://doi.org/10.1016/j.proeng.2015.01.437>
  19. Li, T., Zhang, K., Li, W., Huang, Q.: Research on ROI algorithm of ship image based on improved YOLO. In: *2019 International Conference on Artificial Intelligence and Advanced Manufacturing (AIAM)*, pp 130–133. IEEE (2019)
  20. Van Gansbeke, W., Vandenhende, S., Georgoulis, S., Van Gool, L.: Unsupervised semantic segmentation by contrasting object mask proposals (2021)

# Framework for Biometric User Authentication Based on a Dynamic Handwritten Signature



Igor Anikin  and Ellina Anisimova 

**Abstract** We developed the framework for biometric user authentication based on dynamic handwritten signature recognition using a graphics tablet. The framework was developed in the Scilab mathematical package with integration with C++ development environment. We considered the definition of the dynamic handwritten signature by the set of discrete functions (values of signals) received through different channels. We developed the mathematical models and algorithms for signature receiving, training, and recognition modules, taking into account the fuzzy nature of signatures and the presence of possible fakes. We presented the results of experimental studies of suggested models and algorithms. We determined the best parameters to get the minimum error rate. FAR and EER error rate values were considered. It was found that the recognition accuracy of the developed framework exceeds the accuracy of other well-known methods. The proposed framework can be used as a component of public service delivery, as well as in automated document flow systems.

**Keywords** Biometric authentication · Dynamic handwriting signature · Classification · Machine learning

## 1 Introduction

Biometric user authentication systems are being actively implemented in various spheres of human activity: healthcare, education, business, finance, etc. [1–9]. Such interest is caused by their high accuracy, reliability, ease of use, the impossibility of theft or falsification of human biometric parameters. The most obvious trend in the modern world is the deep penetration of biometric technologies into various

---

I. Anikin (✉)

Kazan National Research Technical University Named After A.N. Tupolev-KAI, 10 K.Marx str,  
Kazan 420111, Russia  
e-mail: [IVAnikin@kai.ru](mailto:IVAnikin@kai.ru)

E. Anisimova

Kazan Federal University, 89 Kazanskaya str, Elabuga 423600, Russia  
e-mail: [ESAnisimova@kpfu.ru](mailto:ESAnisimova@kpfu.ru)

IT products which provide state and municipal services, automated document flow, bank financial services [10]. A dynamic handwritten signature of a person has many advantages in such applications over other biometric characteristics. It can be considered as a replacement for the classic handwritten signature [11–14]. However, the use of dynamic handwritten signatures for biometric authentication is often complicated by their fuzzy character, dependence on the place and time of signing, the position of a person, and his functional state. Also, we have to deal with the presence of skillful fakes, often very similar to genuine samples [15, 16]. In this case, the development of new mathematical models, algorithms, and software for biometric authentication of a person based on his dynamic handwritten signature under the difficulties mentioned above is very actual.

We have developed the framework for biometric user authentication based on a dynamic handwritten signature, as well as mathematical models and algorithms for its modules: signature receiving, training, and recognition. They provide high accuracy of recognition of fuzzy dynamic handwritten signatures even with the presence of skillful fakes.

## 2 Background

A dynamic handwritten signature is generated using a graphics tablet. At the moment of entering the signature, the signal values (dynamic characteristics) are reading through various channels that determine the coordinates on the plane ( $X$ ,  $Y$ ), the force of pressing on the tablet ( $P$ ), the azimuth ( $\Gamma$ ), as well as the angle ( $H$ ) of the pen elevation. Let's take a closer look at each characteristic.

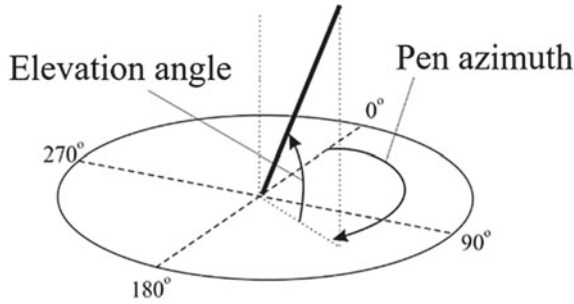
### 2.1 *Characteristics of a Dynamic Handwritten Signature*

The  $X$  ( $Y$ ) coordinate is the position (in mm) of the pen's tip relative to the positive semiaxis  $OX$  ( $OY$ ). Pressing force  $P$  is the amount of pressure (in Newton,  $N$ ) from the pen's tip to the tablet. Azimuth  $\Gamma$  is the angle between the direction of the semiaxis  $OY$  and the projection of the pen. The elevation angle  $H$  is defined as the angle between the pen and its projection (Fig. 1).

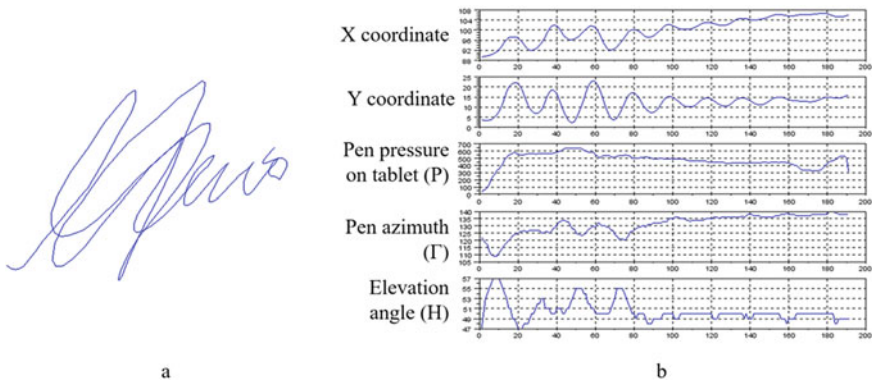
In Fig. 2 we can see the image of a dynamic handwritten signature and its corresponding dynamic characteristics.

### 2.2 *Basic Challenges*

Recognition of dynamic handwritten signatures is complicated by their fuzzy character and the existence of fakes.



**Fig. 1** The elevation angle and azimuth of the pen



**Fig. 2** Dynamic handwritten signature: **a**—signature image; **b**—dynamic characteristics

During biometric authentication, various types of dynamic handwritten signature fakes are possible (Fig. 3). In the case of random fakes, the forger does not know the name of the person whose signature he is forging, nor how his signature looks like. In the case of simple fakes, the forger knows only the username. In the case of skillful fakes, the forger knows the username and the required image of his signature. Existing of skillful fakes greatly complicates the process of recognition of dynamic handwritten signatures.

**Fig. 3** Genuine signature and its fakes: **a**—genuine signature; **b**—a random fake; **c**—a simple fake; **d**—skillful fake



**Fig. 4** The fuzzy character of dynamic handwritten signature images of the user

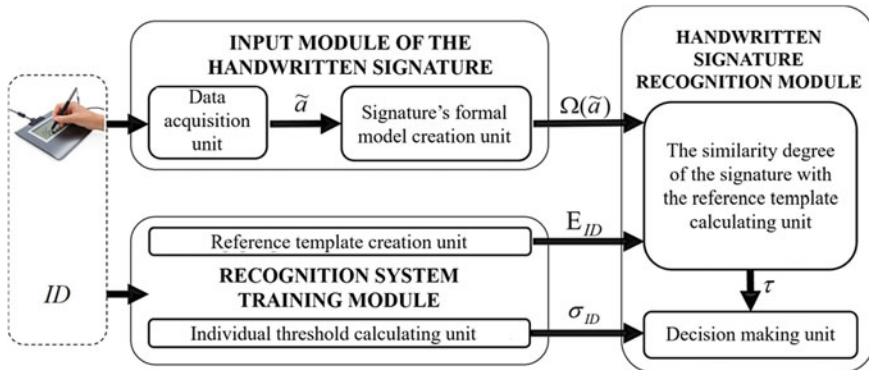


The fuzzy character of handwritten signatures can be associated with the variety of time, place, and position of a person, his psychological and physiological state, which undoubtedly needs to be taken into account in the biometric authentication process (Fig. 4).

Thus, to ensure high efficiency, the biometric user authentication system must be resistant to all kinds of signature fakes as well as to the fuzzy character of handwritten signatures.

### 3 The Architecture of the Developed Framework

We developed the framework for biometric user authentication based on dynamic handwritten signatures (Fig. 5). The framework was developed in the Scilab package with integration with C++ development environment.



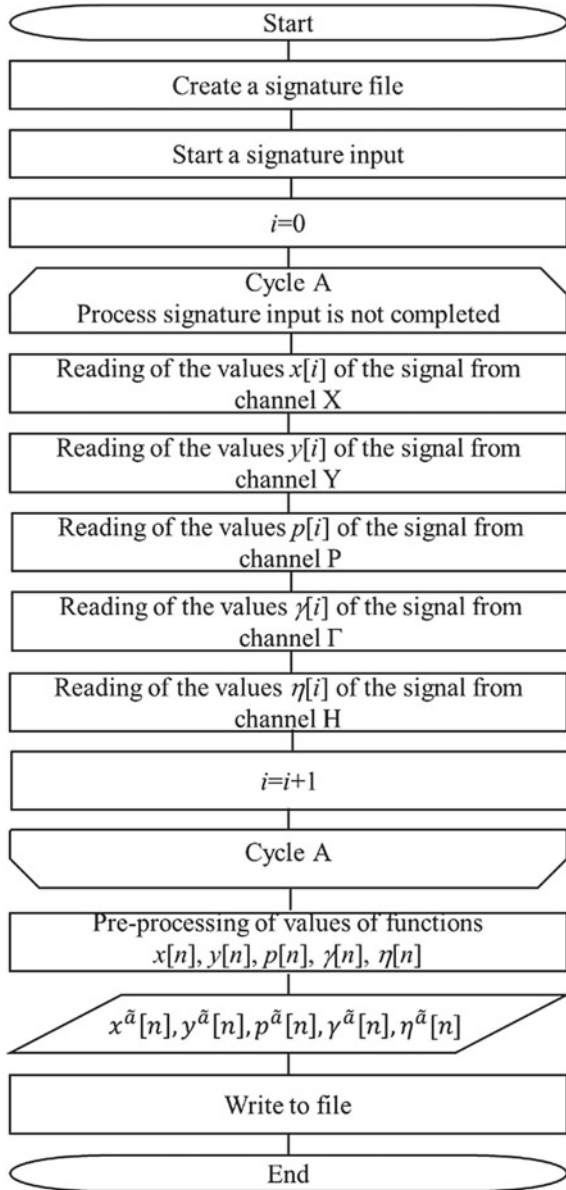
**Fig. 5** The architecture of the developed framework

The architecture of the framework includes the following basic modules.

(1) The input module includes the following basic units:

1.1. Data acquisition unit with the algorithm which is shown in Fig. 6. This unit generates a sequence of discrete functions values

**Fig. 6** Data acquisition unit diagram



$x^{\tilde{a}}[n], y^{\tilde{a}}[n], p^{\tilde{a}}[n], \gamma^{\tilde{a}}[n], \eta^{\tilde{a}}[n]$ , arriving through channels X, Y, P,  $\Gamma$ , H;

- 1.2. Signature’s formal model creation unit [17], which generates the signature model  $\Omega(\tilde{a}) = \{\xi_x^{\tilde{a}}, \xi_y^{\tilde{a}}, \xi_p^{\tilde{a}}, \xi_\gamma^{\tilde{a}}, \xi_\eta^{\tilde{a}}, Angl^{\tilde{a}}[n]\}$  (Fig. 7),

where  $\xi_f^{\tilde{a}} = \left( Sqr_f^{\tilde{a}}, FCLocMax_f^{\tilde{a}}, FuzInc_f^{\tilde{a}}, FuzDec_f^{\tilde{a}}, IncLeng_f^{\tilde{a}}, FuzDec_f^{\tilde{a}}, IncLeng_f^{\tilde{a}}, f^{\tilde{a}}[n], AcsI_f^{\tilde{a}}[n] \right)$ —feature description of a discrete function  $f^{\tilde{a}}[n]$  (Table 1).

The functions  $f^{\tilde{a}}[n]$  are  $x^{\tilde{a}}[n], y^{\tilde{a}}[n], p^{\tilde{a}}[n], \gamma^{\tilde{a}}[n], \eta^{\tilde{a}}[n]$ .

We proposed algorithms for the calculation of the bundle of features presented in Table 1.

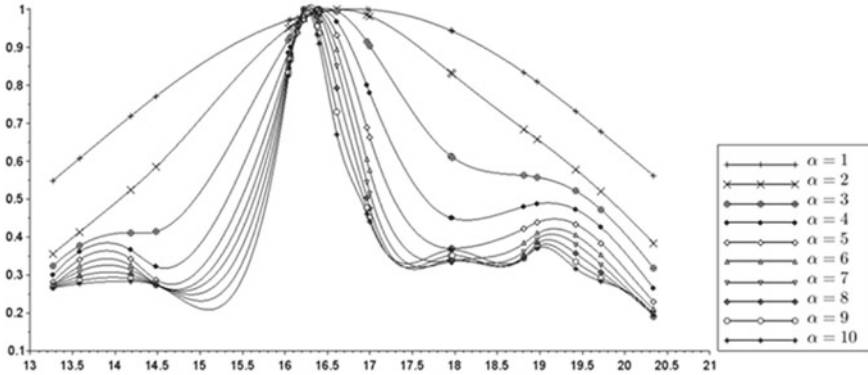
$Sqr_x^{\tilde{a}}$	$FCLocMax_x^{\tilde{a}}$	$FuzInc_x^{\tilde{a}}$	$FuzDec_x^{\tilde{a}}$	$IncLeng_x^{\tilde{a}}$	} Global features values
$Sqr_y^{\tilde{a}}$	$FCLocMax_y^{\tilde{a}}$	$FuzInc_y^{\tilde{a}}$	$FuzDec_y^{\tilde{a}}$	$IncLeng_y^{\tilde{a}}$	
$Sqr_p^{\tilde{a}}$	$FCLocMax_p^{\tilde{a}}$	$FuzInc_p^{\tilde{a}}$	$FuzDec_p^{\tilde{a}}$	$IncLeng_p^{\tilde{a}}$	
$Sqr_\gamma^{\tilde{a}}$	$FCLocMax_\gamma^{\tilde{a}}$	$FuzInc_\gamma^{\tilde{a}}$	$FuzDec_\gamma^{\tilde{a}}$	$IncLeng_\gamma^{\tilde{a}}$	
$Sqr_\eta^{\tilde{a}}$	$FCLocMax_\eta^{\tilde{a}}$	$FuzInc_\eta^{\tilde{a}}$	$FuzDec_\eta^{\tilde{a}}$	$IncLeng_\eta^{\tilde{a}}$	
$x^{\tilde{a}}[0]$	$x^{\tilde{a}}[1]$	...	$x^{\tilde{a}}[N-2]$	$x^{\tilde{a}}[N-1]$	$x^{\tilde{a}}[N]$
$AcsI_x^{\tilde{a}}[0]$	$AcsI_x^{\tilde{a}}[1]$	...	$AcsI_x^{\tilde{a}}[N-2]$	$AcsI_x^{\tilde{a}}[N-1]$	
$y^{\tilde{a}}[0]$	$y^{\tilde{a}}[1]$	...	$y^{\tilde{a}}[N-2]$	$y^{\tilde{a}}[N-1]$	$y^{\tilde{a}}[N]$
$AcsI_y^{\tilde{a}}[0]$	$AcsI_y^{\tilde{a}}[1]$	...	$AcsI_y^{\tilde{a}}[N-2]$	$AcsI_y^{\tilde{a}}[N-1]$	
$p^{\tilde{a}}[0]$	$p^{\tilde{a}}[1]$	...	$p^{\tilde{a}}[N-2]$	$p^{\tilde{a}}[N-1]$	$p^{\tilde{a}}[N]$
$AcsI_p^{\tilde{a}}[0]$	$AcsI_p^{\tilde{a}}[1]$	...	$AcsI_p^{\tilde{a}}[N-2]$	$AcsI_p^{\tilde{a}}[N-1]$	
$\gamma^{\tilde{a}}[0]$	$\gamma^{\tilde{a}}[1]$	...	$\gamma^{\tilde{a}}[N-2]$	$\gamma^{\tilde{a}}[N-1]$	$\gamma^{\tilde{a}}[N]$
$AcsI_\gamma^{\tilde{a}}[0]$	$AcsI_\gamma^{\tilde{a}}[1]$	...	$AcsI_\gamma^{\tilde{a}}[N-2]$	$AcsI_\gamma^{\tilde{a}}[N-1]$	
$\eta^{\tilde{a}}[0]$	$\eta^{\tilde{a}}[1]$	...	$\eta^{\tilde{a}}[N-2]$	$\eta^{\tilde{a}}[N-1]$	$\eta^{\tilde{a}}[N]$
$AcsI_\eta^{\tilde{a}}[0]$	$AcsI_\eta^{\tilde{a}}[1]$	...	$AcsI_\eta^{\tilde{a}}[N-2]$	$AcsI_\eta^{\tilde{a}}[N-1]$	
$AngI^{\tilde{a}}[0]$	$AngI^{\tilde{a}}[1]$	...	$AngI^{\tilde{a}}[N-2]$		

Fig. 7 The dynamic handwritten signature formal model

Table 1 Discrete function features for dynamic handwritten signatures

Feature designation	Feature description
$Sqr_f^{\tilde{a}}$	The area of the curvilinear region bounded by the piecewise linear envelope of the function $f^{\tilde{a}}[n]$ , the straight lines $n = 0, n = N$ , and the $n$ axis
$FCLocMax_f^{\tilde{a}}$	Fuzzy number of local maxima of the function $f^{\tilde{a}}[n]$
$FuzInc_f^{\tilde{a}}$	Evaluation of fuzzy increasing of the function $f^{\tilde{a}}[n]$
$FuzDec_f^{\tilde{a}}$	Evaluation of fuzzy decreasing of the function $f^{\tilde{a}}[n]$
$IncLeng_f^{\tilde{a}}$	The relative length of the sections of increasing of the function $f^{\tilde{a}}[n]$
$AcsI_f^{\tilde{a}}[n]$	The sequence of accelerations of the function $f^{\tilde{a}}[n]$
$AngI^{\tilde{a}}[n]$	A sequence of angle values between vectors connecting every three consecutive dots of a handwritten signature



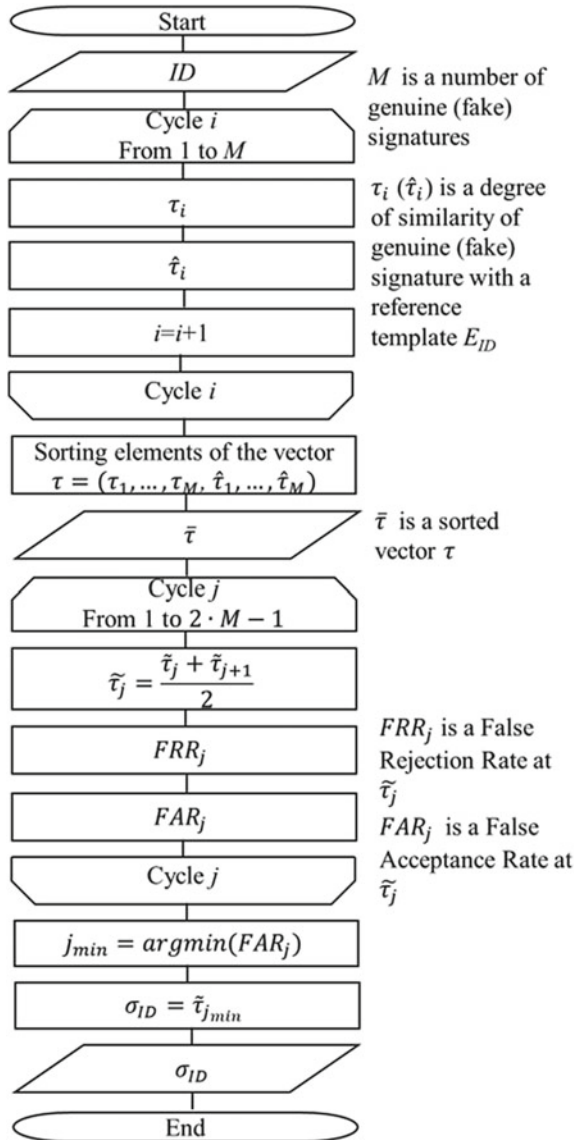


**Fig. 8** The example of membership functions of the  $FCLocMax_p$  feature of a dynamic handwritten signature with compactness values  $1 \leq \alpha \leq 10$  (OX—the characteristic value, OY—the values of the membership function)

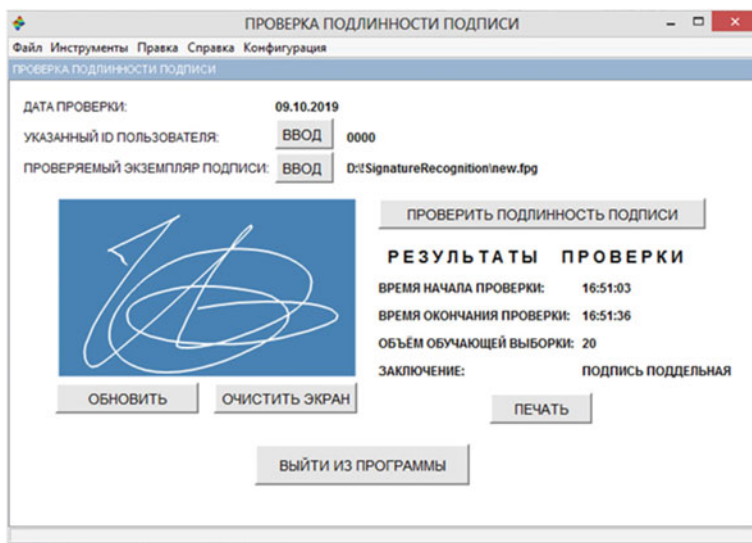
- (2) The recognition system training module includes the following main units:
  - 2.1. The reference template creation unit [18], which generates a reference template  $E_{ID}$  of a user with an identifier  $ID$  based on his previously obtained authentic signatures. The reference template is a set of membership functions (Fig. 8), constructed using the “Method of potentials” based on the values of features calculated for authentic signatures. The quality of the formation of a reference template is experimentally investigated depending on the degree of compactness, which is the main parameter of the “Method of potentials”.
  - 2.2. Individual threshold calculation unit (Fig. 9), which allows, based on genuine and fake samples of a user’s signature, to calculate the value of an individual threshold  $\int_{ID}$  which is used on the signature classification stage.
  
- (3) The handwritten signature recognition module includes the following main units:
  - 3.1. The unit which calculates the similarity  $\tau$  between the signature and reference template units. It works with the previously created formal models  $\Omega(\tilde{a})$ ;
  - 3.2. Decision-making unit which compares the similarity degree  $\tau$  with the value of the individual threshold  $\sigma_{ID}$ .

The user interface of the developed framework is shown in Fig. 10. User  $ID$  is requested to start biometric authentication. Next, the user enters a signature using a graphic tablet. Using the C++ development environment, a file with discrete signature function values is created. The signature image is displayed on the screen. After that, based on the created file, another file of the signature formal model is created.

**Fig. 9** Calculating the individual threshold



Further, based on the user’s training samples, a reference template is created and the value of an individual threshold is set. The degree of similarity of the user’s signature with the reference template is calculated. After that, a decision making about the signature’s validity is made by comparison of the degree of similarity with the individual threshold. Results are displayed on the screen.



**Fig. 10** The user interface

The developed framework can be used effectively used for user authentication in electronic document management systems and hybrid document management systems as well as in the financial sphere etc.

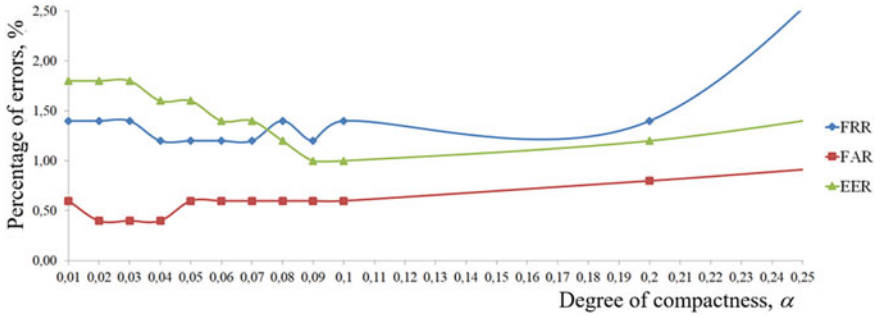
## 4 Experimental Research

We made a group of experiments that investigate the proposed models and algorithms in different situations. The experiments were carried out based on the MCYT\_Signature\_100 signature collection [19].

### 4.1 Selection of the Parameters

#### “Method of Potentials” the Degree of Compactness, $\alpha$

The degree of compactness affects the quality of the formation of the user’s reference template, and, consequently, the recognition accuracy. We have got that the smallest value of Equal Error Rate (EER), is achieved when forming membership functions using the “Method of potentials” with a degree of compactness  $\alpha = 0.09$  (Fig. 11). The smallest value of the False Acceptance Rate (FAR) with the presence of skillful fakes, is achieved when  $\alpha = 0.04$ .



**Fig. 11** Dependence of the recognition accuracy on the degree of compactness,  $\alpha$

**Table 2** Classification accuracy for different membership functions

The volume of the training sample	FAR, %, Gaussian MF	FAR, %, triangular/trapezoidal MF
5	0.45	0.75
10	0.40	0.60
20	0.40	0.60

### Membership Functions

To calculate fuzzy features ( $FCLocmax_f$ ,  $FuzInc_f$ ,  $FuzDec_f$ ,  $IncLeng_f$ ) it is necessary to specify the form of membership functions (MF). We evaluated the classification accuracy for Gaussian/triangular/trapezoidal membership functions. Evaluation results showed the preference for using Gaussian membership functions (Table 2).

### The Rational Features Set

We evaluated the classification accuracy for different sets of parameters that have been used for classification. We used the genetic algorithm for the detection of the rational features set, minimizing FAR. In Table 3 we can see the rational set of the parameters with the FAR = 0.20% and the volume of the training samples = 20.

We also determined the rational set of features (Table 4) with minimal EER = 0.36%.

**Table 3** Rational set of features minimizing FAR

Discrete function	Rational feature set
$x[n]$	$FCLocmax_x, FuzInc_x, Acsl_x[n]$
$y[n]$	$Sqr_y, y[n], Acsl_y[n]$
$p[n]$	$FCLocMax_p, p[n], Acsl_p[n]$
$\gamma[n]$	$Sqr_\gamma, FuzDec_\gamma, IncLeng_\gamma, Acsl_\gamma[n]$
$\eta[n]$	$Sqr_\eta, FCLocMax_\eta, FuzDec_\eta, IncLeng_\eta, Angl[n]$

**Table 4** Rational set of features minimal EER

Discrete function	Rational feature set
$x[n]$	$FuzInc_x, IncLeng_x, Acsl_x[n]$
$y[n]$	$Sqr_y, FCLocMax_y, y[n], Acsl_y[n]$
$p[n]$	$Sqr_p, FCLocMax_p, p[n], Acsl_p[n]$
$\gamma[n]$	$Sqr_\gamma, FCLocMax_\gamma, IncLeng_\gamma, \gamma[n], Acsl_\gamma[n]$
$\eta[n]$	$FCLocMax_\eta, FuzInc_\eta, FuzDec_\eta, \eta[n]$ $Angl[n]$

**Table 5** Comparison with other algorithms

Method	EER	The volume of the training sample
A method for recognizing handwritten signatures based on histogram features [20]	4.02	5
Recognition method based on Fourier descriptors [21]	12.11	No information
Recognition method based on hidden Markov models with signature template protection [22]	10.29	5
Recognition method based on the symbolic representation of the signature [23]	3.80	20
<b>The proposed approach</b>	<b>0.36</b>	<b>20</b>
	<b>1.30</b>	<b>5</b>

## 4.2 Comparison with Other Algorithms

We compared the accuracy of our approach with some other well-known methods (Table 5). We have got better results for different volumes of the training samples.

## 5 Conclusion

We have developed the framework for biometric user authentication based on a dynamic handwritten signature, as well as mathematical models and algorithms for different modules: signature receiving, training, and recognition. The framework was developed in Scilab with integration with C++ development environment. Models and algorithms provide high accuracy of recognition of fuzzy dynamic handwritten signatures even with the presence of skillful fakes.

We made a group of experiments that investigate the proposed models and algorithms in different situations. The experiments were carried out based on the MCYT\_Signature\_100 signature collection. We selected the parameters' values that provided the best recognition accuracy. The best compactness degree value

for “Method of potentials” was found. The Gaussian membership functions were selected. The rational set of features was determined. The recognition accuracy of our approach exceeds the accuracy of well-known modern methods.

## References

1. Wayman, J., Jain, A., Maltoni, D., Maio D.: An introduction to biometric authentication systems. In: Wayman, J., Jain, A., Maltoni, D., Maio, D. (eds.) *Biometric Systems*. Springer, London. (2005). [https://doi.org/10.1007/1-84628-064-8\\_1](https://doi.org/10.1007/1-84628-064-8_1)
2. Ahrabian, K., BabaAli, B.: Usage of autoencoders and Siamese networks for online handwritten signature verification. *Neural Comput. Appl.* **31**, 9321–9334 (2019)
3. Lopez-Lopez, E., Regueiro, C.V., Pardo, X.M., Franco, A., Lumini, A.: Towards a self-sufficient face verification system. *Expert Syst. Appl.* **174** (2021). <https://doi.org/10.1016/j.eswa.2021.114734>
4. Ismail, M., Memon, S., Dhomeja, L.D., Shah, S.M., Hussain, D., Rahim, S., Ali, I.: (2021) Development of a regional voice dataset and speaker classification based on machine learning. *J. Big Data* **8**(1), 2021 (2021). <https://doi.org/10.1186/s40537-021-00435-9>
5. Solano, J., Camacho, L., Correa, A., Deiro, C., Vargas, J., Ochoa, M.: Combining behavioral biometrics and session context analytics to enhance risk-based static authentication in web applications. *Int. J. Inform. Secur.* **20**(2), 181–197 (2021). <https://doi.org/10.1007/s10207-020-00510-x>
6. Rico-Juan, J.R., Inesta, J.M.: Confidence voting method ensemble applied to off-line signature verification. *Pattern Anal. Appl.* **15**, 113–120 (2021)
7. Cantoni, V., Dimov, D., Tistarelli, M. (Eds.): *Biometric authentication*. In: *First International Workshop, BIOMET 2014, Sofia, Bulgaria, June 23–24, 2014*, p. 265. Revised Selected Papers. Springer International Publishing (2014)
8. Dasgupta, D., Arunava, R., Nag, A. (Eds.): *Advances in User Authentication*, p. 360. Springer International Publishing (2017)
9. Zhou, J., Wang, Y., Sun, Z., Xu, Y., Shen, L., Feng, J., Shan, S., Qiao, Y., Guo, Z., Yu, S.: *Biometric recognition*. In: *12th Chinese Conference, CCBR, Shenzhen, China, October 28–29, Proceedings*, p. 762. Springer International Publishing (2017)
10. Lozhnikov, P.S., Sulavko, A.E.: Generation of a biometrically activated digital signature based on hybrid neural network algorithms. *J. Phys.* **1050** (2018)
11. Beresneva A., Epishkina A., Babkin S., Kurnev A., Lermontov V.: Handwritten signature verification: the state of the art. In: Samsonovich, A., Klimov, V. (eds.) *Biologically Inspired Cognitive Architectures (BICA) for young scientists. BICA 2017. Advances in Intelligent Systems and Computing*, vol. 636. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-63940-6\\_33](https://doi.org/10.1007/978-3-319-63940-6_33)
12. Calik, N., Kurban, O.C., Yilmaz, A.R.: Signature recognition application based on deep learning. In: *25th Signal Processing and Communications Applications Conference* (2017)
13. Choudhary, N.Y., Patil, R., Bhadade, U., Chaudhary, B.M.: Signature recognition & verification system using back propagation neural network. *Int. J. IT, Eng. Appl. Sci. Res.* **2**(1), 1–8 (2013)
14. Kudłacik, P., Porwik, P.: A new approach to signature recognition using the fuzzy method. *Pattern Anal. Appl.* **7**, 451–463 (2014)
15. Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Ortega-Garcia, J.: Biometric signature verification using recurrent neural networks. In: *Proceedings of the International Conference on Document Analysis and Recognition*, vol. 1, pp. 652–657 (2018)
16. Yang, W., Wang, S., Shahzad, M., Zhou, W.: A cancelable biometric authentication system based on feature-adaptive random projection. *J. Inform. Secur. Appl.* **58** (2021). <https://doi.org/10.1016/j.jisa.2020.102704>

17. Anikin, I.V., Anisimova, E.S.: Handwritten signature recognition method based on fuzzy logic. In: 2016 Dynamics of Systems, Mechanisms and Machines, Dynamics 2016 (2017). <https://doi.org/10.1109/Dynamics.2016.7818968>
18. Anisimova, E.S., Anikin, I.V.: Fuzzy sets theory approach for recognition handwritten signatures. In: Radionov, A.A., Gasiyarov, V.R. (eds.) Advances in Automation II. RusAutoConf 2020. Lecture Notes in Electrical Engineering, vol. 729. Springer, Cham (2021). [https://doi.org/10.1007/978-3-030-71119-1\\_93](https://doi.org/10.1007/978-3-030-71119-1_93)
19. Ortega-Garcia, J., Fierrez-Aguilar, J., Simon, D., Gonzalez, J., Faundez-Zanuy, M., Espinosa, V., Satue, A., Hernaez, A., Igarza, J., Vivaracho, C., Escudero, D., Moro, Q.-I.: MCVT baseline corpus: a bimodal biometric database. vision, image and signal processing. *IEEE Proc.* **150**(6), 395–401 (2003)
20. Sae-Bae, N., Memon, N.: A simple and effective method for online signature verification. In: BIOSIG 2013—Proceedings of the 12th International Conference of the Biometrics Special Interest Group, pp. 147–158 (2013)
21. Yanikoglu, B., Kholmatov, A.: Online signature verification using fourier descriptors. *EURASIP J. Adv. Sig. Process.* (2009). <https://doi.org/10.1155/2009/260516>
22. Maiorana, E., Martinez-Diaz, M., Campisi, P., Ortega-Garcia, J., Neri, A.: Template protection for HMM-based on-line signature authentication. In: CVPR Workshop, pp. 1–6 (2008)
23. Guru, D.S., Prakash, H.N.: Online signature verification and recognition: an approach based on symbolic representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**(6), 1059–1073 (2009)

# Image Compression Based on the Significance Analysis of the Wavelet Transform Coefficients Using the Energy Feature Model



Maya M. Lyasheva, Stella A. Lyasheva , and Mikhail P. Shleymovich 

**Abstract** Methods of image compression in cyber-physical systems based on computer vision technologies are considered. An approach to image compression is proposed by analyzing the significance of the detail coefficients of the wavelet transform. The approach is based on the use of a model of energy features. The model is a set of weight images of various levels, in which a certain weight is associated with each pixel. To obtain the weights, the relationship of the detailing coefficients of different levels is taken into account. These weights allow us to estimate the need to store the values of the corresponding coefficients of detail for subsequent image reconstruction. The chapter considers various ways of calculating weights when constructing a model of energy features. An image compression scheme based on the proposed approach is presented, the main stages of which are the implementation of the wavelet transform, the construction of a weight model, and the formation of significance maps of detailing coefficients at all levels. To assess the significance of the detail coefficients of the wavelet transform, it is proposed to use the threshold transform. Various ways for selecting threshold values of weights for performing the specified transformation are given. The possibility of improving the quality of image compression by adding elements of significance maps and values of significant detail coefficients of the wavelet transform to the general scheme of encoding operations is considered. The quality indicators of image compression and recovery obtained by applying the proposed approach are given.

**Keywords** Cyber-physical systems · Computer vision technologies · Image processing · Image analysis · Image compression · Image wavelet transform · Image energy features model

---

M. M. Lyasheva (✉) · S. A. Lyasheva · M. P. Shleymovich  
Kazan National Research Technical University Named After A. N. Tupolev-KAI, 10 K.Marx St.,  
Kazan, Tatarstan 420111, Russian Federation  
e-mail: [LyashevaMM@stud.kai.ru](mailto:LyashevaMM@stud.kai.ru)



## 1 Introduction

Currently, the processes of digitalization of the economy are actively developing, aimed, among other things, at improving the efficiency of production through its automation and the use of cyber-physical systems [1].

In modern cyber-physical systems, various technologies are used, which are based on models, methods, and tools of applied and system programming, high-performance computing, artificial intelligence, information protection from random and targeted influences, etc. [2–6].

Among the many technologies used in cyber-physical systems, computer vision technologies are often used, which are based on models, methods, and tools for processing and analyzing digital images [7–10]. The use of computer vision technologies allows you to automatically extract information from images of the surrounding environment and process it. This, in turn, ensures a high speed of solving functional problems by the corresponding systems.

Despite the high level of automation of information processing in systems based on computer vision technologies, as a rule, it is possible to provide images to end-users, i.e. decision-makers. In this case, it is necessary to ensure either the storage of images in storage devices or their transmission through communication channels. In any case, due to the fact that the images are data of significant volumes, it is advisable to apply compression procedures.

In practice, many approaches are used to compress images [11]. For example, methods based on discrete cosine transformation and vector quantization are used [12, 13]. Approaches based on various wavelet transformations are also being actively developed [14–16].

## 2 Wavelet Transform of Images

The wavelet transform allows you to represent the analyzed image in a multiscale form—as a set of copies of different resolutions [17, 18]. This representation allows for a detailed analysis of the characteristic features of the image. Its application is based on the theory of large-scale spaces, developed based on the results of studies of physical processes and biological vision. For example, research in the field of biological vision has shown that the receptive fields of mammalian visual cortex cells can be characterized as spatially localized, oriented, and structurally selective at various spatial scales [19].

The wavelet transform of an image is generally expressed as follows:

$$Wf(\mathbf{u}, s) = \int_{-\infty}^{+\infty} f(\mathbf{x})s^{-1}\psi * ((\mathbf{x} - \mathbf{u})/s)d\mathbf{x}, \quad (1)$$

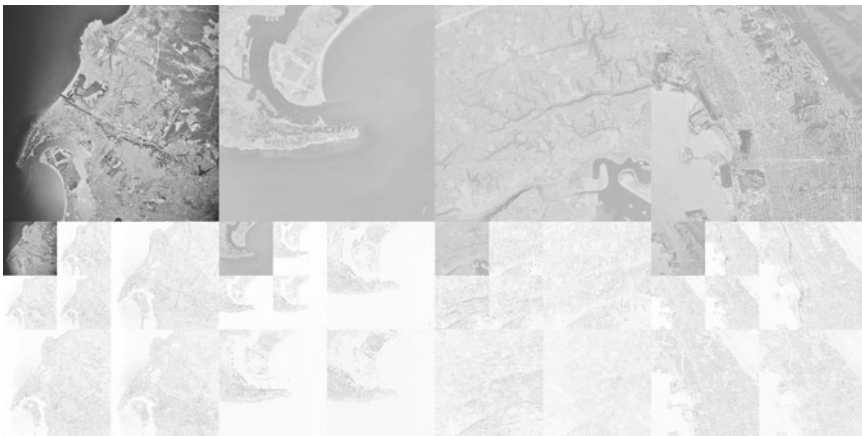
where  $Wf(\mathbf{u}, s)$  is the result of the transformation;  $f(\mathbf{x})$  is a two-dimensional function describing the image;  $\psi^*(\mathbf{x})$  is a complex conjugation of the shifted and scaled wavelet function  $\psi(\mathbf{x})$ ;  $\mathbf{u}$  is a two-dimensional vector of the shift parameters;  $s$  is a scale parameter.

For digital image processing, orthogonal discrete multiple-scale wavelet transformations are often used, the main advantage of which is the existence of fast algorithms for analysis and synthesis. In this case, the image is represented as a set of approximating coefficients of level  $j_0$  and detailing coefficients of levels  $j_0 + 1, j_0 + 2, \dots, J - 1$ , where  $J$  is the maximum possible number of decomposition levels (it is also assumed that  $J$  is the representation level of the original image).

Figure 1 shows an example of an orthogonal discrete multiple-scale wavelet transform for aerial images 2.2.01, 2.2.02, 2.2.03, and 2.2.04 in halftone form with dimensions of  $1024 \times 1024$  pixels taken from the UCS-SIPI collection [20]. When performing the transformation in this case,  $j_0$  and  $J$  are chosen to be 8 and 10, respectively, i.e. the transformation is two-level (the images are represented at scale levels 8 and 9).

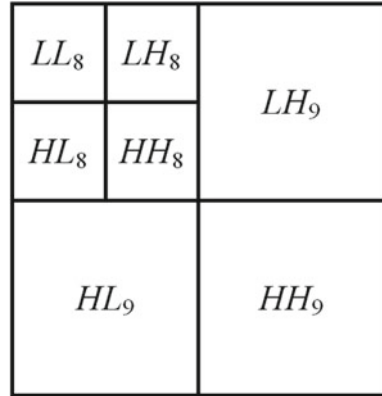
The image decomposition scheme is shown in Fig. 2, where  $LL_8$  is the matrix of approximating coefficients of the 8th level;  $LH_8, LH_9$  are the matrices of horizontal detailing coefficients of the 8th and 9th levels;  $HL_8, HL_9$  are the matrices of vertical detailing coefficients of the 8th and 9th levels;  $HH_8, HH_9$  are the matrices of diagonal detailing coefficients of the 8th and 9th levels.

As you can see in Fig. 1, the detail coefficients of each level show the pixels in which the brightness changes are observed when viewing images at the corresponding scale. The more brightness changes there are in a certain fragment in the image, the more information about the image is contained in that fragment. Thus, the analysis of the detailing coefficients allows us to identify the most significant elements of the image description in terms of the amount of information, and hence perception.



**Fig. 1** Examples of orthogonal discrete multiple-scale wavelet transform

**Fig. 2** The scheme of the wavelet transform



### 3 Model of Energy Features of the Image

The analysis of the detailed coefficients of the wavelet transform can be carried out on the basis of the energy feature model [21, 22]. This model is a set of weight images of various levels, in which each pixel is associated with a certain value (weight), reflecting the estimate of the contribution of the corresponding detailing coefficients to the total energy of the original image. In this case, to obtain the weight values, the relationship of the detailing coefficients of different levels is taken into account.

To obtain the weights of the detailing coefficients, various methods can be used, which are generally expressed as follows:

$$w_{j_0}(m, n) = F_{j_0}(k''_{j_0}, LH_{j_0}(m, n), HL_{j_0}(m, n), HH_{j_0}(m, n)), \tag{2}$$

$$w_j(m, n) = F_j(w_{j-1}(m, n), k'_j, k''_j, LH_j(m, n), HL_j(m, n), HH_j(m, n)), \tag{3}$$

where  $m, n$  are the indices of the detailing coefficients in the corresponding matrices;  $w_{j_0}(m, n), w_{j-1}(m, n), w_j(m, n)$  are the weights of the detailing coefficients at levels  $j_0, j - 1$ , and  $j$  at the position  $(m, n)$ ;  $F_{j_0}(\cdot), F_j(\cdot)$  are the functions for obtaining weights at levels  $j_0$  and  $j$ ;  $k''_{j_0}$  is the coefficient that takes into account the contribution of the detailing coefficients of level  $j_0$ ;  $k'_j, k''_j$  are the coefficients that take into account the contribution of the weight values of level  $j - 1$  and the detailing coefficients of level  $j$ ;  $LH_{j_0}, HL_{j_0}, HH_{j_0}, LH_j, HL_j, HH_j$  are the detailing coefficients at levels  $j_0$  and  $j$ ;  $j = j_0 + 1, \dots, J - 1$ .

Examples of ways to calculate weights:

Method 1:

$$w_{j_0}(m, n) = LH^2_{j_0}(m, n) + HL^2_{j_0}(m, n) + HH^2_{j_0}(m, n), \tag{4}$$

$$w_j(m, n) = w_{j-1}(m, n)/4 + LH_j^2(m, n) + HL_j^2(m, n) + HH_j^2(m, n); \quad (5)$$

Method 2:

$$w_{j_0}(m, n) = \sqrt{LH_{j_0}^2(m, n) + HL_{j_0}^2(m, n) + HH_{j_0}^2(m, n)}, \quad (6)$$

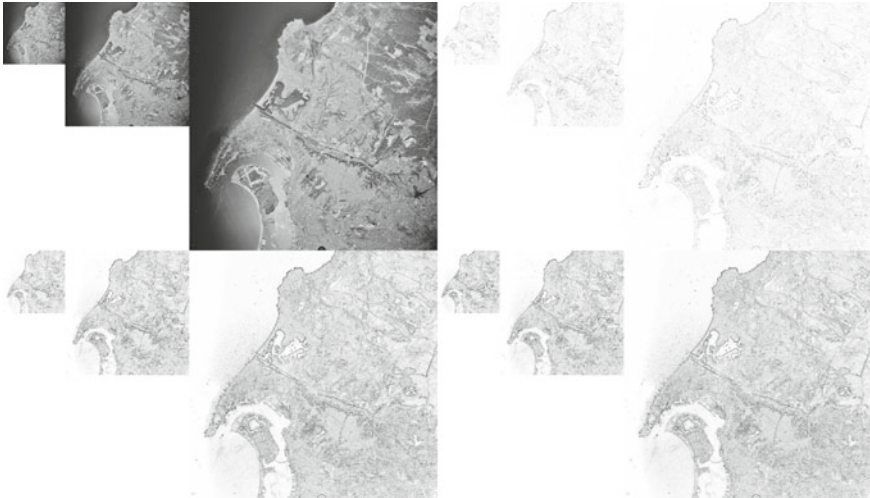
$$w_j(m, n) = w_{j-1}(m, n)/2 + \sqrt{LH_j^2(m, n) + HL_j^2(m, n) + HH_j^2(m, n)}; \quad (7)$$

Method 3:

$$w_{j_0}(m, n) = |LH_{j_0}| + |HL_{j_0}| + |HH_{j_0}|, \quad (8)$$

$$w_j(m, n) = w_{j-1}(m, n)/2 + |LH_j| + |HL_j| + |HH_j|. \quad (9)$$

Examples of weight models obtained by these methods for the same aerial image 2.2.01 from the UCS-SIPI collection in halftone form are shown in Fig. 3. Figure 3 shows the copies of the original image at levels 7–9 and the weights of the detail coefficients at levels 7–9, calculated by the specified methods. The value of the weight is reflected by the brightness—the darker the point, the greater the value of the corresponding weight. When considering the figure, it can be seen that, despite the differences, the weight models quite clearly reflect the nature of brightness changes in the original image at different levels of the wavelet transform.



**Fig. 3** Image weight models

## 4 Image Compression Based on the Energy Feature Model

The weights of the detail coefficients obtained by these methods allow us to assess the need to store their values for subsequent image reconstruction. Thus, you can apply the compression procedure, the general scheme of which consists of the following steps: define the size of the image; set the initial and final levels of the wavelet transform; perform a wavelet transform; build a weight model of the image; generate significance maps of the detail coefficients of all levels of the wavelet transform; save size of the image, the initial and final levels of the wavelet transform, matrices of approximating coefficients, significance maps, and matrices of detailing coefficients.

To reduce the size of the file in which the compressed information is stored, this procedure can be supplemented by encoding the data using some compression method, for example, the Huffman method, the arithmetic method, or the method of encoding the series lengths [23].

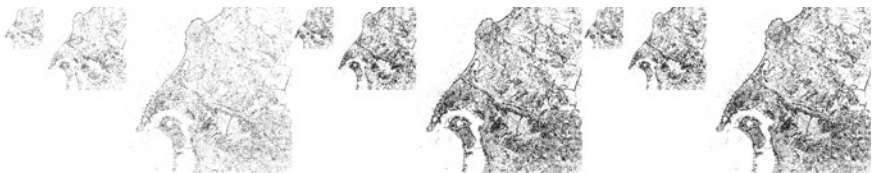
The key point when compressing an image according to the proposed scheme is to evaluate the significance of the detailing coefficients. One of the possible approaches here is based on performing a threshold transformation of weight images—images containing the weights of the detailing coefficients of various levels of the wavelet transform. The threshold transformation of weight images is defined as follows:

$$s_j(m, n) = \begin{cases} 0, & w_j(m, n) < t_j, \\ 1, & w_j(m, n) \geq t_j, \end{cases} \quad (10)$$

where  $s_j(m, n)$  is the result of the threshold transformation for the weight at the position  $(m, n)$  at level  $j$ ;  $w_j(m, n)$  is the weight of the detail coefficients at the position  $(m, n)$  at level  $j$ ;  $t_j$  is the threshold value of the weight of the detail coefficients at the level  $j$ .

Figure 4 shows examples of threshold transformations of weight images of levels 7, 8, 9 (from left to right) obtained by the three specified methods for the 2.2.01 aerial image from the UCS-SIPI collection. When visualizing the results of threshold transformations, the value 0 is reflected in white, and the value 1 is reflected in black.

As already mentioned, to improve the quality of compression, you can use data encoding, for example, elements of significance maps and detail coefficients. To perform this operation, you can use some entropy coding methods. In this work,



**Fig. 4** Examples of results of threshold transformations of weight images

when implementing the coding, adaptive arithmetic compression was applied independently for each data set. The choice of this method is due to the following considerations:

1. The method is single-pass, i.e. encoding is performed as the data is processed without performing the preliminary step of building a general statistical model;
2. The method allows you to provide an additional gain by taking into account statistics obtained only for a relatively small part of the data;
3. The method allows you to take into account the dependencies between data sequences.

When performing adaptive arithmetic coding, a table of the current character frequencies representing the processed data is constructed. The fewer characters that are considered at the same time when forming the next code, the shorter this code will be. Therefore, the encoded data can be further normalized.

In this work, the orthogonal multiple-scale Haar transform, which is time-efficient, was used to construct the weight models. As a result of this transformation, matrices of approximating and detailing coefficients are formed. The values of the approximating coefficients are in the range from 0 to 255, and the values of the detailing coefficients are in the range from  $-255$  to  $255$ . When coding, only significant detailing coefficients are considered. Therefore, there will be no significant loss of accuracy when they are quantized. In accordance with these considerations, the following normalization operation is proposed:

$$\tilde{d}_j(m, n) = \lfloor d_j(m, n)/2 \rfloor + 128, \quad (11)$$

where  $\tilde{d}_j(m, n)$ ,  $d_j(m, n)$  is the normalized and original detailing coefficient from the set of horizontal, vertical, or diagonal detailing coefficients of the level  $j$  of the wavelet transform at the position  $(m, n)$ ;  $\lfloor \cdot \rfloor$  is the operation of obtaining the largest integer not exceeding a given real number. After normalization, the values of the detail coefficients will be in the range from 0 to 255.

As a result of the proposed compression scheme, a file is formed, the volume of which is significantly less than the volume of the original image. To restore the image, you need to perform the reverse steps: read the image dimensions, the initial and final levels of the wavelet transform, the matrices of approximating coefficients, the significance maps, and the matrices of detailing coefficients; generate matrices of detailing coefficients containing zero elements; write the read values corresponding to the non-zero elements of the significance maps to the matrices of the detailing coefficients; perform the inverse wavelet transform.

The compression quality and recovery quality for the images in Fig. 1, obtained as a result of applying the proposed procedure using the specified methods for calculating weights based on the three-level Haar transform, normalization of the detailing coefficients, and adaptive arithmetic coding, are shown in Table 1. To evaluate the quality of image compression and recovery, the ratio of the volume of the original and compressed files and the peak signal-to-noise ratio, respectively, were used. To

**Table 1** Image compression and recovery quality using thresholds corresponding to the 90th, 50th, and 10th percentiles for the 9th, 8th, and 7th levels of the weight model

Images	Image compression quality/image recovery quality		
	Method 1	Method 2	Method 3
2.2.01	6,349/27,209	6,256/27,136	6,338/27,001
2.2.02	10,369/37,324	9,521/37,429	9,649/37,333
2.2.03	8,949/34,529	8,134/34,684	9,058/34,301
2.2.04	7,356/31,035	7,212/30,989	7,171/30,901

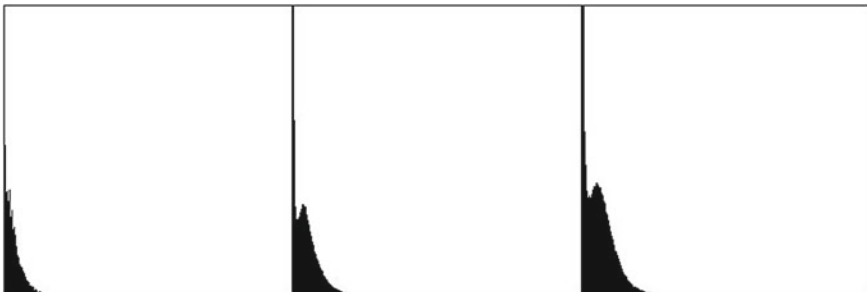
analyze the significance of the detailing coefficients, a threshold transformation was used with thresholds corresponding to the 90th, 50th, and 10th percentiles for the 9th, 8th, and 7th levels of the weight model.

From Table 1 it can be seen that the described approach to image compression provides average compression and recovery quality indicators of 8 and 32, respectively. At the same time, there is a possibility of flexible configuration by choosing the method of calculating weights, threshold values, the type of wavelet transform, and the number of decomposition levels.

Threshold values of weights for constructing significance maps of detail coefficients can be selected based on studies of the weight models of the set of images to be compressed. For such studies, you can build histograms of weights and evaluate the statistical characteristics necessary for the analysis. Examples of histograms of weights of levels 7–9 (from left to right) are shown in Fig. 5.

The process of calculating the threshold values of weights can be automated using adaptive methods. Tables 2, 3 and 4 show the weight thresholds obtained for levels 7–9 of the weight models of the 2.2.01 aerial image from the UCS-SIPI collection using the adaptive methods. In this case, the values of the weights were previously converted to the range from 0 to 255:

$$\tilde{w}_j(m, n) = 255 \times \frac{w_j(m, n)}{\max_j}, \tag{12}$$



**Fig. 5** Examples of image weight histograms

**Table 2** Threshold values for the weight model obtained by method 1

Methods	Threshold values		
	Level 7	Level 8	Level 9
Mean-based method	9	10	10
Median-based method	3	4	5
Method based on the minimum and maximum values	127	127	127
Method based on a histogram with a clipping factor of 0,6	5	7	7
Method based on sequential refinement with an error of 1	31	27	24
The Otsu method	34	31	27

**Table 3** Threshold values for the weight model obtained by method 2

Methods	Threshold values		
	Level 7	Level 8	Level 9
Mean-based method	33	37	39
Median-based method	24	31	35
Method based on the minimum and maximum values	127	127	127
Method based on a histogram with a clipping factor of 0,6	33	41	43
Method based on sequential refinement with an error of 1	46	47	43
The Otsu method	51	47	48

**Table 4** Threshold values for the weight model obtained by method 3

Methods	Threshold values		
	Level 7	Level 8	Level 9
Mean-based method	37	37	38
Median-based method	28	31	35
Method based on the minimum and maximum values	127	127	127
Method based on a histogram with a clipping factor of 0,6	38	40	43
Method based on sequential refinement with an error of 1	52	46	42
The Otsu method	52	49	45

where  $\tilde{w}_j(m, n)$ ,  $w_j(m, n)$  are the reduced and original weight values at the position  $(m, n)$  at the level  $j$  of the weight model, respectively;  $\max_j$  is the maximum weight value at the level  $j$  of the weight model. Of these methods, it is advisable to use the mean-based method, median-based method, or the Otsu method.

Tables 5, 6 and 7, show the values of the compression and recovery quality indicators for the images in Fig. 1, obtained as a result of applying the proposed procedure using the specified methods for calculating weights based on the three-level Haar transform, normalization of the detailing coefficients, and adaptive arithmetic coding. The ratio of the volume of the source and compressed files and the peak



**Table 5** Image compression and recovery using thresholds based on the mean

Images	Image compression quality/image recovery quality		
	Method 1	Method 2	Method 3
2.2.01	4,427/28,248	3,421/30,735	3,296/31,022
2.2.02	9,075/37,599	7,743/38,396	7,297/38,484
2.2.03	6,672/33,886	5,576/34,648	5,745/34,841
2.2.04	4,979/31,394	4,094/33,091	3,887/33,586

**Table 6** Image compression and recovery using thresholds based on the median

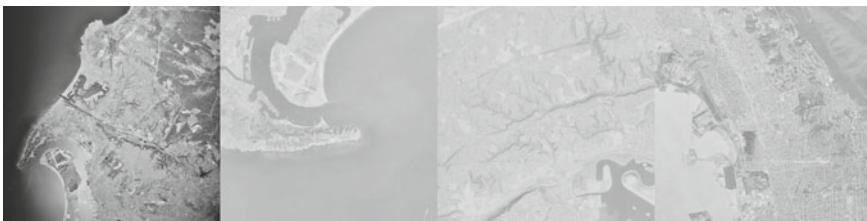
Images	Image compression quality/image recovery quality		
	Method 1	Method 2	Method 3
2.2.01	3,217/32,445	3,205/32,231	3,180/32,226
2.2.02	6,195/38,972	5,758/39,145	5,825/38,989
2.2.03	5,359/35,346	5,576/34,648	5,266/35,083
2.2.04	3,775/34,118	3,815/33,718	3,848/33,844

**Table 7** Image compression and recovery quality using Otsu thresholds

Images	Image compression quality/image recovery quality		
	Method 1	Method 2	Method 3
2.2.01	10,415/23,679	4,252/28,209	4,031/28,428
2.2.02	20,657/33,341	9,798/36,778	9,867/36,977
2.2.03	13,313/31,131	6,847/33,377	6,872/33,279
2.2.04	8,838/28,078	4,456/32,002	4,279/32,311

signal-to-noise ratio were also used to evaluate the quality indicators. Note also that the compression time and image recovery time do not exceed 70 ms when using a 64-bit software implementation on a personal computer with an Intel(R) Core(TM) i5-8300H quad-core processor CPU@2.30 GHz and 8 GB of RAM running the Microsoft Windows 10.

Figure 6 shows the reconstructed images obtained from the aeroimages 2.2.01, 2.2.02, 2.2.03, and 2.2.04 compressed using the proposed approach in grayscale form



**Fig. 6** Examples of recovered images

with dimensions of  $1024 \times 1024$  pixels taken from the UCS-SIPI collection. The median-based method was used to obtain the threshold values.

In general, the restored images contain basic information but have a more smoothed appearance compared to the original representation. However, the weight model can be used to implement procedures for the formation of contour and texture features for the detection and recognition of objects in images in computer vision systems. Therefore, the use of it also for image compression allows you to reduce the time for the overall solution of functional problems. In addition, this approach can be used for image generation and transmission under interference conditions.

It should also be noted that this approach allows you to provide a compression quality that is comparable and in most cases exceeds the compression quality obtained when using methods for representing images in the popular image formats JPEG, TIFF, and PNG.

## 5 Conclusion

The approach to image compression described in this chapter allows for high processing speed and high-quality indicators of compression and recovery. It has the ability to flexibly configure parameters for use in various operating conditions of cyber-physical systems. At the same time, this approach allows us to obtain indicators that are comparable in quality to the corresponding indicators of the methods used to represent images in popular formats, and often exceed them.

Thus, the proposed approach can be effectively applied to image compression during their transmission and storage in cyber-physical systems based on computer vision technologies.

## References

1. Pavlov, B.P., Garifullin, R.F., Babushkin, V.M., Mingaleev, G.F.: Digital transformation of the economy. In: Proceedings of the 33rd International Business Information Management Association Conference, IBIMA 2019: Education Excellence and Innovation Management through Vision 2020, pp. 3359–3364 (2019)
2. Fedorova, A.A., Beliautsou, V.A., Anikin, I.V.: Prediction vehicle's speed with using artificial neural networks. In: Proceedings of the 2020 International Russian Automation Conference (RusAutoCon 2020), pp. 11–15 (2020)
3. Ismagilov, I.I., Murtazin, A.A., Kataseva, D.V., Katasev, A.S., Barinova, A.O.: Formation of a knowledge base to analyze the issue of transport and the environment. *Casp. J. Environ. Sci.* **18**(5), 615–621 (2020)
4. Novikova, S.V., Tutubalin, P.I., Snegurenko, A.P., Yakhina, R.R.: The optimal aircraft gas turbine engine control in low gas mode in the conditions of external additive noise. *J. Phys. Conf. Seri.* **1745**(1), 012039 (2021)
5. Shkinderov, M., Gizatullin, Z.: Technique for noise immunity analysis of access control systems using electromagnetic topology method. In: Proceedings of the 2020 International Russian Automation Conference (RusAutoCon 2020), pp. 144–148 (2020)

6. Gizatullin, Z.M., Shkinderov, M.S., Arkhipov, A.O.: Research of resonant effects in interconnects of multilayer PCB of computing equipment. In: Proceedings of the 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, pp. 116–119 (2020)
7. Li, J., Makarychev, M., Popov, A.: Alternative approach to solving computer vision tasks using graph structures. In: Kravets, A., Bolshakov, A., Shcherbakov, M. (eds.) Cyber-Physical Systems: Industry 4.0 Challenges. Studies in Systems, Decision and Control, vol. 260. Springer, Cham (2020)
8. Alekseev, A.V., Orlova, Y.A., Rozaliev, V.L., Zaboloeva-Zotova, A.V.: Two-stage segmentation method for context-sensitive image analysis. In: Kravets, A., Shcherbakov, M., Kultsova, M., Iijima, T. (eds.) Knowledge-Based Software Engineering. JCKBSE 2014. Communications in Computer and Information Science, vol. 466. Springer, Cham (2014)
9. Eickeler, S., Valdenegro, M., Werner, T., Kieninger, M.: Future computer vision algorithms for traffic sign recognition systems. In: Schulze, T., Müller, B., Meyer, G. (eds.) Advanced Microsystems for Automotive Applications 2015. Lecture Notes in Mobility. Springer, Cham (2016)
10. Xiong, N.N., Shen, Y., Yang, K., Lee, C., Wu, C.: Color sensors and their applications based on real-time color image segmentation for cyber physical systems. *J Image Video Proc.* **2018**, 23 (2018)
11. Salomon, D., Motta, G.: Handbook of Data Compression, 5th edn. Springer, London (2010)
12. Sayood, K.: Introduction to Data Compression, 5th edn. Morgan Kaufmann Publisher, Cambridge (2018)
13. Raid, A.M., Khedr, W.M., El-dosuky, M.A., Ahmed, W.: JPEG image compression using discrete cosine transform—a survey. *Int. J. Comput. Sci. Eng. Surv. (IJCSSES)* **5**(2) (2014)
14. Pearlman, W.A., Said, A.: Set partition coding: part i of set partition coding and image wavelet coding systems. *Found. Trends Signal Proc.* **2**(2), 95–180 (2008)
15. Pearlman, W.A., Said, A.: Image wavelet coding systems: part II of set partition coding and image wavelet coding systems. *Found. Trends Signal Proc.* **2**(3), 181–246 (2008)
16. Marcellin, M.W., Gormish, M.J., Bilgin, A., Boliek, M.P.: An overview of JPEG-2000. In: Proceedings of the IEEE Data Compression Conference, pp. 523–541 (2000)
17. Tang, Y.Y.: Wavelet Theory and Its Application to Pattern Recognition. World Scientific Publishing Company, London (2009)
18. Mallat, S.: A Wavelet Tour of Signal Processing, 3rd edn. Academic Press, New York (2009)
19. Ma, J., Plonka, G.: The curvelet transform. *IEEE Signal Process. Mag.* **27**(2), 118–133 (2010)
20. The USC-SIPI Image Database. <http://sipi.usc.edu/database/database.php>. Accessed 27 Mar., 2021
21. Lyasheva, S.A., Shleymovich, M.P.: Formation of energy features of the image based on wavelet transform. *J. Phys. Conf. Ser.* **1202**, 012006 (2019)
22. Shleymovich, M.P., Medvedev, M.V., Lyasheva, S.A.: Image analysis in unmanned aerial vehicle on-board system for objects detection and recognition with the help of energy characteristics based on wavelet transform. In: Proceedings of the SPIE, vol. 10342, p. 1034210 (2017)
23. Uthayakumar, J., Vengattaraman, T., Dhavachelvan, P.: A survey on data compression techniques: from the perspective of data quality, coding schemes, data type and applications. *J. King Saud Univ.–Comput. Inf. Sci.* (2018). In press. <https://www.sciencedirect.com/science/article/pii/S1319157818301101>. Accessed 27 Mar., 2021

# Construction of a Fuzzy Model for Contour Selection



L. Y. Emaletdinova and M. A. Nazarov

**Abstract** The chapter discusses an algorithm for constructing a model based on Tsukamoto's fuzzy inference, which is used to highlight contour points in a grayscale image. The input data of the fuzzy inference model is a set of vectors generated for each pixel of the image, excluding those that are on the border of the image. Each vector has eight components, corresponding to the eight surrounding pixels. Each component contains the difference in brightness between the considered and the surrounding pixel. Pixel brightness differences are described by three fuzzy sets with trapezoidal membership functions. The output of the model is a sign of belonging of the considered pixel to the contour. The attribute of belonging is described by a fuzzy set. The initial knowledge base of the model consists of  $3^8 \times 2$  production rules. The determination of the parameters of the membership functions is carried out using a genetic algorithm. A reduction of the knowledge base of the fuzzy model is proposed, which made it possible to reduce the number of rules used, which allows one to reduce the cost of computing resources.

**Keywords** Fuzzy logic · Fuzzy model · Fuzzy inference model · Tsukamoto algorithm

## 1 Introduction

Currently, research related to the creation of methods and technical systems for pattern recognition is one of the directions for the creation of artificial intelligence. To select and identify objects in the image, it is necessary to carry out preliminary processing of the original image, the stages, and methods of which depend on the goals and objectives set for the researcher.

One of the most commonly used types of image preprocessing in various tasks is edge extraction. A sufficient number of different algorithms have been developed in this area: the Pruitt operator, the Roberts cross operator, the Sobel operator, the Canny

---

L. Y. Emaletdinova (✉) · M. A. Nazarov  
Kazan National Research Technical University named after A.N. Tupolev–KAI, Kazan, Russia  
e-mail: [lilia@stcline.ru](mailto:lilia@stcline.ru)

operator, etc. [1, 2, 6–13, 18–20]. Described in the literature methods for highlighting contours differ in the used anti-aliasing filters and methods of calculating the edge strength. Many of the edge selection methods are based on the computation of the image gradient, however, they differ in the types of filters used to compute the gradients in the horizontal and vertical directions [16, 17].

Let us consider the formal setting of the problem of extracting contours in the original image [4].

Let be:

A full-colour image  $I$  is given, described by the display:

$$\beta : X \times Y \rightarrow R \times G \times B$$

where  $X \times Y$ —a set of coordinates of pixels of the image on the plane;

$R \times G \times B$ —many colour components of pixels of a full-colour image.

It is required, using a colour image, to highlight the contours—the points of the digital image in which the brightness changes sharply or there are other types of inhomogeneities.

To solve this problem, they traditionally switch from a full-colour image to a grayscale image using the transformation:

$$\varphi : R \times G \times B \rightarrow Z$$

Here  $Z = \{L_1, L_2, \dots\}$ —is the set of brightness values calculated based on the full-colour image by the equation:  $L_i = 0.2989 * r_i + 0.5870 * g_i + 0.1140 * b_i$ , where  $(r_i, g_i, b_i)$ —are the red, green, blue components of the  $i$ -th pixel of the full-colour image with coordinates  $(x_i, y_i)$ ,  $i = \overline{1, K}$ . Thus  $L_i = f(x_i, y_i) = \varphi(\beta(x_i, y_i))$ —is the brightness value of the  $i$ -th pixel of the halftone image. In what follows, we will consider the normalized brightness values in the interval  $[0, 1]$ , for which we divide the brightness  $L_i$  by 256.

You can use one of the above methods as a method for selecting contours, however, after using them, contours with numerous breaks can be obtained, which can lead to difficulties in selecting and identifying an object. To solve this problem, it is necessary to select the filtering threshold, which is quite difficult in the case of automation of this process. In this regard, it is proposed to use an approach in the selection of boundaries based on the fuzzy logical inference of Tsukamoto [3, 14], and to build a fuzzy model for deciding the belonging of the considered pixel to the contour based on the change in brightness in the considered pixel and its environment.

## 2 Building a Fuzzy Inference Model

Consider a certain pixel of the image, its brightness  $L$ , and the values of the brightness  $L_i, i = \overline{1, 8}$  neighbouring pixels (Fig. 1). Since the value of the brightness of the pixel lies in the range  $[0, 1]$ , then the values of the difference in brightness  $d_i = L - L_i$  will be in the range  $[-1, 1]$ .

We introduce four fuzzy sets in the interval  $[-1, 1]$ :

- A—«The brightness of the pixel from the neighbourhood is much less than the brightness  $X$  of the central pixel»:  $A = \{(d, \mu_A(d)); d \in [-1, 1]\}$ ;
- B—«The brightness of the pixel from the neighbourhood is close to the brightness  $X$  of the central pixel»:  $B = \{(d, \mu_B(d)); d \in [-1, 1]\}$ ;
- C—«The brightness of a pixel from the neighbourhood is much greater than the brightness  $X$  of the central pixel»:  $C = \{(d, \mu_C(d)); d \in [-1, 1]\}$ ;
- E—«The considered center pixel belongs to the contour»:  $E = \{(z_1, \mu_E(z)); z_1 \in [0, 1]\}$ ,
- F—«The considered center pixel does not belong to the contour»:  $F = \{(z_2, \mu_F(z)); z_2 \in [0, 1]\}$ ,

where  $\mu_A, \mu_B, \mu_C : [-1, 1] \rightarrow [0, 1]$ —are the membership functions of fuzzy sets, respectively  $A, B, C$ ;  $\mu_E, \mu_F : [0, 1] \rightarrow [0, 1]$  membership functions of fuzzy sets, respectively  $E, F$  (Fig. 2).

Then the knowledge base including of predicate rules and having 8 inputs ( $d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8$ ) and one output  $z$  will consist of the following rules:

$$\begin{aligned} \text{If } d_1 = G_1 \ \& \ d_2 = G_2 \ \& \ d_3 = G_3 \ \& \ d_4 = G_4 \ \& \ d_5 = G_5 \ \& \ d_6 \\ & = G_6 \ \& \ d_7 = G_7 \ \& \ d_8 = G_8, \text{ then } z_1^k = E, \end{aligned} \tag{1}$$

$$\begin{aligned} \text{If } d_1 = G_1 \ \& \ d_2 = G_2 \ \& \ d_3 = G_3 \ \& \ d_4 = G_4 \ \& \ d_5 = G_5 \ \& \ d_6 \\ & = G_6 \ \& \ d_7 = G_7 \ \& \ d_8 = G_8, \text{ then } z_2^k = F, \end{aligned} \tag{2}$$

where  $G_i \in \{A, B, C\}, i = \overline{1, 8}, k = \overline{1, 3}$ .

Thus, the knowledge base will consist of  $3^8 \times 2 = 13122$  rules, since the enumeration of possible combinations of values  $G_i, i = \overline{1, 8}$  is carried out.

Tsukamoto’s inference algorithm, adapted to the solution of the problem under consideration, consists of the following steps:

**Fig. 1.** The frame of the analyzed pixels of the image

$L_1$	$L_2$	$L_3$
$L_8$	$L$	$L_4$
$L_7$	$L_6$	$L_5$

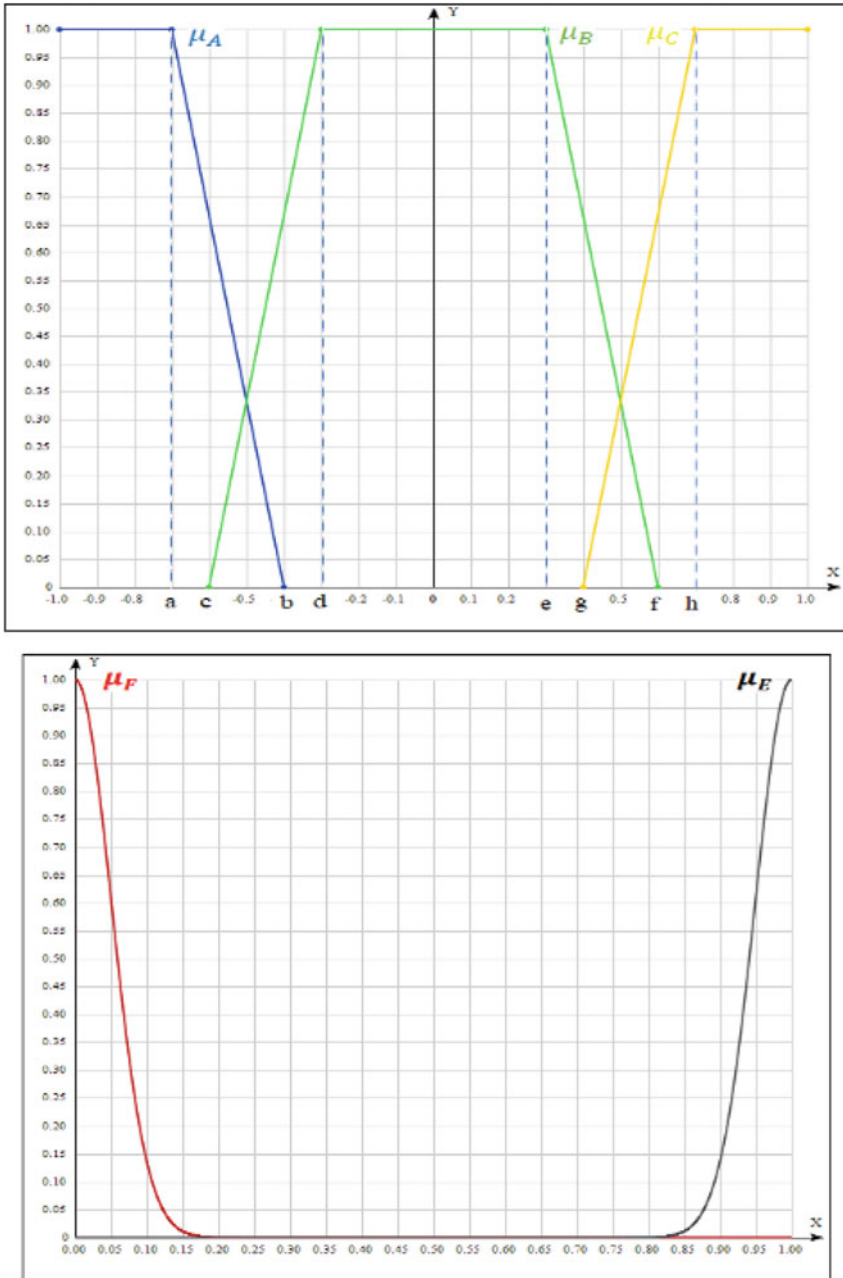


Fig. 2 Membership functions of fuzzy sets A, B, C, E, F

1. The degrees of truth  $\mu_A(d_0)$ ,  $\mu_B(d_0)$ ,  $\mu_C(d_0)$  are determined for the premises of each rule (1), (2).
2. The “cut-off” levels  $\alpha_i$  for the premises of each of the rules are found,  $i = \overline{1, 13122}$  using the “minimum” operation.
3. For each k-th rule ( $k = \overline{1, 3^8}$ ) defines a clear value  $z_1$  as a solution to the equations  $\alpha_k = \mu_E(z_1)$ . The resulting value  $z_1$  determines the value of the output k-th rule  $z_1^k = z_1$ .
4. For each k-th rule ( $k = \overline{1, 3^8}$ ) defines a clear value  $z_2$  as a solution to the equations  $\alpha_k = \mu_F(z_2)$ ,  $k = (3^8 + 1), 2 \times 3^8$ . The resulting value  $z_2$  determines the value of the output k-th rule  $z_2^k = z_2$ .

Determine the clear value of the output variable  $z_0$  as a weighted average:

$$z_0 = \frac{\sum_{k=1}^{3^8} \alpha_k z_1^k + \sum_{k=1}^{3^8} \alpha_{k+3^8} z_2^k}{\sum_{k=1}^{3^8 \times 2} \alpha_k}$$

Having set the boundary value  $\varepsilon$ , the condition  $z_0 \geq \varepsilon$  is checked and a decision is made whether the pixel belongs to ( $z_0 = 1$ ) or not ( $z_0 = 0$ ) to the contour:

$$\begin{cases} \text{Если } z_0 \geq \varepsilon, & \text{то } z_0 = 1, \\ \text{Если } z_0 < \varepsilon, & \text{то } z_0 = 0. \end{cases}$$

### 3 Reduction of the Knowledge Base of the Fuzzy Model of Contour Selection

The resulting knowledge base is redundant and contains production rules that may never be used when making decisions. These rules can be excluded by an expert. However, this approach is laborious and a large number of rules can lead to expert errors. The following approach is proposed to automate the reduction process:

Generate a binary image from a full-colour image using traditional methods.

Moving horizontally from left to right, consider sequentially  $3 \times 3$  pixel frames and form a set  $\{(d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, z_0)\}$ , where  $d_i$ ,  $i = \overline{1, 8}$ —the difference in brightness between pixels surrounding the center pixel of the frame and the center-most frame;  $z_0 = 1$ , if the pixel is contour;  $z_0 = 0$ , otherwise.

For each vector of values  $(d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, z_0)$  from the generated set, search for a suitable rule and mark it as used.

Rules that were not used when processing vectors  $(d_1, d_2, d_3, d_4, d_5, d_6, d_7, d_8, z_0)$  of the generated set are excluded from the model.



## 4 Determination of the Parameters of the Membership Functions of a Fuzzy Model

In the obtained fuzzy model, the parameters of the membership functions  $\mu_A, \mu_B, \mu_C$  are unknown:

$$\mu_A(x; a, b) = \begin{cases} 1, & \text{для } x \leq a \\ \frac{b-x}{b-a}, & \text{для } a < x \leq b \\ 0, & \text{для } x > b \end{cases}$$

$$\mu_B(x; c, d, e, f) = \begin{cases} 0, & \text{для } x \leq c \\ \frac{x-c}{d-c}, & \text{для } c < x \leq d \\ 1, & \text{для } d < x < e \\ \frac{f-x}{e-f}, & \text{для } e \leq x \leq f \\ 0, & \text{для } x > f \end{cases}$$

$$\mu_C(x; g, h) = \begin{cases} 0, & \text{для } x \leq g \\ \frac{x-g}{h-g}, & \text{для } g < x \leq h \\ 1, & \text{для } x > h \end{cases}$$

Membership functions  $\mu_E, \mu_F$  have the following form:

$$\mu_E(x, \mu_1, \sigma_1) = \exp\left(-\frac{(x)^2}{2*(0.5)^2}\right) \text{ для } \forall x \geq 0$$

$$\mu_F(x, \mu_2, \sigma_2) = \exp\left(-\frac{(x-1)^2}{2*(0.5)^2}\right) \text{ для } \forall x \leq 1$$

To find the parameters of membership functions, it is proposed to use a genetic algorithm [5, 15]:

1. Prepare training set.

- Generate a grayscale image from a full-colour image.
- Draw visually defined contours on a halftone image.
- Moving horizontally from left to right, consider sequentially  $3 \times 3$  pixel frames and form a set  $\left\{ \left( d_1^j, d_2^j, d_3^j, d_4^j, d_5^j, d_6^j, d_7^j, d_8^j, z_0^j \right), j = \overline{1, N} \right\}$ , where  $d_i^j, i = \overline{1, 8}$ —the difference in brightness between pixels surrounding the center pixel of the frame and the center-most frame;  $z_0^j = 1$ , if the pixel is contour;  $z_0^j = 0$ , otherwise;  $N$ —number of frames.

2. Stage of coding.

Decimal values of the optimized parameters  $a, b, c, d, e, f, g, h$ , are encoded by a binary string (chromosome) of length  $\Omega$ .

3. Formation of the initial population of 30 chromosomes  $\vartheta_i, i = \overline{1, 30}$ :
  - 3.1 Chromosome creation based on random number generation.
  - 3.2 Obtaining decimal values of the required parameters recorded in each chromosome by converting parts of chromosomes from binary notation to decimal.
4. Condition check: If  $(a < b)$  and  $(b > c)$  and  $(c < d)$  and  $(d < e)$  and  $(e < f)$  and  $(f > g)$  and  $(g < h)$  and  $(a < d)$  and  $(e < h)$  and  $(a < 0)$  and  $(b < 0)$  and  $(c < 0)$  and  $(d < 0)$  and  $(e > 0)$  and  $(f > 0)$  and  $(g > 0)$  and  $(h > 0)$ , then the corresponding chromosome is used in the population. Otherwise, clause 3.1 is performed.
5. Calculation of the value of the fitness function of the chromosomes of the population.
  - (a) Conversion of the  $\vartheta_i$  chromosome genotype to the phenotype.
  - (b) For each sample, using a fuzzy inference model, determine the value of the output variable.
  - (c) Calculation of the fitness function:  $P(\vartheta_i) = - \sum_{j=1}^N (z_{i0}^j - \tilde{z}_{i0}^j)^2$ ,  
 where  $z_{i0}^j, \tilde{z}_{i0}^j$ —are the values of the output variable, respectively, given in the example and calculated by the model.
6. Formation of a new population based on selection, crossing, and mutation of chromosomes.
7. Iterative execution of points 4 and 5 until the formed population of a given quality is reached.

A fragment of the original image and the result of the fuzzy model are shown in Fig. 3.

This approach made it possible to highlight the contours in the image and use them for further recognition using the neural network approach described in [4].

## 5 Conclusion

Using the Tsukamoto fuzzy inference algorithm to isolate the contours of objects in an image is an alternative to common gradient methods that allows you to select contours with fewer missing pixels. To build a fuzzy model for extracting contours in an image, it is necessary to form training samples that allow you to build and reduce the knowledge base, as well as use a genetic algorithm to refine the parameters of the membership functions of fuzzy sets used in the model. Reduction of the knowledge base of a fuzzy model makes it possible to reduce the number of rules used, which makes it possible to reduce the cost of computational resources. When defining the

**Fig. 3** Fragment of the original image and the result of the fuzzy model



parameters of the membership functions, it is advisable to take into account the ratios of their values relative to each other to exclude chromosomes from consideration, which have lower values of the fitness function.

## References

1. Gonzalez, R., Woods, R.: *Digital image processing*, 1072 p. Prentice-Hall, Inc., Upper Saddle River, NJ (2005)
2. Katasyov, A.S., Katasyova, D.V., Kirpichnikov, A.P.: Neuronet biometric imaging system of human face recognition. *Vestnik of Kazan Technological University*. #18 (2016)
3. Ershov, M.D., Georgieva, S.S.: Research on approaches to object contour detection based on preliminary filtration and fuzzy logic. "Digital Signal Processing"-M.: "Russian Scientific and Technical Society of Radio Engineering, Electronics and Communications named after A.S. Popov" (2019)
4. Emaletdinov, L.Y., Nazarov, M.A.: Neuronetic algorithm of object recognition on the image based on the reference contour. In: Bolshakov, A. (ed.) *Mathematical methods in technology and technology: Sat International Scientific Conferences in 12 t. T. 4/under the general*. St. Petersburg: Ed-vo Polytechnic. Un-ta, 137–141 s. (2020)
5. Goldberg, D.E.: *Genetic Algorithms in Search, Optimization and Machine Learning*, 13th ed., 432 p. Addison-Wesley Longman Publishing Co., USA (1989)
6. Muthukrishnan, R.: Edge detection techniques for image segmentation. *Int. J. Comput. Sci. Inf. Technol.*, 259–267 (2011)
7. Canny, J.F.: A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 679–698 (1986)

8. Duda, R., Hart, P.: *Pattern Classification and Scene Analysis*, 512 p. "Wiley" Pub., New York (1973)
9. Mutneja, V.: *Methods of image edge detection: a review*. *J. Elect. Electron. Syst.* (2015)
10. Savant, S.: *A review on edge detection techniques for image segmentation*. (*IJCSIT*) *Int. J. Comput. Sci. Inf. Technol.* **5**(4) (2014)
11. Ghosh, C., Majumder, S., Ray, S., Datta, S., Mandal, S.N.: *Different EDGE detection techniques: a review*. In: *Electronic Systems and Intelligent Computing. Lecture Notes in Electrical Engineering*, vol. 686. Springer, Singapore (2020)
12. Magnier, B., Abdulrahman, H., Montesinos, P.: *A Review of supervised edge detection evaluation methods and an objective comparison of filtering gradient computations using hysteresis thresholds*. *J. Imaging* **4**, 74 (2018)
13. Ansari, M., Kurchaniya, D., Dixit, M.: *A comprehensive analysis of image edge detection techniques*. *Int. J. Multimed. Ubiquitous Eng.*, 12 (2017)
14. Napitupulu, S., Nababan, E.B., Sihombing, P.: *Comparative analysis of fuzzy inference tsukamoto mamdani and sugeno in the horticulture export selling price*. In: *2020 3rd International Conference on Mechanical, Electronics, Computer, and Industrial Technology (MECnIT)*, Medan, Indonesia, pp. 183–187 (2020)
15. Sivanandam, S.N., Deepa, S. N.: *Introduction to Genetic Algorithms*, 442 p. Springer, Berlin Heidelberg (2008)
16. Jason, D.B.: *Chapter 4-Memory optimization and video processing*. *Embedded Systems*, pp. 147–185. Morgan Kaufmann (2016)
17. McReynolds, T., Blythe, D.: *Advanced Graphics Programming Using OpenGL*, 644 p. Morgan Kaufmann (2005)
18. Koschan, A., Abidi, M.: *Digital Color Image Processing*, 394 p. Wiley (2019)
19. Dawson-Howe, K., Fitzgibbon, A., Williams, C., Robertson, C., Trucco, E., Fisher, R., Breckon, T.: *Dictionary of Computer Vision and Image Processing*, 388 p. Wiley (2018)
20. Parker, R.J.: *Algorithms for image processing and computer vision*, 506 p. Wiley (2018)

# Identification of Vehicle Trajectory Anomalies on Streaming Video



Igor Anikin  and Aigul Mardanova 

**Abstract** We suggested the approach for identifying a vehicle's trajectory anomalies on streaming video. Video surveillance cameras installed on street intersections in Kazan city were used as the source data. We applied YOLO v3 artificial neural network and KCF object tracking algorithm to get an initial vehicle's trajectories from raw video streams. Then we used polynomial approximation, thinning, and clustering for training our model and classifying all trajectories into normal or anomaly. The Ramer-Douglas-Peucker algorithm was used for trajectory's thinning. We suggested the modified Longest Common SubSequence (LCSS) metric for the comparison of trajectories under spatial perspective. In the training stage, we used the agglomerative hierarchical clustering for the creation of two types of clusters – with normal and anomaly trajectories. In the classification stage, we used cluster models and the modified LCSS metric to classify certain vehicle trajectories as normal or abnormal. The framework has been developed for the identification of vehicle trajectory anomalies on video streams. This framework could be used as the key element of intelligent transport systems.

**Keywords** Cyber-physical systems · Intelligent transport systems · Machine learning · Video processing · Clustering · Anomaly detection

## 1 Introduction

Nowadays, cyber-physical systems (CPS) have become very perspective in different Society 5.0 applications [1, 2]. Intelligent transportation systems (ITS) are one of them [3]. ITS is an essential component of the modern smart city. One of their main goals is to ensure the efficiency and safety of transportation systems based

---

I. Anikin (✉)

Kazan National Research Technical University Named After A.N. Tupolev-KAI, 10 K.Marx str,  
Kazan 420111, Russia  
e-mail: [IVAnikin@kai.ru](mailto:IVAnikin@kai.ru)

A. Mardanova  
Zalando SE, Erfurt, Germany

on automatic monitoring and management of urban traffic [4, 5]. The widespread introduction of such systems into everyday life is intended to improve the road safety and capacity of road networks, quality of traffic, to decrease the negative impacts of transportation systems on the environment.

Video surveillance systems are one of the key elements of modern ITS [6]. They provide streaming video from the roads which could be used to solve different tasks related to traffic monitoring: detection of the over speed and traffic congestions, violation of traffic rules, etc. Detection of vehicle trajectory anomalies is one of the most important tasks for video surveillance systems in ITS [7]. Such trajectories may indicate the presence of different adverse events in the road traffic, such as violation of traffic rules, car incidents, lane congestions, inadequate drivers, etc. The information about such events should be used further at the high level of ITS to improve road safety and capacity.

In this chapter, we suggested a new approach for the identification of vehicle trajectory anomalies on the video streams from street intersection road cameras in Kazan city. We also developed the framework for identification such anomalies. This framework could be used as the key element of intelligent transport systems.

## 2 Background

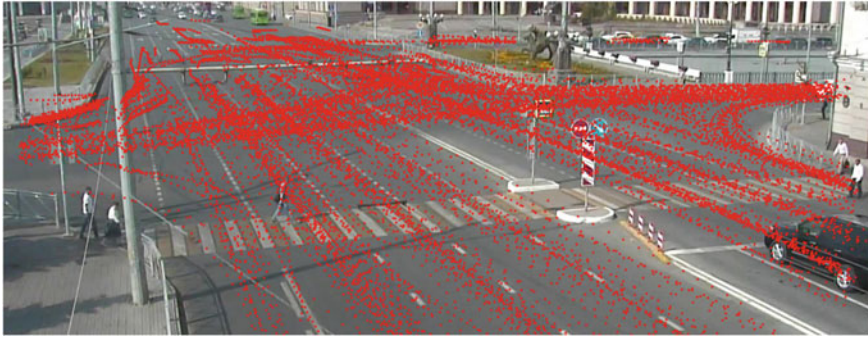
Generally, trajectories can be described as multi-dimensional sequences containing a temporally ordered list of locations along with any additional information [8]. So, since a trajectory, denoted as  $\tau$ , represents consecutive positions of a moving target object in the temporal domain, in a case of a single-camera surveillance data it can be defined as a sequence of 3D points (1):

$$\tau = (x_1, y_1, t_1), (x_2, y_2, t_2), \dots, (x_n, y_n, t_n) \quad (1)$$

Here  $(x_i, y_i)$  defines the position of the target object on the image at time  $t_i$ . We can see in Fig. 1 the example of vehicle trajectories obtained from the street intersection video surveillance camera.

Trajectory anomalies can be described as traffic flow patterns, which significantly deviate from some normal behavior pattern or are inconsistent with the rest of traffic behavioral patterns. They are supposed to have great local or global differences with the majority of trajectories in terms of a chosen similarity metric [9]. Our work is focused on determining spatial trajectory anomalies where the classification process takes into consideration only spatial information of moving object trajectories ( $(x_i, y_i)$  coordinates). Such anomalies may indicate the presence of different adverse events in the road traffic: car incidents, violation of traffic rules (U-turns, double line crossing or moving in an opposite direction), inadequate drivers, etc.

Detection of trajectory anomalies on source video streams (Fig. 1) is complicated by the existence of spatial perspective and the relative size of the objects (vehicles). If



**Fig. 1** Examples of vehicle trajectories

the vehicle is more distant, it becomes smaller. In this case, more distant trajectories become closer to each other than the trajectories which are closer to the video camera. This fact has to be taken into account in the distance metric which will be used for trajectories comparison.

Another fact which complicates the process of trajectory anomalies detection is a big number  $n$  of points in the sequence (1). It affects the performance of anomaly detection algorithms. The distribution of these points on the image (Fig. 1) is not uniform due to the spatial perspective and different vehicle speeds in different locations. Taking into account these facts, we need to make a thinning of sequences (1) to reduce the number of points and leave only key ones.

We suggested an approach for the identification of vehicle trajectory anomalies on the video stream from street intersection road cameras in Kazan city. This approach takes into account above mentioned challenges.

### 3 Suggested Approach

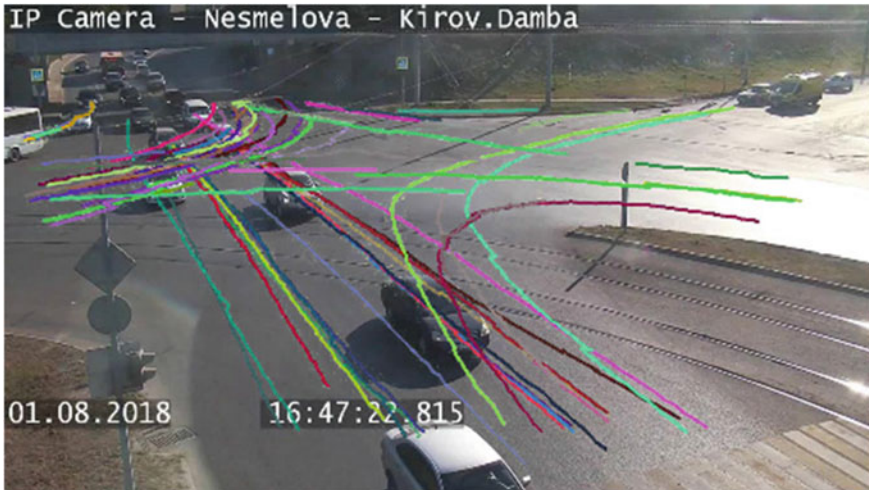
The suggested approach consists of two basic stages: training and classification.

On the training stage we implement the following basic steps:

- Step 1.1. Extraction of the number of sequences (1) from the source video stream and their filtration.
- Step 1.2. Thinning of the sequences (1). Reduction of the number of points in the sequences (1) and detection of the key points for further analysis.
- Step 1.3. Comparison of thinning sequences and their clustering. Creation of two types of clusters—with normal and anomaly trajectories.

On the classification stage we implement the following basic steps:

- Step 2.1. Extraction of the vehicle’s trajectory (1) from the source video stream and its thinning.



**Fig. 2** Vehicle trajectories after YOLO v3 and KCF

Step 2.2. Classification of the trajectory as normal or abnormal based on cluster models created during the training stage.

Let's consider these steps more deeply.

### ***3.1 Trajectories Extraction***

To extract the sequences (1) from the source video stream we apply the approach suggested by authors in [10, 11]. We used source video stream framing, then apply YOLO v3 artificial neural network [12] to detect and recognize vehicles on every frame. Finally, we used Kernelized Correlation Filters (KCF) algorithm for tracking the recognized vehicles [13]. The sequences (1) are created as the result of such tracking. Some of these trajectories are senseless and look deficient due to some faults of the KCF tracker on the raw video stream. To remove such trajectories, we applied filtering and ignore short trajectories with small covered distances. We can see in Fig. 2 the example of obtained vehicle trajectories with linear-spline interpolation.

### ***3.2 Trajectories Thinning***

Sequences (1) contain a lot of points and most of them are not important to detect the anomalies [14]. To delete unimportant points and increase the system performance we used the thinning of the sequences (1), reducing the number of points, and leaving



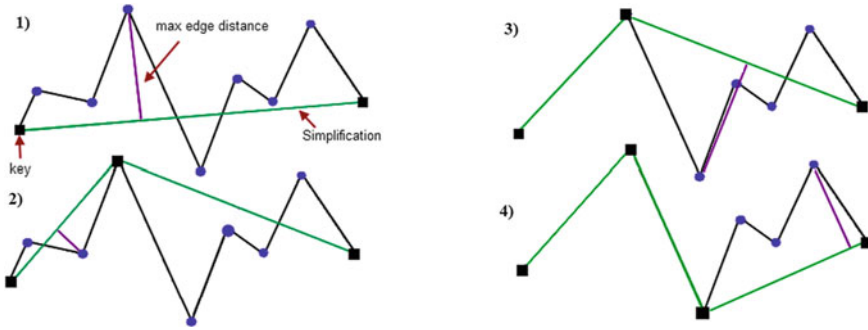


Fig. 3 Sequence (1) thinning with using RDP algorithm

only key points which are important for further analysis. We selected the key points by using the following methods.

- (1) Approximation of points in the sequence (1) by cubic or quartic polynomials  $f(\tau)$  and detection of the key points with  $f'(\tau) = 0$  or  $f''(\tau) = 0$ . Preliminary evaluation results are shown that these kinds of polynomials are most suitable for trajectories approximation. Linear and quadratic functions cannot fit the source data properly due to the complexity of trajectory forms.
- (2) Sequence (1) thinning using Ramer-Douglas-Peucker (RDP) N algorithm [15]. The main concept of the basic RDP algorithm is the following. We have a source trajectory curve represented by line segments (polyline). The RDP algorithm seeks to obtain a representative polyline with a smaller amount of points (Fig. 3). The resulting simplified trajectory consists of a subset of the original sequence (1), with indispensably kept first and last points. The algorithm uses the notion of ‘dissimilar’, which is calculated as the maximum distance between the original trajectory curve and its approximation. Unfortunately, the traditional RDP algorithm does not provide a guaranteed amount of resulting approximation points [16]. That is why we used its modification–RDP N algorithm for sequence (1) thinning. RDP N algorithm uses a point count tolerance instead of a point-to-edge distance tolerance as an objective for approximation.

### 3.3 Trajectories Comparison

We used Longest Common SubSequence (LCSS) metric [17–19] as a similarity measure between trajectories, defined by reduced sequences (1). LCSS tries to match two trajectory sequences based on the longest common sub-sequence between them. The task of finding the longest common sub-sequence is usually solved recursively. LCSS metric allows two trajectories to stretch, enables some elements to remain unmatched, and is quite robust to the presence of outliers. The LCSS distance is calculated according to the expressions (2, 3):

$$D_{LCSS}(\tau_1, \tau_2) = 1 - \frac{LCSS_{\delta, \varepsilon}(\tau_1, \tau_2)}{\min(m, n)} \quad (2)$$

$$\begin{aligned} LCSS_{\delta, \varepsilon}(\tau_1, \tau_2) &= \{0, \text{if } m = 0 \text{ or } n = 0\} + LCSS_{\delta, \varepsilon}(\text{Head}(\tau_1), \text{Head}(\tau_2)), \\ &\text{if } |t_{1x, m} - t_{2x, n}| < \varepsilon \text{ and } |t_{1y, m} - t_{2y, n}| < \varepsilon \text{ and } |m - n| \\ &< \delta \max\{LCSS_{\delta, \varepsilon}(\text{Head}(\tau_1), \tau_2) LCSS_{\delta, \varepsilon}(\tau_1, \text{Head}(\tau_2)) \\ &\text{otherwise} \end{aligned} \quad (3)$$

Here  $\tau_1$  and  $\tau_2$ —trajectories (sequences (1)),  $m$ ,  $n$ —trajectory lengths (the number of points in sequences (1)).

Parameter  $\delta$  defines the maximum distance in terms of time between two trajectory points in which we can look to match a given point from one trajectory to another. It could be considered as the maximum index difference between two input trajectories allowed in calculations.

Parameter  $\varepsilon$  defines the size of proximity to look for matches in sequences (1) in terms of space. It is a floating-point number that represents the maximum allowed distance between trajectory points in each dimension to consider them as equivalent.

The  $\text{Head}(\tau)$  function is defined to return the first  $M - 1$  points from the trajectory  $\tau$ , representing the trajectory with the last point removed.

Traditionally  $\delta$  and  $\varepsilon$  parameters are constant and defined in advance. However, in our case, we need to take into account the spatial perspective. These parameters are functionally dependent on the position of a moving vehicle on the image in respect to the video camera. Since the bottom part of the image represent the region located closer to the surveillance camera, moving objects on the upper part of the image are more distant from the camera and as a result are more densely located concerning the representation of each other on the image. As a result,  $\delta$  and  $\varepsilon$  parameters should be adaptive and dependent on the distance from the moving object to the video camera. These parameters should be decreased when the trajectory point gets farther from the camera and increases the trajectory point gets closer.

We suggested the modification of the LCSS metric to take into account the distance from the vehicle to the video camera. This modification used adaptive  $\delta$  and  $\varepsilon$  values, defined by expressions (4 and 5).

$$\delta = \text{coeff}_{\delta} * \min(\tau_1.\text{length}, \tau_2.\text{length}) \quad (4)$$

$$\begin{aligned} \varepsilon_x &= \text{coeff}_{\varepsilon} * \frac{(\max X - \min X)}{\text{distToCP}} \\ \varepsilon_y &= \text{coeff}_{\varepsilon} * \frac{(\max Y - \min Y)}{\text{distToCP}} \end{aligned} \quad (5)$$

Here  $\text{distToCP}$  value is calculated as Euclidean distance between the trajectory point on the image and video camera position. The values  $\text{coeff}_{\delta}$ ,  $\text{coeff}_{\varepsilon}$  are selected experimentally.

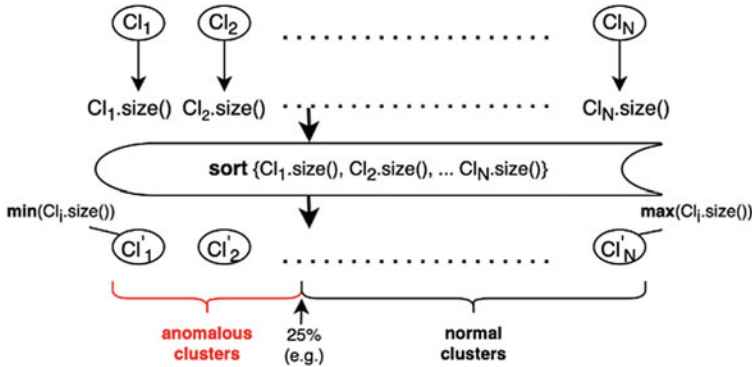


Fig. 4 Normal and anomalous clusters

### 3.4 Trajectories Clustering

Defining the LCSS metric for comparison of trajectories make it possible to compare their similarity and make clustering. We used the agglomerative hierarchical clustering approach [20] to create two types of clusters—with normal and anomaly trajectories. The general scheme of defining the normal and anomalous clusters is presented in Fig. 4. We used the Dunn’s Validity Index (DI) [21, 22] to evaluate the quality of clustering. It is an internal evaluation metric that is intended to identify compact clusters with a small variance between cluster members which are well-separated between each other.

We used the following concept in the clustering methodology. A normal cluster is a cluster containing a relatively large amount of trajectories in comparison with others. Trajectories in such clusters define a frequent vehicles’ behavior. Similarly, the anomalous cluster contains a relatively small amount of trajectories. In this work, we used 0, 25 quantile to separate normal and anomalous clusters.

We also define the cluster models (CM) for normal and abnormal clusters which are a compact representation of them. We used the approach suggested in [23] and take a CM as the trajectory with a minimum average LCSS distance to other trajectories in this cluster (Fig. 5).

### 3.5 Trajectory Classification

In the classification stage, we used the modified LCSS metric and previously created cluster models to classify certain vehicle trajectories as normal or abnormal. The anomalies can be detected as trajectories related to anomalous clusters or which are not associated with any of the known clusters.

Fig. 5 CM detection

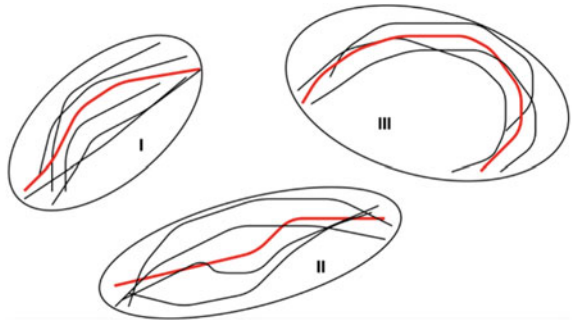
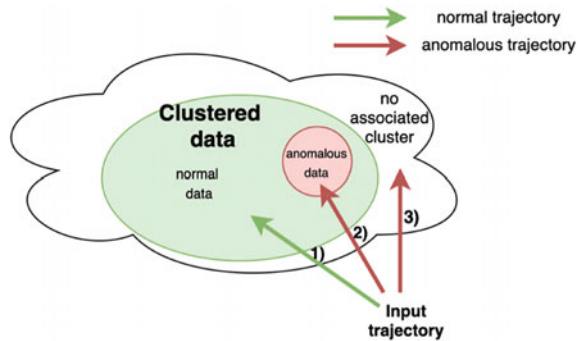


Fig. 6 Vehicle's trajectory classification

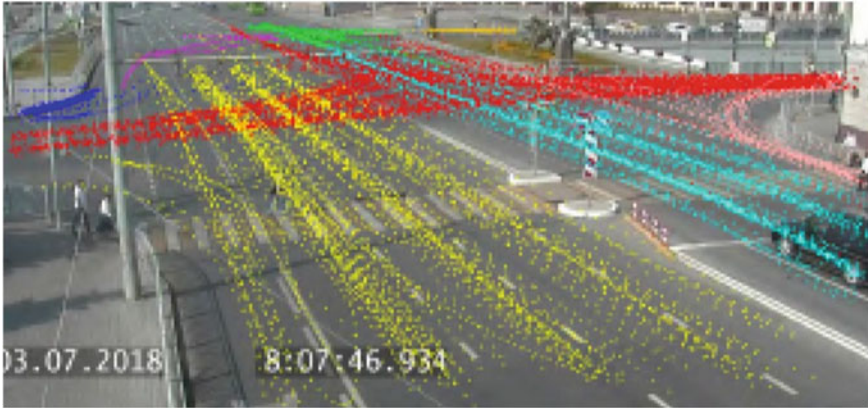


We used the following steps for real vehicle trajectory classification based on cluster models (Fig. 6).

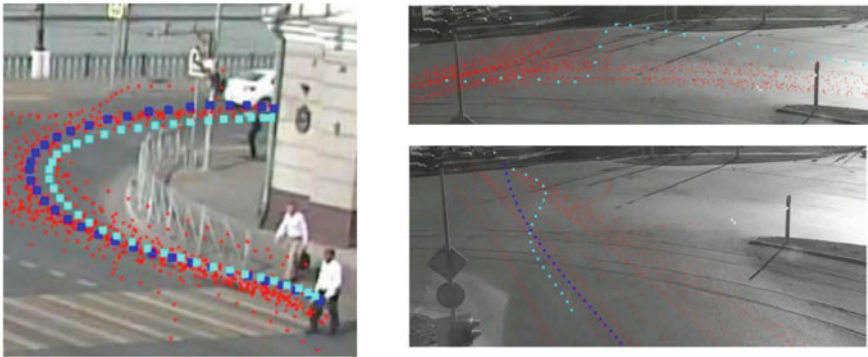
- Step 1. Extract the vehicle's trajectory  $\tau$  from the source video stream and its thinning according to the above-described techniques.
- Step 2. Calculate the LCSS distance from  $\tau$  to each CM.
- Step 3. Select the closest cluster (normal or anomalous). If the input trajectory does not comply with any of the known clusters, it is considered to be anomalous.

#### 4 The Framework for Identification of vehicle's Trajectory Anomalies on Streaming Video

The framework has been developed for the identification of vehicle trajectory anomalies. We used the following stack of technologies in this framework: Java-11OpenJDK, ApacheMaven-3.6.3, CommonsMath: The Apache Commons Mathematics Library-3.4.1, Java AWT, Javax Swing. We integrated it with another framework of vehicle detection and tracking on streaming video which was developed by authors in [7, 8].



**Fig. 7** The example of trajectories clustering



**Fig. 8** The examples of the normal trajectory (left side) and anomaly trajectories (right side)

In Fig. 7 we can see the example of trajectories clustering.

In Fig. 8 we can see the example of trajectory classification and detection of normal and abnormal trajectories.

The developed framework could be used as the key element of intelligent transport systems.

## 5 Evaluation

We evaluated the suggested approach on real video streams using the developed framework. As a result, we evaluated the quality of algorithms that we used, selected the rational values of their parameters, and did some suggestions and conclusions. The quality of the clustering process was assessed using Dunn's Validity Index.

To make a justified decision about the degrees of approximation polynomials we try to fit the trajectories using polynomials of 3rd, 4th, 5th degrees respectively. We used the R2 metric to assess the quality of such approximation and got the following conclusions from the experiments:

- Using linear and quadratic functions cannot fit the source data properly due to the complexity of trajectory forms.
- Using 5th-degree polynomials did not affect significantly the final results.
- Cubic and quartic polynomials work well even with slow-moving or inactive vehicle trajectories (including trajectories of vehicles waiting at the intersections). Such kinds of trajectories are common for the city traffic in rush hours. These polynomials work well on vehicle trajectories with a non-constant speed or an acceleration on some sections of a path. They show acceptable results for trajectories with complex shapes (sharp turns, Pascal snails).

We compared the results of cubic and quartic polynomial approximation with trajectory thinning by the RDP N algorithm. We can say that polynomial approximation works better if the points in trajectory (1) are greatly affected by the time component, for example for trajectories from traffic jams. RDP N algorithm works better if the points in trajectory (1) are not much affected by time.

We tested our framework with the implementation of trajectory filtering and without it. We can say that the quality of the clustering process (according to Dunn's Validity Index) is better when the filtering of source trajectories was used. It helps to delete short outlier trajectories that appear as a result of vehicle loss by the KCF tracking algorithm.

Using LCSS and modified LCSS metrics is very computationally expensive and significantly increases the time of model training [19]. Thinning the trajectories (1) using polynomial approximation or RDP N algorithm can significantly increase the performance at this stage, but the final number of points should not be high (not more than 10). For example, clustering of 438 trajectories (95,703 trajectory pairs), defined by a maximum of 8 points, took approximately 19 min on PC with Intel Core i5 which is borderline. The classification process does not demand so many LCSS comparisons and could be implemented in real-time.

We tried to test our framework by using static and adaptive  $\delta$  and  $\epsilon$  parameters in the LCSS metric. The quality of the clustering process (according to Dunn's Validity Index) was much better for the adaptive values.

## 6 Conclusion

We implemented the framework for the identification of vehicle trajectory anomalies on video streams. We also used it to detect the anomalies on the video from street intersection road cameras in Kazan city. We can say that the suggested approach is working well in this case.

We integrated the developed framework with another one (developed by authors) which separately solves the task of vehicle detection and tracking on the video stream. The framework applies trajectory filtering and automatically selects the way for trajectory thinning (cubic/quartic polynomials or RDP N) according to the R2 metric and specific trajectory data. Thinning of trajectories significantly increases the performance of the system on the training and classification stages. In the training stage, we used agglomerative hierarchical clustering to create clusters with normal and anomaly trajectories. In the classification stage, we used the modified LCSS metric and previously created cluster models to classify certain trajectories as normal or abnormal.

Finally, we can conclude that the suggested approach and the framework can be effectively used as the key element of modern intelligent transport systems. The information about detected anomalies could be used further at the high level of ITS to improve road safety and capacity. Further research will be devoted to increasing the quality of vehicle detection and tracking on video streams, the continuous adaptation of cluster models to current conditions, and increasing the system performance.

## References

1. Kravets, A.G., Bolshakov, A.A., Schervakov, M.V. (eds.): *Cyber-Physical Systems: Advances in Design & Modelling*, p. 340. Springer Nature, Switzerland (2020)
2. Kravets, A.G., Bolshakov, A.A., Schervakov, M.V. (ed.): *Society 5.0: Cyberspace for Advanced Human-Centered Society*, p.280. Springer International Publishing (2021)
3. Popov, S., Kurochkin, M., Kurochkin, L.M., Glazunov, V.: Hardware and software equipment for modeling of telematics components in intelligent transportation systems. In: *Lecture Notes in Computer Science*, 2014, vol. 8638, pp. 598–608. (2014)
4. Krushel, E.A., Stepanchenko, I.V., Panfilov, A.E., Berisheva, E.D.: An experience of optimization approach application to improve the urban passenger transport structure. In: Kravets, A., Shcherbakov, M., Kultsova, M., Lijima, T. (eds.) *Knowledge-Based Software Engineering, 11th Joint Conference, JCKBSE 2014*, pp. 27–39. Springer, Cham, Heidelberg, New York, Dordrecht, London (2014)
5. Fedorova, A.A., Beliautsov, V.A., Anikin, I.V.: Prediction vehicle's speed with using artificial neural networks. In: *International Russian Automation Conference, RusAutoCon, 2020*. <https://doi.org/10.1109/RusAutoCon49822.2020.9208089> (2020)
6. Mondal, A., Dutta, A., Dey, N., Sen, S.: Visual traffic surveillance: a concise survey. *Front. Artif. Intell. Appl.* **323**, 32–41 (2020)
7. Sun, Y., Zhu, H., Liao, Y., Sun, L.: Vehicle anomaly detection based on trajectory data of ANPR system. In: *IEEE Global Communications Conference, GLOBECOM 2015* (2015)
8. Atluri, G., Karpatne, A., Kumar, V.: Spatio-temporal data mining: a survey of problems and methods. *ACM Comput. Surv.* **51**(4), 2017 (2017)
9. Meng, F., Yuan, G., Lv, S., Wang, Z., Xia, S.: An overview on trajectory outlier detection. *Artif. Intell. Rev.* **52**(4), 437–2456 (2019)
10. Minnikhanov, R., Dagaeva, M., Anikin, I., Bolshakov, T., Makhmutova, A., Mingulov, K.: Detection of traffic anomalies for a safety system of smart city. In: *CEUR Workshop Proceedings*, vol. 2667, pp. 337–342 (2020)

11. Makhmutova, A., Minnikhanov, R., Dagaeva, M., Anikin, I., Bolshakov, T., Khuziakhmetov, I.: Intelligent detection of object's anomalies for road surveillance cameras. In: SIBIRCON 2019—International Multi-Conference on Engineering, Computer and Information Sciences, Proceedings, pp. 762–767 (2019)
12. Miao, Y., Liu, F., Hou, T., Liu, L., Liu, Y.: A nighttime vehicle detection method based on YOLO v3. In: Proceedings - 2020 Chinese Automation Congress, CAC 2020, pp. 6617–6621 (2020)
13. Zhang, K., Ren, H., Wei, Y. Gong, J.: Multi-target vehicle detection and tracking based on video. In: Proceedings of the 32nd Chinese Control and Decision Conference, CCDC 2020, pp. 3317–3322 (2020)
14. Chen, M., Xu, M., Franti, P.: A fast  $o(n)$  multiresolution polygonal approximation algorithm for gps trajectory simplification. In: IEEE Transactions on Image Processing, vol. 21(5), pp. 2770–2785 (2012)
15. Douglas, D., Peucker, T.: Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Can. Cartograph.* **10**(2), 112–122 (1973)
16. Koning, E.: Polyline simplification. <https://www.codeproject.com/Articles/114797/Polyline-Simplification> (2011)
17. Toohey, K.: R package documentation. In: Similarity Measures. LCSS. <https://rdr.io/cran/SimilarityMeasures/man/LCSS.html>. Last accessed 14 Oct 2021
18. Toohey, K., Duckham, M.: Trajectory similarity measures. *Sigspat. Spec.* **7**(1), 43–50 (2015)
19. Zhang, Z., Huang, K., Tan, T.: Comparison of similarity measures for trajectory clustering in outdoor surveillance scenes, vol. 3, pp. 1135–1138 (2006)
20. Contreras, P., Murtagh, F. (2015): Hierarchical clustering. In: *Handbook of Cluster Analysis*. CRC Press pp. 103–124 (2015)
21. Dunn, J.C.: Well-separated clusters and optimal fuzzy partitions. *J. Cybernet.* **4**(1), 95–104 (1974)
22. Ansari, Z., Azeem, M.F., Ahmed, W., Babu, A.: Quantitative evaluation of performance and validity indices for clustering the web navigational sessions. *World Comp. Sci. Inform. Technol. (WCSIT) J.* **1**(5), 217–226 (2011)
23. Ghrab, N.B., Fendri, E., Hammami, M.: Abnormal events detection based on trajectory clustering. In: 13th International Conference on Computer Graphics, Imaging and Visualization (CGiV), pp. 301–306 (2016)



# Analyzing the Content of Business Documents Recognized with a Large Number of Errors Using Modified Levenshtein Distance



Oleg Slavin , Vera Farsobina, and Aleksei Myshev

**Abstract** The chapter discusses methods for analyzing test material of business documents, lists tasks that use word comparison. The features of the analysis of recognized texts are indicated. A mechanism for identifying recognized words based on textual feature points is described. The advantages and disadvantages of Levenshtein position are listed. Other distances between string objects are described: Jaro-Winkler similarity, multiset metric, MFKC metric. The Levenshtein standard distance is compared to other distances between two string objects. A modification of the Levenshtein position is proposed focused on the features of the recognized characters. Experimental results are presented that demonstrate the effect of using the proposed distance in comparison with the normalized Levenshtein distances. The experiments investigate the extraction of data from the document and the classification of documents. We also compared the time spent on calculating the modified Levenshtein metric and the multiset metric. The proposed method can be applied in a modern CAD system in the recognition component to analyze the information of recognized text documents. Also, the method can be in the system of the analysis of the recognized text using the methods of computational linguistics.

**Keywords** Document analysis · Document recognition · Levenshtein distance

---

O. Slavin (✉) · V. Farsobina  
Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, 9  
Prospect 60-Letiya Oktyabrya, Moscow 117312, Russia  
e-mail: [oslavin@isa.ru](mailto:oslavin@isa.ru)

O. Slavin  
Moscow Institute of Physics and Technology (National Research University), Moscow Region, 9  
Institutskiy Per., Dolgoprudny 141701, Russia

O. Slavin · A. Myshev  
Smart Engines Service, 9 Prospect 60-Letiya Oktyabrya, Moscow 117312, Russia

O. Slavin · V. Farsobina · A. Myshev  
National Research Nuclear University MEPhi, Moscow, Russia

## 1 Introduction

Analyzing textual content of documents is one of the tasks of computational linguistics which is designing a textual or a linguistic model, formally describing the attributes of the analyzed text. The main tasks of computational linguistics are:

- extracting information from texts (highlighting the named entities and keyword fields);
- classifying and rubricating texts and topic modeling;
- subject indexing;
- sentiment analysis;
- machine translation.

Many operations in the abovementioned tasks are based on comparing words with the alphabet in the database [1, 2]. The texts, devised for analysis implicitly, contain a negligible number of errors.

Levenshtein distance, based on inserting, deleting, and substituting characters, is a popular method of comparing words [3–5]. There are well-known algorithms of computing Levenshtein distance which help to find both the number of operations for an edit distance (hereinafter referred to as edit operations) and their sequence. The original Levenshtein distance between two strings  $S_1$  and  $S_2$  is defined in [6] as the minimum number of editing operations for transforming  $S_1$  into  $S_2$ . The original Levenshtein distance is calculated as follows:

$$\rho_{LEV}(S_1, S_2) = M_{LEV}(|S_1|, |S_2|)$$

$$\forall j M_{LEV}(0, j) = 0, \forall i M_{LEV}(i, 0) = 0 \quad (1)$$

$$M_{LEV}(i, j) = \min \left( \begin{array}{l} M_{LEV}(i, j - 1) + 1, M_{LEV}(i - 1, j) + 1, \\ M_{LEV}(i - 1, j - 1) + \text{substCost}(S_{1i}, S_{2i}) \end{array} \right)$$

where  $\text{substCost}(S_{1i}, S_{2i})$  is cost of substituting character  $S_{1i}$  with character  $S_{2i}$ ,  $|S_1|$  and  $|S_2|$  are the lengths of words  $S_1$  and  $S_2$ . The algorithm of computing Levenshtein distance has a quadratic time complexity.

The subject of the chapter refers to the systems of Document Image Processing and Document Understanding. The relevance of the study is characterized by a significant number of studies on the topic, see [7, 8]. The proposed modification is devised for administrative documents [9].

## 2 Relevance

Levenshtein distance is an effective method for comparing both long sequences of strings, for example, to analyze DNA strands, and short ones, like words of natural languages. Nonetheless, when analyzing and processing texts, there arises a need for an equation of two widely differing words. For example, a large Levenshtein distance between the words  $s1s2s3s4s5c1c2c3c4$  and  $c1c2c3c4s1s2s3s4s5$  identifies them as not similar, while during text classification an expert may call these words semantically similar.

To reduce the margins of error, other metrics can be applied. For two words  $S1$  and  $S2$ , which have  $m$  matching characters and  $t$  transpositions, the Jaro–Winkler similarity [3] with a given  $m = 0$  equals 0 and otherwise is defined as follows:

$$s_{JW}(S_1, S_2) = \frac{1}{3} \cdot \left( \frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{m - t}{m} \right)$$

where two characters  $s1i \in S1$  and  $s2j \in S2$  are considered matching if their value is equal ( $s1i = s2j$ ), the number of transpositions is defined as the number of matching characters with different character sequences, the following is true:

$$|i - j| < \left\lceil \frac{1}{2} \cdot \max(|S_1|, |S_2|) \right\rceil - 1$$

The Jaro–Winkler similarity gives higher assessments to the words with the finite prefix. By specifying the finite prefix  $l$  and coefficient  $p$ , we can compute Jaro–Winkler distance

$$\rho_{JW}(S_1, S_2) = (1 - lp)(1 - s_{JW}(S_1, S_2))$$

which does not satisfy the identity axiom and does not obey the triangular inequality. The advantage of  $\rho_{JW}(S1, S2)$  is the opportunity to equate words with a significant number of transpositions.

The distance between strings may be based on string hashing. The study [10] represents the word as a sequence  $k$  ( $k = 2$ ) of the most frequent characters in this word (most frequent  $k$  characters, MFKC) with a specified frequency rate. For example, the word  $s1s2s1s3s2s1$  is represented by a vector  $s13s22$ . The study [10] represents the distance between the two words as follows:

$$s_1(S_1)^{k_1(S_1)}, s_2(S_1)^{k_2(S_1)}, s_1(S_2)^{k_1(S_2)}, s_2(S_2)^{k_2(S_2)}$$

The distance is computed according to the following algorithm:

def similarity := 0

if  $s_1(S_1)^{k_1(S_1)} == s_1(S_2)^{k_1(S_2)}$  then similarity: = similarity +  $k_1(S_1) + k_1(S_2)$

if  $s_1(S_1)^{k_1(S_1)} == s_2(S_2)^{k_2(S_2)}$  then similarity: = similarity +  $k_1(S_1) + k_2(S_2)$

if  $s_2(S_1)^{k_2(S_1)} == s_1(S_2)^{k_1(S_2)}$  then similarity: = similarity +  $k_2(S_1) + k_1(S_2)$

if  $s_2(S_1)^{k_2(S_1)} == s_2(S_2)^{k_2(S_2)}$  then similarity: = similarity +  $k_2(S_1) + k_2(S_2)$

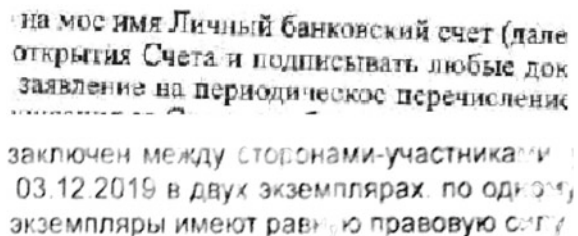
The result should be the mean of  $\rho\text{SDF}(S_1, S_2) = \text{limit} - \text{similarity}$ , where the limit is pre-determined. Thus, the computed distance would not satisfy the identity axiom and would not obey the triangle inequality. Due to the simplified representation, the accuracy of the method is less than that of the method of computing Levenshtein distance; this is proved by the results given in the article [10]. It is evident that with growing parameter  $k$  the accuracy grows as well. The advantages of MFKC are less complex algorithm ( $O(n \cdot \log(n))$ ) compared to the algorithm of computing Levenshtein distance.

The MFKC method with no limitation of the number  $k$  for the strings  $S_1$  и  $S_2$  is equivalent to calculating the metric of the multiset [11]. The multisets  $\text{MS}(S_1)$  and  $\text{MS}(S_2)$  supporter  $\text{Supp}(\text{MS}(S_1)) = \text{Supp}(\text{MS}(S_2))$  are the alphabet of valid characters. The frequencies of each element in the multiset correspond to the character frequency per string. Therefore, the difference between the frequencies of each element can be assessed with the power of symmetric difference of the multisets [11]:

$$\rho_{\text{MS}}(S_1, S_2) = |\text{MS}(S_1) \Delta \text{MS}(S_2)| \quad (2)$$

When analyzing texts extracted from recognized documents, first and foremost, we should take into consideration the possibility of a significant number of errors in the characters recognized. The recognition mistakes arise from the imperfection of OCR, as well as from the defects of text digitization with a scanner or a smartphone camera such as aberrations, distortions, projective distortions, noisy pictures from the scanner, distortions caused by folded or bruised paper, and small-sized fonts. The examples of noisy pictures are represented in Fig. 1. New works are being published on the topic of processing noisy images, for example, see [12, 13]. The flaws of OCR are characterized by the number of misrecognized margins of words, strings, and paragraphs, rather than the number of errors in a specific character. A significant number of potential errors should be considered when analyzing recognized texts as misrecognized characters may decrease the accuracy of text analysis. Levenshtein distance may be more adequate for considering additional errors than the distances which allow significant changes in edit operations.

**Fig. 1** The examples of noisy and lighted text pictures



на мое имя Личный банковский счет (далее  
открытия Счета и подписывать любые док  
заявление на периодическое перечисление

заключен между сторонами-участниками  
03.12.2019 в двух экземплярах. по одному  
экземпляры имеют равную правовую силу

The study proposes modified Levenshtein distance which considers potential errors in recognizing administrative documents [9]. Administrative documents, therefore, are understood to mean documents in a company's administrative work, which have specific attributes. The extracted attributes (fields) are used in the company's information system. One of the characteristics of administrative documents is a limited thesaurus of a static text. Besides, administrative documents are flexible in their structure: fonts, word order, and allowed insignificant changes in the text. The study presupposes that the static texts are printed, not hand-written ones. The chapter develops the results of the study [14, 15].

### 3 Problem Statement

A model of a one-page document  $D$  is a set of strings  $\{L_i^d\}_{i=1}^{|D|}$ , each string  $L_i^d$  is a set of words  $\{W_j^l\}_{j=1}^{|L_i^d|}$ . Strings can be grouped into paragraphs or chapters. A word  $W$  is defined in [4] as a textual feature point

$$W = (T(W), L(W), B(W), Z(W), F(W)) \quad (3)$$

where.

- $T(W)$  – *the nucleus* of a textual feature point, i.e. the sequence of characters in a word, consisting of characters of a specific alphabet or a sequence of character locations with assessments of compliance between a character location and a character;
- $L(W)$  – parameters for modified Levenshtein distance,  $L(W) = \{d(W), FA(W), M(W), D(W)\}$ ;
- $B(W)$  – *the frame* of the textual feature point, consisting of boundary coordinates  $B_{x1}(W), B_{y1}(W), B_{x2}(W), B_{y2}(W)$ ; boundary coordinates can be normalized within a certain range;
- $Z(W)$  – the zone consisting of the boundary coordinates  $Z_{x1}(W), Z_{y1}(W), Z_{x2}(W), Z_{y2}(W)$ , on the allowable location of the textual key point within a normalized page;
- $F(W)$  – *attributes* of the textual feature point, e.g., the typeface and modified versions of a font.

The parameters of the textual key point  $L(W)$  can be specified either for a specific point or a group of points.

The textual feature point is similar to the graphic feature point which is understood to mean a point that satisfies the following criteria:

- the difference from the points in its neighborhood,
- resistance to image noise,
- resistance to specific transforms (e.g., affine transforms or scaling, see [16]).

The attributes of feature points are as follows:

- repeatability – the feature point must be in the same area of the image of the object, despite viewpoint and illumination changes during digitization,
- distinctiveness/informativeness – the neighborhoods of the feature points must have significant differences between each other,
- locality – the feature point and its neighborhood must occupy a small area of an image,
- quantity – the number of detected feature points must be large enough to detect the objects,
- accuracy – detected feature points must have an accurate location, both in the original image and in the scaled picture,
- efficiency – the feature point detection time must be allowable for time-sensitive applications.

The feature points are compared by a similarity measure  $d$ , which must be close to zero when two points, corresponding to the same area in the image of a document, are compared and must take big values when comparing points from different image areas.

The textual feature point presupposes differences from the next textual feature points in its neighborhood. If a neighborhood is understood to mean a text string, then many words in administrative documents differ from the adjacent words in the string. Several identical words in the same string are not textual feature points. However, if a neighborhood is understood to mean one or two adjacent words, then two adjacent words in the same string, with different words next to them, are also textual feature points.

A feature point descriptor is an identifier used in comparing feature points. The descriptor must be invariable when comparing feature points concerning image transformations. Comparing textual feature points may be based on the nucleus  $T(W)$  and attributes  $F(W)$ .

The method of extracting feature points from the image is called the detector. The detector of the textual feature point is a recognition procedure with the help of some OCR, which extracts feature point descriptors from the image of the document. The image of a document page is scanned from a paper page by a scanner or a digital camera. The abovementioned attributes of feature points are fair for every textual feature point if modern OCR can compensate for different image distortions. The uniqueness of textual feature point descriptors is determined by the document structure (one-to-one division of a document into constellations – chapters, paragraphs, and strings) and the attributes of the natural language (a rare coincidence of two adjacent words within the same document).

Comparing two textual feature points in this study is based on modified Levenshtein distance  $\rho\text{LEV}$ . To compare a textual feature point  $W$  and a recognized word  $W_r$ , a parameter  $d(W)$  is used, being a component of parameter array  $L(W)$ . If  $\rho\text{LEV}(T(W), W_r) < d(W)$ , then the word  $W_r$  and the textual feature point  $W$  are identical.

Different relations between textual feature points (higher-lower relations, right-left relations, or geometric distance between the frames) allow unifying points into constellations with clustering algorithms.

### 4 Modified Levenshtein Distance

Let us list some of the known specifics of the original Levenshtein distance. We will consider comparing recognized words with potential errors and correct words from the vocabulary given. The number  $d(W)$  of allowable edit operations must be different for the words of different lengths.  $d(W)$  must not exceed the length of short words, less a certain value. While for a short word with a length equal to 1 the limit  $d(W)$  must be equal to 0, for long words the limit  $d(W)$  may be sufficiently large. To consider this condition, the normalized Levenshtein distance can be applied, e.g. the one from the following study [6]:

$$\tilde{\rho}_{LEV}(S_1, S_2) = \frac{2 \cdot \rho(S_1, S_2)}{|S_1| + |S_2| + \rho(S_1, S_2)}$$

The normalized Levenshtein distance is also a metric, i.e. it satisfies the identity axioms, the symmetry axiom, and the triangular inequality. There are two reasons for considering potential errors in recognized words. First, besides making insignificant mistakes in specific characters in bold-type documents, some OCR programs allow numerous word recognition and margin recognition errors in noisy, lighted, and distorted texts. That is, we cannot determine the limit  $d(W) = 1$  for recognized texts due to the risks of ignoring a significant number of recognized words. Second, the sensible limit  $d(W) = 2$  will lead to the identity of the compared words from Table 1, located far enough from each other.

Similar considerations refer to the distance  $\rho_{MS}$ . Table 2 illustrates the differences in the values of distances between two adjacent words, computed with the help of  $\rho_{LEV}$ ,  $\rho_{SDF}$ , and  $\rho_{MS}$ . The distance  $\rho_{MS}$  in many cases of word distortion gives the

**Table 1** Examples of the words identical according to Levenshtein metric  $\rho_{LEV}(W_1, W_2)$

Reference word $W_1$	Lengths of word $ W_1 / W_2 $	Compared word $W_2$	$\rho_{LEV}(W_1, W_2)$	$\tilde{\rho}_{LEV}(S_1, S_2)$
ADDRESS	6/6	UNDRESS	2	0.29
STREET	5/4	FACES	1	0.20
ISSUE	6/5	ISSUED	2	0.30
OPENING	8/6	OPEN	2	0.25
TF-3201-296	11/11	TF-3201-258	2	0.17
TF-3201-302	11/11	TF-3201-306	1	0.08

**Table 2** Examples of computing  $\rho_{LEV}$ ,  $\rho_{SDF}$  (for  $k = 2$ , limit = 8) and  $\rho_{MS}$

Reference word $W_1$	Compared word $W_2$	$\rho_{LEV}(W_1, W_2)$	$\rho_{SDF}(W_1, W_2)$	$\rho_{MS}(W_1, W_2)$
$s_1s_2s_3s_4$	$s_1s_5s_5s_3s_4$	2	6	3
$s_1s_2s_3s_4s_5$	$s_1s_2s_3s_4s_6$	1	4	2
$s_1s_2s_3$	$s_4s_5s_3$	2	8	4
$s_1s_2s_3s_4s_5s_6s_7s_8$	$s_9$	9	8	9
$s_1s_2$	$s_1s_2s_3s_4s_5s_6s_7$	5	4	5

result, similar to  $\rho_{LEV}$ . Edit operations like substitution are taken into consideration when computing  $\rho_{LEV}$  as  $\text{substCost}(S_{1i}, S_{2i})$ , whereas computing  $\rho_{MS}$  is considered 2, deletion and insertion are considered 1 and 2 respectfully. Unlike  $\rho_{LEV}$ , changes in the character order of the words and the distance  $\rho_{MS}$  are not considered.

Describe the proposed modified Levenshtein distances.

The first change is related to the cost of edit operations (See (1)). The cost of deleting and inserting characters remains equal to 1, but the cost of substitution  $\text{substCost}'(s_1, s_2)$  is computed depending on the values of the characters  $s_1$  and  $s_2$ . It is known that characters with similar lettering can substitute each other during OCR. In the Latin alphabet, these are “B8”, “DO”, “II”. When computing Levenshtein distance to compensate for such errors, a lower cost of substituting characters with similar lettering is proposed. The costs of character substitution for the characters with similar lettering are chosen during the training process. The choice of pairs of similar characters is determined by the features of the used OCR. Similar characters and substitution costs are described with the set  $FA(W)$ . According to definition (3), different groups of textual feature points can have different costs of substituting different characters.

A document page may contain pairs of words with the major character match and a difference in a small number of characters in a certain area, most commonly, in the suffix of the string. The examples are given in Table 1. Such words can be identifiers or abbreviations. There are other possible causes of misidentification of similar words with a common stem and different suffixes and prefixes. To exclude such cases, we propose to use the following patterns  $M(W)$

$$b_1b_2 \dots b_k \dots m_1m_2 \dots m_p \dots e_1e_2 \dots e_q$$

with given characters in the beginning, in the middle, and at the end of the word. If compared characters of a recognized word do not satisfy the pattern, then Levenshtein distance grows by the given penalty. The patterns  $M(W)$  constitute the set of parameters  $L(W)$  of the textual feature point.

As the array of word pairs can be quite large ( $O(10,000)$  word pairs for a page), to reduce time costs on word comparison, we propose using limits for differences in the lengths of words compared. For a textual feature point, a limit  $D(W)$  may be given



$$||S1| - |S2|| < D(W)$$

The parameter both optimizes high-speed performance and prohibits false identification of the words with one prefix and an optional suffix.

The proposed modified Levenshtein distance is computed with a recurrence formula:

$$\begin{aligned} \rho_{LEV}(S_1, S_2) &= M_{LEV}(|S_1|, |S_2|) + \text{Pen}(M(S_1), S_2) + \text{Pen}(M(S_2), S_1) \\ \forall j M_{LEV}(0, j) &= 0, \forall i M_{LEV}(i, 0) = 0 \end{aligned} \quad (4)$$

$$M_{LEV}(i, j) = \min \left( M_{LEV}(i, j-1) + 1, M_{LEV}(i-1, j) + 1, M_{LEV}(i-1, j-1) + \text{substCost}(S_{1i}, S_{2j}) \right)$$

where  $\text{Pen}(M(S_i), S_j)$  is the penalty for unmatching the recognized word  $S_j$ , computed with the help of the pattern  $M(S_1)$  of modified Levenshtein distance.

In the same way, we can figure out penalties for the metric of the multiset (See Formula (2)). The number of patterns for  $\rho_{MS}$ , compared with the number of patterns, for  $\rho_{LEV}$  will increase significantly, because potential changes in the word order, allowable for  $\rho_{MS}$ , may require concealing.

## 5 Results of Experiments

The method was proved on our own test set, consisting of images of typical document samples: Agreement, Letter of attorney, and Questionnaire (KYC) scanned with a resolution of from 100 to 300 dpi and different digitalization quality. Such sets of administrative documents are not publicly available due to personal data in them [17]. The subject of the research was the accuracy of recognizing fields of flexible documents; recognition was based on snapping field margins of textual key points. To recognize the texts, we used OCR Smart ID Reader, based on neural network recognition algorithms [18], optimized for different architectures [19]. The experiment assessed changes in the accuracy of recognizing fields of flexible documents with original Levenshtein distance substituted with its modification. Tables 3 and 4 demonstrate the results of the experiment, which illustrate the effectiveness of the proposed method of identifying the distance between reference and recognized words.

Moreover, the application of modified Levenshtein distance was proved when classifying types of documents. The test relies on our own test set consisting of images of Power of attorney document images of two subclasses, which differ by an insignificant number of keywords. The amounts of documents in the subclasses D1 and D2 were 120 and 131 respectfully. The classification algorithm was based on

**Table 3** Comparison of errors when using original (1) and modified (2) Levenshtein distance

Document	Number of documents	Number of fields in a document	Number of errors		Error rate%	
			1	2	1	2
Agreement	167	35	549	533	13,57	12,38
Letter of attorney	130	111	273	73	6,13	1,63
KYC	443	167	4188	3466	5,69	4,71

**Table 4** Comparison of snap errors when using original (1) and modified (2) Levenshtein distance

Document	Number of documents	Number of fields in a document	Number of errors		Error rate%	
			1	2	1	2
Agreement	167	35	311	246	7,69	6,08
Letter of attorney	130	111	269	71	6,04	1,59
KYC	443	167	1528	233	2,07	0,32

searching a structured set of words in the text of the recognized document. An ideal set was parameterized with a set of textual feature points, united with relations of order, and grouped into a disjunctive normal form.

The allocation of words was defined as an ordered set of words  $R = \{W_1, W_2, \dots\}$ , for which each word from the recognized texts would be checked and allowable distance  $d_1(W_i, W_i + 1)$  may be given for each pair  $W_i$  and  $W_i + 1$ .

The assessment of location  $\rho(R)$  was computed as  $\max(\rho(W_1), \rho(W_2), \dots)$ . An allowable limit of every location was set.

The combination of words was defined as a set of locations  $S = \{R_1, R_2, \dots\}$ , for which each word location from the recognized texts would be checked. The order of locations in the combination does not matter. An allowable assessment limit of every location was set.

The assessment of combination  $\rho(S)$  was computed as  $\max(\rho(R_1), \rho(R_2), \dots)$ .

The model of classification was defined as a set of combinations  $M = \{S_1, S_2, \dots\}$  for which each combination from the recognized texts would be checked. The order of combinations in the model does not matter, it was only required that the assessment of at least one combination should not exceed the given limit.

The model of the model  $\rho(M)$  was computed as  $\min(\rho(S_1), \rho(S_2), \dots)$  [14].

The described model considers the word order and corresponds to the notion of statistically resistant n-grams and vector representations. When training models M, collocation search methods like TopMine [20] or SegPhrase [21] may be used. Being a simple method, TopMine scales linearly and help to form the vocabulary, where every n-gram has three attributes:

**Table 5** Comparison of quality of classification when using the original Levenshtein distance

D <sub>1</sub>		D <sub>2</sub>	
$tp = 119$	$fp = 1$	$tp = 130$	$fp = 1$
$tn = 115$	$fn = 16$	$tn = 120$	$fn = 0$
$Precision = 0,99$	$Recall = 0,82$	$Precision = 0,99$	$Recall = 1$

**Table 6** Comparison of quality of classification when using modified Levenshtein distance

D <sub>1</sub>		D <sub>2</sub>	
$tp = 119$	$fp = 0$	$tp = 131$	$fp = 0$
$tn = 131$	$fn = 0$	$tn = 120$	$fn = 0$
$Precision = 1$	$Recall = 1$	$Precision = 1$	$Recall = 1$

- possessing a high frequency in the collection;
- containing words, not randomly often constituting  $n$ -grams;
- being absent in  $(n + 1)$ -grams, which possess the abovementioned attributes.

The quality of the algorithm was evaluated by the following criteria:

- $Precision = tp / (tp + fp)$ ,
- $Recall = tp / (tp + fn)$ ,

where the following values were used:

- the number of correctly classified documents of their class  $tp$ ;
- the number of rejected classified documents of their class  $fp$ ;
- the number of rejected classified documents of the other class  $tn$ ;
- the number of incorrectly classified documents of the other class  $fn$ .

The results of classification are demonstrated in Tables 5 and 6.

The data from Tables 5 and 6 characterize the effectiveness of the application of modified Levenshtein distance.

The described modification can be applied when compensating for the errors in the metric of the multiset  $\rho_{MS}$ . The number of descriptions of penalty functions for the metric  $\rho_{MM}$  will be significantly bigger than the number of similar descriptions of  $\rho_{LEV}$ . This is clear from the invariability of the metric  $\rho_{MS}$  to the changes in character order in words.

The advantage of the metric  $\rho_{MS}$  is the linear complexity of its algorithm. Like the Hamming metric, the metric  $\rho_{MS}$  can be effectively applied in different computing architectures [22]. Nonetheless, when applied to words of natural languages, the quadratic time complexity of the computing algorithm for  $\rho_{LEV}$  does not always take much time on the application in a specific CPU instruction system. The reason is related to the limits of lengths of words in natural languages.

Table 7 shows a comparison of time spent on computing the  $\rho_{LEV}$  and  $\rho_{MS}$  metrics. The assessment was conducted on the arrays of static words in administrative documents. There were  $1363^2 = 1\ 857\ 769$  comparisons made. The experiments were

**Table 7** Comparison of high-speed performance of the original and modified Levenshtein distances

metric	time
$\rho_{MS}$	$1,78 \cdot 10^{-4}$ ms
$\rho_{LEV}$	$1,42 \cdot 10^{-4}$ ms

carried out on a computer with the following characteristics: Intel® Core™ i9-9900 3.60 GHz, DDR 2666 MHz. To build the test, written in C++, we used Microsoft Visual Studio Community 2019, with /O2 optimization parameter.

Table 7 shows that the high-speed performance of  $\rho_{LEV}$  metric slowed by 25%, compared to the  $\rho_{MS}$  metric. The total time cost on computing distances in the Power of attorney documents equals approximately 10–20 ms. Given the fact that recognition time cost was approximately 1 s per document, the advantage of the  $\rho_{MS}$  metric is regarded as insignificant.

## 6 Conclusion

The study proves that the proposed method of computing modified Levenshtein distance can be applied to the analysis of recognized documents. When recognizing textual documents with flexible structure, the use of the proposed word distance increases both accuracies of snapping fields and recognition accuracy. The high-speed performance of computing modified Levenshtein distance, assessed with the help of examples from administrative documents, turned out to be comparable to the high-speed performance of the multiset metric.

**Acknowledgements** The research is carried out with partial financial support of The Russian Foundation for Basic Research (Project 20-07-00934).

## References

1. Kravets, A.G., Salnikova, N.A., Shestopalova, E.L.: Development of a module for predictive modeling of technological development trends. *Cyber Phys. Syst.* 125–136 (2021). [https://doi.org/10.1007/978-3-030-67892-0\\_11](https://doi.org/10.1007/978-3-030-67892-0_11)
2. Sabitov, A., Minnikhanov, R., Dagaeva, M., Katasev, A., Asliamov, T.: Text classification in emergency calls management systems. *Cyber Phys. Syst.* 199–210 (2021). [https://doi.org/10.1007/978-3-030-67892-0\\_17](https://doi.org/10.1007/978-3-030-67892-0_17)
3. Deza, M.M., Deza, E.: *Encyclopedia of distances*. Springer-Verlag, Berlin (2009)
4. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. *Rep. USSR Acad. Sci.* **163**(4), 845–848 (1965)
5. Sankoff, D., Kruskal, J.: Review: time warps, string edits, and macromolecules: the theory and practice of sequence comparison. *J. Log. Comput.* **11**(2), 356–356 (1983). <https://doi.org/10.1093/logcom/11.2.356>
6. Yujian, L., Bo, L.: A normalized Levenshtein distance metric. *IEEE Trans. Pattern Anal. Mach. Intell.* **29**(6), 1091–1095 (2007). <https://doi.org/10.1109/TPAMI.2007.1078>

7. Moysset, B., Kermorvant, C., Wolf, C.: Learning to detect, localize and recognize many text objects in document images from few examples. *IJDAR* **21**, 161–175 (2018). <https://doi.org/10.1007/s10032-018-0305-2>
8. Nagy, G.: Document analysis systems that improve with use. *IJDAR* **23**, 13–29 (2020). <https://doi.org/10.1007/s10032-019-00344-x>
9. Rusiñol, M., Frinken, V., Karatzas, D., Bagdanov, A.D., Lladós, J.: Multimodal page classification in administrative document image streams. *IJDAR* **17**(4), 331–341 (2014)
10. Şecker, ŞE., Altun, O., Ayan, U., Mert, C.: A novel string distance function based on most frequent k characters. *Int. J. Mach. Learn. Comput.* **4**(2), 177–183 (2014). <https://doi.org/10.7763/IJMLC.2014.V4.408>
11. Petrovsky, A.B.: Metrics in multiset spaces. *J. Intell. Fuzzy Syst.* **36**(4), 3073–3085 (2019). <https://doi.org/10.3233/JIFS-18525>
12. Hjouji, A., EL-Mekkaoui, J., Jourhmane, M.: Image classification by mixed finite element method and orthogonal legendre moments. *Pattern Recogn. Image Anal.* **30**, 655–673 (2020). <https://doi.org/10.1134/S1054661820040185>
13. Karkishchenko, A.N., Mnukhin, V.B.: On the metric on images invariant with respect to the monotonic brightness transformation. *Pattern Recogn. Image Anal.* **30**, 359–371 (2020). <https://doi.org/10.1134/S1054661820030104>
14. Slavina, O.A.: Using special text points in the recognition of documents. In: *Studies in Systems, Decision and Control*, vol. 259, pp. 43–53. Springer Nature Switzerland AG (2020). [https://doi.org/10.1007/978-3-030-32579-4\\_4](https://doi.org/10.1007/978-3-030-32579-4_4)
15. Andreeva, E., Arlazarov, V.V., Slavina, O., Mishev, A.: Comparison of scanned administrative document images. In: *Proceedings of SPIE, 2020: Twelfth International Conference on Machine Vision, ICMV 2019*, vol. 11433, pp. 16–18. Amsterdam, Netherlands. (2019). <https://doi.org/10.1117/12.2559369>
16. Schmid, C., Mohr, R.: Local gray value invariants for image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **19**(5), 530–535 (1997). <https://doi.org/10.1109/34.589215>
17. Awal, A.M., Ghanmi, N., Sicre, R., Furon, T.: Complex document classification and localization application on identity document images. In: *Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition*, pp. 427–432. (2017). <https://doi.org/10.1109/ICDAR.2017.77>
18. Chernyshova, Y.S., Sheshkus, A.V., Arlazarov, V.V.: Two-step CNN framework for text line recognition in camera-captured images. *IEEE Access.* **8**, 32587–32600 (2020). <https://doi.org/10.1109/ACCESS.2020.29740511>
19. Limonova, E.E., Neiman-zade, M.I., Arlazarov, V.L.: Special aspects of matrix operation implementations for low-precision neural network model on the elbrus platform. In: *Bulletin of the South Ural State University. Ser. Mathematical Modelling, Programming & Computer Software (Bulletin SUSU MMCS)*, vol. **13**(1), 118–128 (2020). <https://doi.org/10.14529/mmp200109>
20. El-Kishky, A., Song, Y., Wang, C., Voss, C. R., Han, J.: Scalable topical phrase mining from text corpora. In: *Proc. VLDB Endowment*, vol. **8**(3), pp. 305–316 (2014). <https://doi.org/10.14778/2735508.2735519>
21. Liu, J., Shang, J., Wang, C., Ren, X., Han, J.: Mining quality phrases from massive text corpora. In: *Proc. of the 2015 ACM SIGMOD International Conference on Management of Data—SIGMOD* vol. 45, pp. 1729–1744. ACM, New York, NY, USA. (2015). <https://doi.org/10.1145/2723372.2751523>
22. Limonova, E., Skoryukina, N., Neiman-zade, M.: Fast hamming distance computation for 2D art recognition on VLIW-architecture in case of Elbrus platform. In: *Proc. SPIE, Eleventh International Conference on Machine Vision*, vol. 11041. Art. ID: 110411N. (2018). <https://doi.org/10.1117/12.2523101>