

Smart Innovation, Systems and Technologies 220

Nikita Voinov  
Tobias Schreck  
Sanowar Khan *Editors*



# Proceedings of International Scientific Conference on Telecommunications, Computing and Control

TELECCON 2019



 Springer

# **Smart Innovation, Systems and Technologies**

Volume 220

## **Series Editors**

Robert J. Howlett, Bournemouth University and KES International,  
Shoreham-by-Sea, UK

Lakhmi C. Jain, KES International, Shoreham-by-Sea, UK

The Smart Innovation, Systems and Technologies book series encompasses the topics of knowledge, intelligence, innovation and sustainability. The aim of the series is to make available a platform for the publication of books on all aspects of single and multi-disciplinary research on these themes in order to make the latest results available in a readily-accessible form. Volumes on interdisciplinary research combining two or more of these areas is particularly sought.

The series covers systems and paradigms that employ knowledge and intelligence in a broad sense. Its scope is systems having embedded knowledge and intelligence, which may be applied to the solution of world problems in industry, the environment and the community. It also focusses on the knowledge-transfer methodologies and innovation strategies employed to make this happen effectively. The combination of intelligent systems tools and a broad range of applications introduces a need for a synergy of disciplines from science, technology, business and the humanities. The series will include conference proceedings, edited collections, monographs, handbooks, reference books, and other relevant types of book in areas of science and technology where smart systems and technologies can offer innovative solutions.

High quality content is an essential feature for all book proposals accepted for the series. It is expected that editors of all accepted volumes will ensure that contributions are subjected to an appropriate level of reviewing process and adhere to KES quality principles.

Indexed by SCOPUS, EI Compendex, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST), SCImago, DBLP.

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

More information about this series at <http://www.springer.com/series/8767>

Nikita Voinov · Tobias Schreck · Sanowar Khan  
Editors

Proceedings of International  
Scientific Conference  
on Telecommunications,  
Computing and Control  
TELECCON 2019

 Springer

*Editors*

Nikita Voinov  
Higher School of Software Engineering  
Peter the Great St. Petersburg Polytechnic  
University  
St. Petersburg, Russia

Tobias Schreck  
Department of Computer Science  
and Biomedical Engineering  
Graz University of Technology  
Graz, Austria

Sanowar Khan  
School of Maths, Computer Science  
and Engineering  
City, University of London  
London, UK

ISSN 2190-3018

ISSN 2190-3026 (electronic)

Smart Innovation, Systems and Technologies

ISBN 978-981-33-6631-2

ISBN 978-981-33-6632-9 (eBook)

<https://doi.org/10.1007/978-981-33-6632-9>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

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 Singapore Pte Ltd.

The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

# Preface

This volume contains the papers that were presented at the First International Scientific Conference on Telecommunications, Computing and Control (TELECCON-2019) organized by Peter the Great St. Petersburg Polytechnic University during November 18–19, 2019. It provided a great platform for researchers from across the world to report, deliberate, and review the latest progress in the cutting-edge research pertaining to telecommunications, signal processing, artificial intelligence, intelligent control systems, modeling, and simulation.

We would like to express our appreciation to the members of the program committee for their support and cooperation in this publication. We are also thankful to the team from Springer for providing a meticulous service for the timely production of this volume.

Special thanks to all the guests who have honored us in their presence at the conference. Our thanks are due to all session chairs, track managers, and reviewers for their excellent support. Last but not least, our special thanks go to all the authors who submitted papers and all the attendees for their contributions and fruitful discussions that made this conference a great success.

St. Petersburg, Russia  
Graz, Austria  
London, UK

Nikita Voinov  
Tobias Schreck  
Sanowar Khan

# Contents

<b>Generative Adversarial Network for Detecting Cyber Threats in Industrial Systems</b> .....	1
Vasiliy Krundyshev and Maxim Kalinin	
<b>Detection and Prediction of Safety Faults in Inter-Device Networks Applying a Set of Data-Driven Methods</b> .....	15
Maxim Kalinin, Vasiliy Krundyshev, Viacheslav Belenko, and Valery Chernenko	
<b>Development of an Algorithm for Determining the Railway Tracks on Video Image</b> .....	27
Ivan Deylid, Sergey Molodyakov, and Boris Tyutin	
<b>Hardware and Software System for Collection, Storage and Visualization Meteorological Data from a Weather Stand</b> .....	37
Pavel Pankov, Igor Nikiforov, and Yufeng Zhang	
<b>Plant Disease Recognition Based on Multi-dimensional Features of Leaf RGB Images</b> .....	49
Basim Al-Windi and Vladimir Tutygin	
<b>Methodology of Service Development with a Single Application Programming Interface</b> .....	67
Vitaly Monastirev, Pavel Drobintsev, and Petar Kochovski	
<b>Using Symbolic Computing to Find Stochastic Process Duration Distribution Laws</b> .....	77
Georgiy Zhemelev and Alexandr Sidnev	
<b>Comparison of the Shape of Digital Models of Pump Components</b> .....	99
Evgeniy Ivanov, Aleksandr Zharkovskii, Igor Borshchev, and Arsentiy Klyuyev	

<b>Parametric Oscillations of Viscoelastic Orthotropic Rectangular Plates of Variable Thickness</b> .....	109
Rustamkhan Abdikarimov, Bakhodir Normuminov, Dadakhan Khodzhaev, and Davron Yulchiyev	
<b>Methods and Technologies for Protecting Pharmaceutical Products in Polymer Packaging from Counterfeiting</b> .....	119
Tamara Chistyakova, Roman Makaruk, Ilya Sadykov, and Christian Kohlert	
<b>Solving Multicriteria Optimization Problem for an Oil Refinery Plant</b> .....	131
Dmitri Kostenko, Vyacheslav Shkodyrev, and Vadim Onufriev	
<b>Methods and Techniques for Increasing the Accuracy of Continuous Non-invasive Blood Pressure Measurement Under Dynamic Loads</b> .....	141
Gleb Zaitsev, Alexei Vassiliev, and Quang-Kien Trinh	
<b>Computer Modeling of Robust Control of Vibrationless Movement of Multi-mode Flexible Structures</b> .....	153
Vladimir A. Prouzin, Kiseon Kim, and Georgy Shevlyakov	
<b>Feature-Based Plant Seedlings Classification</b> .....	165
Dmitri Jakovlev, Iuliia Kamaletdinova, and Georgy Shevlyakov	
<b>Medical Training Simulation in Virtual Reality</b> .....	177
Vladimir Ivanov, Sergey Strelkov, Alexander Klygach, and Dmitry Arseniev	
<b>Application of the Hybrid Model to Numerical Modeling of the Urban Transport Network Topology</b> .....	185
Vadim Glazunov, Mikhail Chuvatov, Leonid Kurochkin, Mikhail Kurochkin, Alexander Chernyshev, and Leonid Hanin	
<b>Synchronization Scheme for UWB Wireless Sensor Network System</b> ...	195
Iuliia Tropkina, Sergey Zavjalov, and Dong Ge	
<b>The Deep Survival Forest and Elastic-Net-Cox Cascade Models as Extensions of the Deep Forest</b> .....	205
Lev Utkin, Andrei Konstantinov, Anna Meldo, Victoria Sokolova, and Frank Coolen	
<b>An Explanation Method for Siamese Neural Networks</b> .....	219
Lev Utkin, Maxim Kovalev, and Ernest Kasimov	
<b>Hierarchical Multi-agent System for Production Control Using KPI Reconciliation</b> .....	231
Vladislav Kovalevsky, Vadim Onufriev, and Anton Dybov	
<b>Semi-supervised Learning for Medical Image Segmentation</b> .....	245
Mikhail Kots, Mikhail Pozigun, Andrei Konstantinov, and Viacheslav Chukanov	



**Developing a New Generation of Reconfigurable Heterogeneous Distributed High Performance Computing System** ..... 255  
Alexander Antonov, Vladimir Zaborovskij, and Ivan Kisilev

**Usage of a BART Algorithm and Cognitive Services to Research Collaboration Platforms** ..... 267  
Sergey Saradgishvili and Ilia Voronkov

**A Computer-Aided Diagnosis System in the Diagnosis of Multiple Sclerosis** ..... 277  
Polina Andropova, Dmitriy Cheremisin, and Anna Meldo

**Predicting Students’ Performance on MOOC Using Data Mining Algorithms** ..... 285  
Sergey Nesterov, Elena Smolina, and Tigran Egiazarov

**On the Implementation of the Planar3D Model Using the Explicit Time Integration Scheme and the Statistical Front Tracking Method** ..... 293  
Egor Starobinskii, Nikita Mushchak, Svetlana Kraeva, Sergei Khlopin, and Egor Shel

**Fast Fourier Transform in Planar3D Model Using an Explicit Numerical Integration Scheme** ..... 307  
Nikita Mushchak, Egor Starobinskii, Sergei Hlopin, and Egor Shel

**Employee Performance Analytics Approach Based on Anomaly Detection in User Activity** ..... 321  
Aleksey Lukashin, Mikhail Popov, Dmitrii Timofeev, and Igor Mikhalev

**Deep Predictive Control** ..... 333  
Dmitry Baskakov and Vyacheslav Shkodyrev

**On the Computational Complexity of Deep Learning Algorithms** ..... 343  
Dmitry Baskakov and Dmitry Arseniev

**Enactivism in the Conceptual Basis of the Non-classical Theory of Management of Ergatic Systems** ..... 357  
Sergey Sergeev, Vladimir Ivanov, and Oleg Ipatov

**The Solution of “If-Problem” in Computations with Multi-valued Variables Based on Operator Overloading** ..... 365  
Vyacheslav Sal’nikov and Konstantin Semenov

**The Interval Method of Bisection for Solving the Nonlinear Equations with Interval-Valued Parameters** ..... 373  
Konstantin Semenov and Anastasia Tselishcheva

**Complex Monitoring Systems for Landfills** ..... 385  
Aleksandr Titov, Sergey Krasnov, Andrey Timofeev, and Victor Denisov

<b>Modeling the Control Object in the Management System of the Regional Socioeconomic System</b> .....	395
Elena Averchenkova	
<b>Transformation, Visualization and Analysis Different Kind of Study Information Contained in the Students' Electronic Portfolio</b> .....	407
Elena Ilina, Yuliya Kocherzhinskaya, Nikita Dyakonov, Daria Arefeva, Tat'yana Antonova, and Il'ya Levandovskii	
<b>An Automated Measuring Complex for Research Parameters of Unmanned Aerial Vehicle</b> .....	419
Oleg Drozd, Pavel Avlasko, Semen Bordyugov, and Denis Kapulin	
<b>Research and Evaluation of the Most Significant Quantitative Characteristics of MPLS Equipment</b> .....	431
Andrey Krasov, Pavel Karelsky, Igor Zuyev, Max Kovzur, and Aleksander Tasyuk	
<b>Study of the Microstrip Waveguide Prototype Model for Use as a Retunable Diffraction Grating</b> .....	445
Dmitry Nikulin, Valery Reichert, Sergey Shergin, Igor Karmanov, Vladimir Korneyev, and Polina Zvyagintseva	
<b>The Use of Digital Cameras for Multispectral Registration with an Unmanned Aircraft</b> .....	451
Evgenij Gritskevich, Sergei Novikov, Polina Zvyagintseva, Diana Makarova, Marina Egorenko, and Aelita Shaburova	
<b>Computer Model for Analysis of the Process of Image Construction in Optical-Electronic Visualization Systems</b> .....	461
Evgenij Gritskevich, Marina Egorenko, Diana Makarova, Sergei Novikov, Alexey Polikanin, and Aelita Shaburova	
<b>Solution of Partial Differential Equations on Radial Basis Functions Networks</b> .....	475
Mohie Alqezweeni and Vladimir Gorbachenko	
<b>Predicting Personality from Image Preferences: Tendencies, Models and Implementation</b> .....	491
Stanislav Krainikovskiy, Mikhail Melnikov, and Roman Samarev	
<b>Power Consumption Meter for Energy Monitoring and Debugging</b> .....	499
Nikita Kulikov, Elena Yaitskaya, Arina Shvedova, and Vladimir Zhalnin	
<b>LoRaWAN Gateway Coverage Evaluation for Smart City Applications</b> .....	513
Vadim Shpenst and Andrei Terleev	

<b>Fire Resistance Evaluation of Tempered Glass in Software ELCUT . . . .</b>	<b>523</b>
Marina Gravit, Nikolay Klimin, Alina Karimova, Evgenia Fedotova, and Ivan Dmitriev	
<b>Author Index . . . . .</b>	<b>539</b>

# About the Editors

**Nikita Voinov** has been Associate Professor of the Higher School of Software Engineering at the Institute of Computer Science and Technology of Peter the Great St. Petersburg Polytechnic University since 2012, becoming Deputy Director for Science and Research in 2018. Dr. Voinov’s research interests run a gamut of topics including the technologies for developing mobile applications, information management and data storage, cloud infrastructure and services, and data science and analysis of big data. The current research projects he is participating in are “Technologies and toolset for reliable control of production areas of Internet of Things” (joint project with the Indian Institute of Technology Bombay, Mumbai, India) and “Methods and technologies for verification and development of software for modeling and calculations using HPC platform with extramassive parallelism.

**Tobias Schreck** has been Professor at the Institute of Computer Graphics and Knowledge Visualization at the Faculty for Computer Science and Biomedical Engineering of Graz University of Technology since 2015. His research interests concern visual data analysis and applied 3D object retrieval. Professor Schreck has acted as a principal investigator in several funded research projects, describing how large collections of data can be visually explored and analyzed for data understanding and decision-making. Another focus of research is how users can interactively search for visual patterns in data, assisted by similarity functions, visual quality measures, and novel search interfaces, including approaches based on user sketching and eye tracking, and how to interactively visualize and steer machine learning approaches like clustering and regression analysis by users, supporting effective and scalable data exploration. Potential applications include, among others, data exploration for engineering and industrial applications, in research data repositories, and digital libraries. Professor Schreck has served as a program co-chair for the IEEE Conference on Visual Analytics Science and Technology in 2017 and 2018, and is currently Associate Editor for IEEE Transactions on Visualization and Computer Graphics.

**Sanowar Khan** has been with the School of Mathematics, Computer Science and Engineering of City, University of London, since 1989. Professor Khan’s research interests cover mathematical modeling and CAD of sensors, actuators and devices,

computational electromagnetics, magnetic shape memory smart materials, finite element modeling and numerical methods, and forward and inverse problems in tomographic imaging. Professor Khan is Honorary Professor of SPPU, Fellow of the Institution of Engineering and Technology, the Institute of Measurement and Control, and a member of the IEEE, IEEE Magnetics, Engineering in Medicine and Biology, and Instrumentation and Measurement societies. He is also a founder member of the International Compumag Society and a member of the Journal Editorial Committee of Measurement and Control. Professor Khan is Editor-in-Chief of the journal Sensor Review. He has been Associate Editor of the journal Measurement, and the International Journal on Measurement Technologies and Instrumentation Engineering.

# Generative Adversarial Network for Detecting Cyber Threats in Industrial Systems



Vasily Krundyshev and Maxim Kalinin

**Abstract** The transition from the information economy to the digital presents new challenges to the community related to the development of breakthrough technologies, a network of cyber-physical systems, artificial intelligence, and big data. When creating digital platforms, a number of difficulties arise: the large dimension of the digital infrastructure and its heterogeneity, poorly established information interaction between the segments, the lack of a common approach to ensuring cybersecurity, and high dependence on personnel qualification and reliability of equipment. The introduction of the digital economy leads to an increase in the risk of cyber threats associated with problems of access control between systems, regulation of information, and control flows. In this paper, for solving cyber threat detection tasks, it is proposed to use generative adversarial neural networks. The paper presents training and testing algorithms of the neural network. The result of the experiments demonstrated high accuracy at cyber threat detection.

**Keywords** Artificial intelligence · Cyber threats · Generative adversarial networks · Neural networks · Industrial systems · Intrusion detection · ANN · GAN

## 1 Introduction

Today, the digital revolution no longer seems to be another slogan of marketers and fiction of science fiction writers. Energy, engineering, transport, and other industries are actively entering the era of digital transformation and are under the influence of changes in Industry 4.0, which includes the Internet of things, blockchain technology, modeling using augmented reality, and much more. According to World Bank estimates, Industry 4.0 could bring the world economy up to \$30 trillion [1]. Digitalization involves accelerating processes and automating production to the level of “one-button problem solving” when the operator is not constrained in movement

---

V. Krundyshev (✉) · M. Kalinin  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [vmk@ibks.spbstu.ru](mailto:vmk@ibks.spbstu.ru)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021  
N. Voinov et al. (eds.), *Proceedings of International Scientific Conference on Telecommunications, Computing and Control, Smart Innovation, Systems and Technologies* 220, [https://doi.org/10.1007/978-981-33-6632-9\\_1](https://doi.org/10.1007/978-981-33-6632-9_1)

and can manage almost the entire enterprise from anywhere. Given the specifics of most enterprises, high security requirements are one of the key conditions in the design of information and telecommunication systems.

In the industry of automated process control systems, a unique situation arose: Several high-profile incidents [2, 3] that led to targeted cyberattacks on a very limited number of objects completely formed an idea of the potential threat, both among developers of protective equipment and among potential users of these tools. Accessible information about many of these incidents was presented exclusively to researchers and developers of protection tools against traditional computer threats. As a result, the main efforts were focused on the technical analysis of the IT component of the corresponding attacks, and the analysis of the cyber-physical component was not given due attention. The resulting reports contain numerous misinterpretations of almost all known incidents that are constantly encountered in the professional environment of engineers, and also proved to be difficult for potential users to understand the means of protection. Emerging manufacturers of new specialized tools for protecting industrial automation systems have developed products that protect, perhaps not so much from real everyday attacks as from synthetic scenarios invented by information security researchers themselves sometimes without relying on practical experience, and their active marketing activities have generated demand for such products. Thus, there was a dangerous situation when industrial automation systems not only became vulnerable to accidental attacks of malicious software not specifically targeted at them, but also attracted the attention of traditional cybercriminals to industrial enterprises, as evidenced by the results of studies published on the ics-cert portal kaspersky.com [4].

Due to the features of modern digital infrastructures and the rapidly growing volume of data being processed, traditional protection methods become ineffective, so researchers are faced with the task of creating new methods for ensuring cybersecurity that take into account the specifics of the object of protection—a modern high-tech enterprise. Many researchers have focused on the use of artificial neural network (ANN) apparatus to identify hidden patterns and deviations [5–7]. Several solutions have already shown their effectiveness in solving the problem of detecting cyber threats, and the fact that the world’s largest companies spend impressive amounts on research and development in this area only confirms the promise of this approach [8–10].

The purpose of the paper is to assess the possibility of using modern generative adversarial networks (GANs) to solve the problem of detecting cyber threats in industrial systems. This paper is organized as follows: Sect. 2 presents mathematical model of GAN; Sect. 3 provides an analysis of studies on the use of GAN in various areas; Sect. 4 describes the proposed architecture of GAN; Sect. 5 presents the training and testing algorithms of the neural network; Sect. 6 presents the results of assessing the quality of the developed solution, and finally a conclusion is provided in Sect. 7.

## 2 Generative Adversarial Network

Generative adversarial network (GAN) is a relatively new class of artificial neural networks, the main purpose of which is a generation of certain objects. Ian Goodfellow from Google proposed this model in 2014 [11]. GAN consists of a combination of two neural networks, one of which generates the objects, and the second one estimates them. The first network is called a generator, and the second one is called a discriminator. GAN is based on the mathematics of game theory. The main idea is to oppose the work of the generator and the discriminator, so that the game was played not to anyone’s victory, but to a mutual balancing of the capabilities of the enemy and bringing the game to a draw. The task of the discriminator, according to the data obtained in the analysis of the business problem, is to build a solution to the problem. The task is reduced to classification, regression, and others. The task of the generator, using noise data  $z$  and sometimes some auxiliary information about the system, is to construct data for the discriminator such that for the discriminator they will become indistinguishable from the data from the true dataset (Fig. 1).

Therefore, the generator also learns to distinguish the fake data provided by the generator from the real data from the original dataset. This is an antagonistic game, a zero-sum game. The task is in mutual learning of the discriminator and the generator, and not in the victory of someone.

Let  $p_z$  be the distribution of noise data  $z$ .  $p_g$  is distribution of data  $x$ , received from the generator.  $p_r$  is the original distribution of data.

The task of the discriminator is to maximize the value of the expectation [12, 13]:

$$E_{x \sim p_r(x)}[\log D(x)] \tag{1}$$

At the same time, to minimize the value of the data obtained from the generator:

$$E_{z \sim p_z(z)}[\log(1 - D(G(z)))] \tag{2}$$

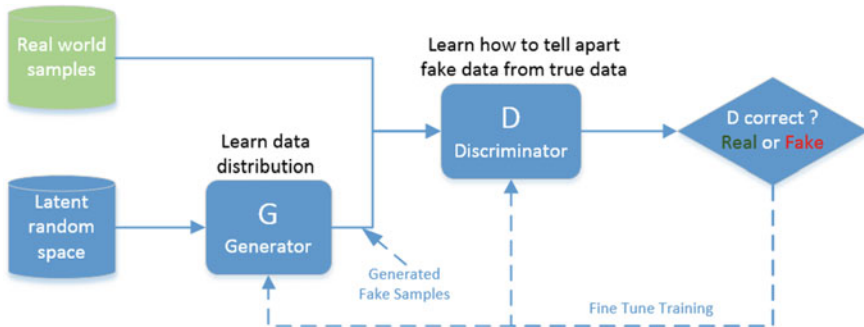


Fig. 1 GAN architecture



Of course, at the same time, the task of the generator is to increase the chances that the discriminator recognizes a fake:

$$E_{z \sim p_z(z)}[\log(1 - D(G(z)))] \quad (3)$$

You can combine these tasks as an error function for minimax games:

$$\begin{aligned} \min_G \max_D L(D, G) &= E_{x \sim p_r(x)}[\log D(x)] + E_{z \sim p_z(z)}[\log(1 - D(G(z)))] \\ &= E_{x \sim p_r(x)}[\log D(x)] + E_{x \sim p_g(x)}[\log(1 - D(x))] \end{aligned} \quad (4)$$

Determine the best discriminator result:

$$L(G, D) = \int_x (p_r(x) \log(D(x)) + p_g(x) \log(1 - D(x))) dx \quad (5)$$

Let  $\tilde{x} = D(x)$ ,  $A = p_r(x)$ ,  $B = p_g(x)$ .

Then:

$$f(\tilde{x}) = A \log \tilde{x} + B \log(1 - \tilde{x}) \quad (6)$$

$$\frac{df(\tilde{x})}{d\tilde{x}} = A \frac{1}{\ln 10} \frac{1}{\tilde{x}} - B \frac{1}{\ln 10} \frac{1}{1 - \tilde{x}} = \frac{1}{\ln 10} \left( \frac{A}{\tilde{x}} - \frac{B}{1 - \tilde{x}} \right) = \frac{1}{\ln 10} \frac{A - (A + B)\tilde{x}}{\tilde{x}(1 - \tilde{x})} \quad (7)$$

Equating the value of the derivative to zero, we get:

$$D^*(x) = \tilde{x}^* = \frac{A}{A + B} = \frac{p_r(x)}{p_r(x) + p_g(x)} \in [0, 1] \quad (8)$$

The total global minimum for the generator and the discriminator is obtained:

$$\begin{aligned} L(G, D^*) &= \int_x (p_r(x) \log(D^*(x)) + p_g(x) \log(1 - D^*(x))) dx \\ &= \log \frac{1}{2} \int_x p_r(x) dx + \log \frac{1}{2} \int_x p_g(x) dx = -2 \log 2 \end{aligned} \quad (9)$$

Let us denote the Jensen–Shannon distance:

$$D_{JS}(p||q) = \frac{1}{2} D_{KL} \left( p \parallel \frac{p+q}{2} \right) + \frac{1}{2} D_{KL} \left( q \parallel \frac{p+q}{2} \right) \quad (10)$$

Distance between distributions  $p_r$  and  $p_g$ :

$$\begin{aligned}
D_{JS}(p_r||p_g) &= \frac{1}{2}D_{KL}\left(p_r||\frac{p_r+p_g}{2}\right) + \frac{1}{2}D_{KL}\left(p_g||\frac{p_r+p_g}{2}\right) \\
&= \frac{1}{2}\left(\log 2 + \int_x p_r(x) \log \frac{p_r(x)}{p_r+p_g(x)} dx\right) \\
&\quad + \frac{1}{2}\left(\log 2 + \int_x p_g(x) \log \frac{p_g(x)}{p_r+p_g(x)} dx\right) = \frac{1}{2}(\log 4 + L(G, D^*))
\end{aligned}
\tag{11}$$

Thus,  $L(G, D^*) = 2D_{JS}(p_r||p_g) - 2 \log 2$ .

To achieve the maximum accuracy of the discriminator, it is necessary to maximize the distance between the real distribution and the distribution of the generator.

### 3 Related Works

The history of artificial intelligence as a new scientific direction begins in the middle of the twentieth century. However, only since the 2010s, with the advent of high-performance computers, researchers and developers have adopted the technology of artificial intelligence to solve urgent applied problems. Information security experts are seriously considering the possibility of replacing traditional statistical anomaly detectors with artificial neural networks. Among the main advantages of ANN are the following:

- no need to formalize knowledge (replaced by training);
- ability to learn automatically and in the process;
- probability of detecting unknown attacks;
- possibility of parallelization of work.

In works [14–17], it is proposed to use a multilayer perceptron, previously trained on the basis of attacks (e.g., KDD [18]). In the work of Zhang et al. [19], it is proposed to use a neural network and wavelet transform. Kang and Kang [20] use deep learning to detect problems in a vehicle’s onboard CAN network. Emilianova et al. [21] propose to use a recirculating neural network or the principal component method for compressing the feature space, after which they investigate the use of both a double-layer perceptron and a self-organizing map (SOM). In [22], existing approaches are investigated and indicate the possibility of using perceptrons with different numbers of layers, single-layer classifiers for detecting the normal state, as well as a hybrid network consisting of a SOM and a perceptron.

The number of GAN-related jobs has grown exponentially in the past two years. Now, more than 300 different approaches, algorithms, and improvements have been recorded. The scope of the GAN is very wide.

The first area is image processing, video, audio, and textual information. For example, in [23], the authors present TextGAN, a model capable of generating realistic texts. It can be noted that the long short-term memory (LSTM) network is used as a generator, and convolutional neural network (CNN) is used as a discriminator. In [24], the authors presented the AnoGAN model based on DCGAN, the purpose of which is to determine anomalies in computed tomography images. Such works open up broad prospects for image processing. You can unravel the style of written characters and the characters themselves, the background of the image and the object, hair, emotions, and the fact that there are points in the sample from the selfie, etc. In addition, serious topics such as the diagnosis of diseases and the search for drugs are addressed.

The second area is the detection of anomalies. For example, in [25], the authors proposed an efficient anomaly detection algorithm based on the GAN. This model is based on teaching methods that allow you to simultaneously teach both an encoder and a generator together with a discriminator. As an example, the authors cite the model bidirectional GAN (BiGAN) [26]. This approach helps to avoid the computationally expensive operation of restoring a hidden view.

The third area is the generation of data for training. The authors of [27] investigated another problem—the generation of plausible data to increase the training sample. As a result of training with the help of these samples, the random forest classifier was strengthened, having trained to find such domain families that it did not initially detect. The basis of the model of autoencoder allows you to detect domain names generated using domain generation algorithm (DGA).

Thus, an analysis of the literature has shown that management and monitoring tools are evolving toward complex solutions. Modern systems, in general, strive to not only perform the narrow task of intrusion detection, but also help in diagnosing network malfunctions, while implementing both anomaly detection methods and methods for detecting known violations. This increases the amount of data that you want to process, and their dimension. The literature mainly deals with neural network methods based on perceptrons or SOMs. Recently, however, the attention of researchers and developers has focused on the use of competitive neural networks or convolutional networks. These methods are well proven in related areas where complex analysis is required.

At the moment, there is a clear gap between the needs of the industrial Internet of things (IIoT) market and the offer in the form of existing solutions [28, 29]. Therefore, it makes sense to conduct research in this direction.

## 4 Architecture of GAN

GAN is an architecture consisting of a generator and a discriminator that are configured to work against each other. Discrimination algorithms attempt to classify input data. Given the characteristics of the data, they try to determine the category to

which they relate. Generative algorithms are reversed. Instead of predicting a category according to the available images, they try to match images to this category. That is, in other words, discriminatory models study the boundary between classes, and generative models model the distribution of individual classes.

An open neural network library, Keras [30], was used to implement the neural network. It contains numerous implementations of widely used building blocks of neural networks, such as layers, target and transfer functions, and optimizers.

Consider the discriminator model. It uses a sequential Keras model with dense layers (fully connected layers). The number of layers and neurons in the layers is set when creating a class of the neural network model.

The discriminator model can be represented as two blocks. The first block consists of a set of layers, each of which uses a pre-configured matrix of weights and the activation function ReLu. The `kernel_initializer = 'uniform'` parameter implies the use of an initializer that generates uniformly distributed tensors. Within the framework of the problem being solved, this initializer allows to achieve a significant increase in the accuracy of the neural network. The activation function was chosen as the ReLu function—one of the simple activation functions that looks like  $f(x) = \max(0, x)$ , since it efficiently reduced the error compared to the sigmoid and Tanh functions [31]. The second block contains one layer, which contains the number of neurons equal to the number of investigated signs and using the sigmoidal activation function.

Instead of the classical procedure of stochastic gradient descent, Adam optimizer [32] is used to update iterative weights in the network based on training data. It is well suited for tasks that are large in terms of data and parameters.

The generator model is designed in a similar way. It also uses the activation functions ReLu, uniform initializer, and Adam optimizer.

GAN training should take place in a special way. While the discriminator is training, the generator values are held constant. Similarly, during generator training, the discriminator values should remain constant. Everyone must train against a static opponent. This will allow the generator to better read the gradient by which it should learn, and the discriminator will be able to establish this gradient. If the discriminator is too well trained, it will return values very close to 0 or to 1, so the generator will have difficulty reading the gradient. If the generator is too good, then it will constantly use the discriminator's flaws leading to incorrect negatives.

The trained neural network needed to be saved with the possibility of further loading. Keras makes it quite easy to save a trained neural network and load it again. For this, the model used is converted to JSON format, and all weights are written in.h5 format. After loading, the model must be compiled.

## 5 Training and Testing Algorithms

For training and testing the developed ANN, it is necessary to prepare datasets. Existing datasets, such as KDD and DARPA, do not take into account the specifics of a dynamic industrial system, so it was decided to synthesize the data using the

network simulator NS-3 [33], which allows you to create a digital twin of the network infrastructure of the enterprise. In order for the developed intrusion detection model to be fully and adequately run the program repeatedly, each time in a new experiment the location of the nodes, their speed, and transmitter power changed. In total, 500,000 launches of 4 scenarios were carried out (without attack, with a black hole attack, with a gray hole attack, and with a DDoS attack). Each attack has its own characteristics, which the neural network takes into account.

The training algorithm is as follows:

1. From the routing tables, we obtain information on the number of hops between nodes. From the statistics files (flowmon) is allocated information on the transfer of packets and their loss. Packets are extracted from the script file for NetAnim, and the received, transmitted, and dropped packets are calculated for each node. According to all these data, statistics is considered and a feature vector is prepared.
2. The information from point 1 is combined into one marked sample: Matrix  $X$  is a set of feature vectors for each analyzed traffic.
3. A neural network is built according to the model of generative-competitive neural networks. For this purpose, 3 neural networks are compiled. The first is a generator consisting of three layers: the first of 5 neurons, the input dimension is 10, the second of 3, the third, the output, of 70 (the size of the feature vector). The activation function for 1, 2 layers is ReLu, and for the output—sigmoid. The second network is a discriminator consisting of three layers: 5, 3, and 1 neurons, respectively. The input dimension is 70. The activation functions are ReLu, ReLu, and sigmoid, respectively. The third network (stacked) is the combination of the first and second network into one, where discriminator training is disabled.
4. Neural network must be compiled.
5. The formed samples are fed to the input of the neural network for learning. The neural network trains 500 epochs. At each epoch, a random part of the input data consisting of 1000 vectors is taken. Thousand random vectors of dimension 10 are generated. The latter are transmitted to the generator input to predict fake attributes. Fake and real data are concatenated into one table. All fake data are marked as non-malicious traffic. The resulting table is fed to the input of the discriminator learning. The discriminator takes each vector, drives the calculations through the neurons, counts the differentials for each layer in a reverse pass, and recalculates the weights for the neurons. Then, a generator is trained on a set of random vectors that are marked as malicious. This is done by training the integrated neural network. The generator is trained according to the discriminator principle, and the discriminator is not trained at this moment.
6. The neural network is saved to a file for future reuse.
7. The neural network is ready for operation; that is, you can proceed to testing it.

The testing algorithm is as follows:

1. The neural network is loaded from the file.
2. At the input of the neural network, data are supplied.

3. For marking, dataset used trained discriminator.
4. Calculate the metrics to assess the quality of the neural network.

## 6 Quality Assessment of Neural Network

Quality assessment of the developed neural network was carried out using the sklearn module metrics library. To evaluate the effectiveness of the developed solution, several metrics were used: accuracy, precision, recall,  $F$ -measure. Precision–recall curve, ROC curve, and a confusion matrix were also constructed.

In the simplest case, such a metric can be the share of scenarios for which the neural network made the right decision.

$$\text{Accuracy} = \frac{P}{N} \quad (12)$$

According to the results of the experiments, the maximum accuracy of detecting cyber threats amounted to 95%.

Precision and recall are metrics that are used in evaluating most of the information extraction algorithms. Sometimes, they are used on their own, sometimes as a basis for derived metrics such as  $F$ -measure or  $R$ -precision. The essence of precision and recall is very simple.

The precision of a system within a class is the proportion of scenarios that actually belong to this class relative to all the scenarios that the system has assigned to this class.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (13)$$

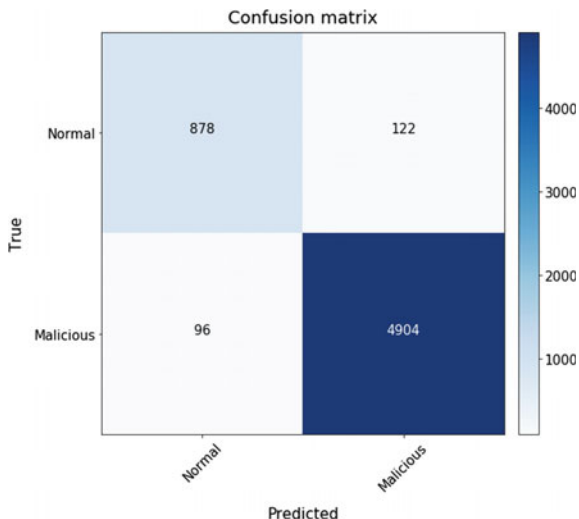
The recall of a system is the proportion of scenarios found by the classifier that belong to the class relative to all scenarios of this class in the test set.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (14)$$

Obviously, the higher the precision and recall, the better. However, in practice, maximum precision and recall are not achievable at the same time and you have to look for a balance. Therefore, it is advisable to have a certain metric that combines information about the precision and recall of our algorithm. Such a metric is the  $F$ -measure.  $F$ -measure is the harmonic average between precision and recall. It tends to zero if precision or recall tends to zero.

An important role in the description of these metrics in terms of classification errors is played by the error matrix. The classification error matrix for the developed neural network will be as follows (Fig. 2).

**Fig. 2** Confusion matrix



**Table 1** Consistency table

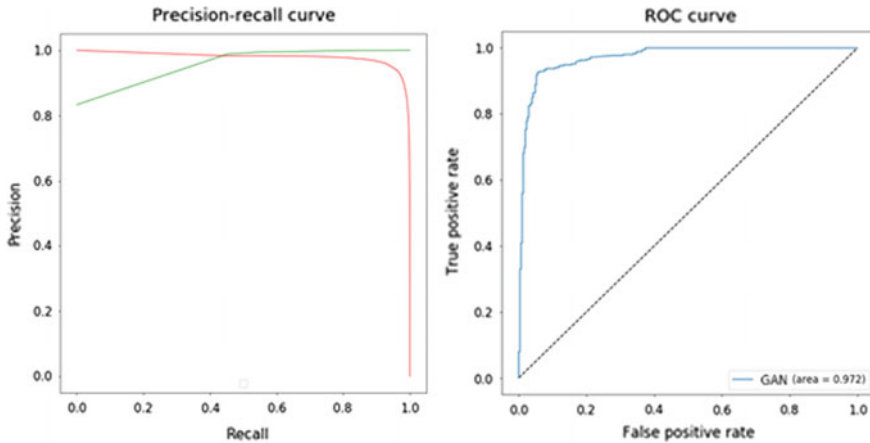
	Precision	Recall	<i>F</i> -measure	Support
Normal	0.90	0.88	0.89	1000
Malicious	0.98	0.98	0.98	5000
Micro-avg	0.96	0.96	0.96	6000
Macro-avg	0.94	0.93	0.93	6000
Weighted-avg	0.96	0.96	0.96	6000

The values of the previously described metrics for developed neural network are presented in Table 1.

Figure 3 presents precision–recall curve and ROC curve.

A method for comparing ROC curves is to estimate the area under the curves. Since the model is always characterized by a curve located above the positive diagonal, they speak of changes from 0.5 (“useless” classifier) to 1.0 (“ideal” model). This estimate can be obtained directly by calculating the area under the polyhedron bounded by the coordinate axes on the right and bottom and on the top left by the experimentally obtained points. The numerical indicator of the area under the curve is called area under curve (AUC), and in this case it is 0.972, which indicates the high effectiveness of the developed model for detecting cyber threats.

As a result of the study, ANN was developed to detect and prevent cyber threats in real time and limited resources on the industrial Internet of things. According to the obtained results, we can conclude that the developed ANN for the initial problem gives a stable result equal to 95%. The probabilities of FP and FN do not exceed 5%. When trying to increase the input data, the generator reduced the accuracy of the discriminator, so it was not possible to improve the result. It was also noticed that



**Fig. 3** Precision–recall curve and ROC curve

with an increase in network traffic, the processing time of the signs that the ANN receives at the input increases by several times. Training ANN with a large amount of network traffic is much faster than parsing input data. Considering the fact that a large amount of data was generated for training, the network retraining should be carried out quite rarely in the future.

## 7 Conclusion

As a result of the study, various modifications of generative adversarial networks were considered. Areas in which they are already successfully applied are identified. Algorithms for training and testing the developed neural network for detecting cyber attacks in IIoT are proposed. The best neural network configuration for solving the task has been determined. A method of synthetic dataset generation for training and testing has been developed. The experimental results confirmed the effectiveness of the proposed method based on GAN.

The study showed that using neural networks in the task of detecting attacks in dynamic networks of IIoT is a better strategy than classical signature-based analysis. Firstly, neural networks are a more flexible solution, since by increasing the input vector with new features, this is a simple task and does not require rewriting the network code, and any adaptation to new features occurs automatically. Secondly, the neural network is not inferior in performance to complex signature methods, which, having a huge database of signatures, are capable of heavily loading resources.

In the future, it is planned to expand the number of detected cyber threats, supplement the signs of attacks, and conduct a larger study using the computing resources of the supercomputer.



**Acknowledgements** The work was funded by the Russian Federation Presidential grants for support of young scientists and postgraduate students (SP-443.2019.5).

## References

1. World Economic Forum: The Fourth Industrial Revolution Davos 2016. <https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-itmeans-and-how-to-respond>. Last accessed 2019/09/05
2. Coppolino, L., D'Antonio, S., Formicola, V., Romano, L.: Enhancing SIEM technology to protect critical infrastructures. In: Critical Information Infrastructures Security 7th International Workshop, CRITIS 2012, pp. 10–21, Norway (2012)
3. Baltimore's 911 Emergency System Hit by Cyberattack. <https://www.nbcnews.com/news/us-news/baltimore-s-911-emergency-system-hitcyberattack-n860876>. Last accessed 2019/09/05
4. Goncharov, E.: Challenges of Industrial Cybersecurity. <https://ics-cert.kaspersky.com/reports/2019/01/17/challenges-of-industrial-cybersecurity/>. Last accessed 2019/09/05
5. Berman, D., Buczak, A., Chavis, J., Corbett, C.: A survey of deep learning methods for cyber security. *Information* **10**, 122 (2019)
6. Igor, H., Bohuslava, J., Martin, J., Martin, N.: Application of neural networks in computer security. *Procedia Eng.* **69**, 1209–1215 (2013)
7. Swarup, K.: Artificial neural network using pattern recognition for security assessment and analysis. *Neurocomputing* **71**(4–6), 983–998 (2008)
8. Deng, L., et al.: Recent advances in deep learning for speech research at Microsoft. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 8604–8608, Vancouver (2013)
9. With QuickType: Apple wants to do more than guess your next text. It wants to give you an AI. <https://www.wired.com/2016/06/apple-bringing-ai-revolution-iphone/>. Last accessed 2019/09/05
10. Xiong, W., Wu, L., Allewa, F., Droppo, J., Huang, X., Stolcke, A.: The Microsoft 2017 Conversational Speech Recognition System [Technical Report]. <https://www.microsoft.com/en-us/research/publication/microsoft-2017-conversational-speech-recognition-system/>. Last accessed 2019/09/05
11. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014)
12. Hui, J.: Why it is so hard to train generative adversarial networks! Medium. *Data Sci.* (2018)
13. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. *Int. J. Eng. Trends Technol. (IJETT)* (2015)
14. Mustafaev, A.: Neirosetevaya sistema obnaruzheniya kompyuternyh atak na osnove analiza setevogo trafika. *Voprosy bezopasnosti*. 2016. № 2, pp. 1–7 (2016)
15. Halenar, I., et al.: Application of neural networks in computer security (2013)
16. Govindarajan, M., Chandrasekaran, R.: A hybrid multilayer perceptron neural network for direct marketing (2014)
17. Gallant, S.: Perceptron-based learning algorithms. *IEEE Trans. Neural Netw.* **1**(2), 179–191 (1990)
18. KDD Cup 1999: <https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>. Last accessed 2019/09/05
19. Zhang, H., Huang, Q., Fangwei, L., Jiang, Z.: A network security situation prediction model based on wavelet neural network with optimized parameters. *Digital Commun. Netw.* **2**(3), 139–144 (2016)
20. Kang, M., Kang, J.: Intrusion detection system using deep neural network for in-vehicle network security. *PLoS ONE* **11**(6), e0155781 (2016). <https://doi.org/10.1371/journal.pone.0155781>

21. Emilianova, U., Talalaev, A., et al.: Neyrosetevaya tehnologiya obnaruzheniya setevykh atak na informacionnye resursy. Programmnye sistemy: teoriya i prilozheniya **3**(7), 3–15 (2011)
22. Kornev, P., Pylkin, A., Sviridov, A.: Using artificial intelligence in intrusion detection systems (2015)
23. Zhang, Y., Gan, Z., Fan, K., Chen, Z., Henao, R., Shen, D., Carin, L.: Adversarial feature matching for text generation. arXiv preprint [arXiv:1706.03850](https://arxiv.org/abs/1706.03850) (2017)
24. Schlegl, T., Seeböck, P., Waldstein, S., Schmidt-Erfurth, U., Langs, G.: Unsupervised anomaly detection with generative adversarial networks to guide marker discovery. In: International Conference on Information Processing in Medical Imaging, pp. 146–157. Springer, Cham (2017)
25. Zenati, H., Foo, C., Lecouat, B., Manek, G., Chandrasekhar, V.: Efficient GAN-based anomaly detection. arXiv preprint [arXiv:1802.06222](https://arxiv.org/abs/1802.06222) (2018)
26. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. arXiv preprint [arXiv:1605.09782](https://arxiv.org/abs/1605.09782) (2016)
27. Zheng, Y., Zhou, X., Sheng, W., Xue, Y., Chen, S.: Generative adversarial network based telecom fraud detection at the receiving bank. Neural Netw. (2018)
28. Pavlenko, E., Zegzhda, D.: Sustainability of cyber-physical systems in the context of targeted destructive influences. In: 2018 IEEE Industrial Cyber-Physical Systems (ICPS), St. Petersburg, pp. 830–834 (2018)
29. Lavrova, D., Poltavtseva, M., Shtyrkina, A., Zegzhda, P.: Detection of cyber threats to network infrastructure of digital production based on the methods of Big Data and multifractal analysis of traffic. In: SHS Web of Conferences, vol. 44, p. 00007, CC-TESOC2018WoS (2018)
30. Keras: <https://keras.io/>. Last accessed 2019/09/05
31. Shulga, D.: Exploring Activation Functions for Neural Networks (2017)
32. Brownlee, J.: Gentle Introduction to the Adam Optimization Algorithm for Deep Learning (2017)
33. Network Simulator NS-3: <https://www.nsnam.org/>. Last accessed 2019/09/05

# Detection and Prediction of Safety Faults in Inter-Device Networks Applying a Set of Data-Driven Methods



Maxim Kalinin , Vasily Krundyshev , Viacheslav Belenko, and Valery Chernenko

**Abstract** Digital transformation concerns the safety and reliability for smart homes, industrial IoT, smart building, VANET, WSN, and mesh networks. For common network environment, data confidentiality, integrity, and availability were treated as safety of information, but last decade, due to appearance of dynamic device-to-device cyber spaces, the cyber security is focused on safety, reliability, and sustainability of the connected cyber physical units. Growing variability and amount of controlled data make traditional analytical methods ineffective for safety ensuring. The paper presents our method of cyber faults detection and prediction in smart inter-device infrastructure by data-driven analysis. The proposed approach is founded on the modified method of  $k$ -nearest neighbors ( $k$ NN) improved with Dempster–Shafer (DS) theory and spatial-temporal correlation of the connected devices. This method shows above 99% level of effectiveness comparing to common safety assurance methods.

**Keywords** Dempster–Shafer · Safety · Data-driven analysis · Fault · Spatial-temporal correlation · IoT ·  $k$ NN

## 1 Introduction

Last decade, due to the active development of dynamic device-to-device networks (e.g., IoT [1], IIoT [2], WSN [3], MANET [4], VANET [5], as well as a popular concept of smart homes, smart buildings, and smart cities), the object of protection acquires a new glance as an element of the cyber space, in which traditional read and write operations have real-life physical consequences. Now, Internet of things (IoT) is a core of the concept of smart buildings and smart cities in which we live.

---

M. Kalinin (✉) · V. Krundyshev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [max@ibks.spbstu.ru](mailto:max@ibks.spbstu.ru)

V. Belenko · V. Chernenko  
LG-Polytechnic, Branch Office of LG Electronics Inc. (Korea), St. Petersburg, Russia

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021  
N. Voinov et al. (eds.), *Proceedings of International Scientific Conference on Telecommunications, Computing and Control*, Smart Innovation, Systems and Technologies 220, [https://doi.org/10.1007/978-981-33-6632-9\\_2](https://doi.org/10.1007/978-981-33-6632-9_2)

Technology forecasts suggest that about 30.7 billion IoT devices will be installed and connected to Internet by 2020. Most of these devices will be deployed in critical networked infrastructures of cyber physical space [6].

Rob van Kranenburg has introduced four levels of IoT by its coverage [7]: Body area network (BAN), local area network (LAN), wide area network (WAN), very wide area network (VWAN). On each level, IoT is a cyberspace equipped with integrated technological systems. Traditionally, these systems were installed separately. Due to the interoperability and interdependence of data between these systems, the concept of fully integrated smart network has emerged [8]. This infrastructure is aimed at improving energy efficiency and reducing operating costs [9].

While cyber physical infrastructure integrates preferences and actions to provide new services, there is also a growing risk of cyber attacks targeting both cyber subsystem of IoT and all outer systems connected to it. Cybercriminals who were previously concentrated on corporate networks and online services are now increasingly targeted at industrial control systems (ICS) and IoT-based areas [10]. Critically important infrastructures (e.g., smart hospitals, automated power plants, sensors field, smart farms) can now be their targets. It can lead to large economic losses or increase the risk of emergency. Below there are some kinds of threats that can be detected in inter-device networks and possible consequences of their activity:

- hardware failure: abort of service and system malfunction;
- power outages: system disruption;
- impact of viruses and trojans: malfunctions in the system software, disruption of work or disabling of the system;
- user mistakes: possible system failures due to improper use of equipment;
- interception of information transmitted over communications: violation of the confidentiality of information, it is possible to seize control of the system;
- presence of inner intruders (security staff, service men, cleaners, etc.): system malfunctions due to improper maintenance of equipment, the level of danger depends on the degree of insider access to the system.

Thus, a new mechanism is needed that would allow timely monitoring of failures and predicting their possible occurrence in the safety trace. Today, there are many methods for detecting safety failures and malfunctions in cyber physical environment of connected devices, e.g. [11–13]. However, it is important not only to timely detect the safety failures with high level of accuracy, but also to predict the further occurrence of faults. The goal of this paper is to develop a highly effective method for analysis and predicting of safety failures in inter-device network using a complex of new machine learning methods of data-driven analysis. To present our work, the paper is structured in the following manner: Sect. 2 lists the known methods for detecting and predicting failures, Sect. 3 describes our failure analysis solution, Sect. 4 sets our experimental results, and, finally, Sect. 5 gives a conclusion to our work.

## 2 The Related Works

A survey of the known methods for fault management allows us to erect the following background:

1. *k*-nearest neighbors (*k*NN) method [14]:
  - mathematical structure at the heart of the method—calculate distance function;
  - fault detection rate—81%;
  - used to detect malfunctions of an oil-filled power transformer.
2. modified *k*NN (*k*NN + DS method) [15]:
  - calculating the distance function with application of Dempster–Shafer (DS) theory;
  - fault detection rate—95%;
  - used to detect malfunctions of an oil-filled power transformer.
3. support vector machine (SVM) + *k*NN [16]:
  - construction of a hyperplane for a set of points, finding the distance function;
  - fault detection rate—75.3%;
  - utilized to detect failures in various machines. For detection, the SVM method is used, and for its improvement, the KNN method is utilized.
4. artificial neural networks (ANN) [17, 18].
  - Mahalanobis distance calculation.
  - ability to predict the deterioration of the details considered in the article environment.
5. SVM + back propagation neural network (BPNN) [19].
  - search for a hyperplane for a set of points;
  - fault detection rate: SVM—91.93%, and BPNN—85.57%;
  - detection of failures at a thermal power plant. SVM and BPNN are used for classification.
6. fuzzy neural network (FNN) [20].
  - counting fuzzy Gaussian function of membership;
  - fault detection rate—93.37%;
  - applied in wastewater treatment systems. Based on the predicted sensor validity index (SVI) values and sensor outputs. To predict SVI values, FNN is entered with multiple inputs and one output (MISO). On the basis of various water quality sensors, water quality is assessed and, based on this, a possible failure in the treatment system is predicted.

The listed methods are quite diverse, however, none of them have a sufficiently high accuracy of fault detection, except *k*NN + DS method. The further section

presents our proposal to increase accuracy of  $k\text{NN} + \text{DS}$  method. Our method is based on the data-driven analysis received from the devices of the digital infrastructure and further building the Spatial-Correlation Consistency Regions (SCCR) for the neighbor devices. We have called our method as  $k\text{NN} + \text{DS} + \text{SCCR}$ .

### 3 Safety Failure Analysis and Prediction

#### 3.1 Data-Driven Failure Analysis

The proposed method for data-driven failure analysis consists of a module for calculating regions of data consistency and modules for predicting the failures (Fig. 1).

The proposed system contains the following units:

- module for calculating the regions of consistency;
- module for consistency checking;
- databases with regions of consistency;
- fault detection module;
- failure prediction module.

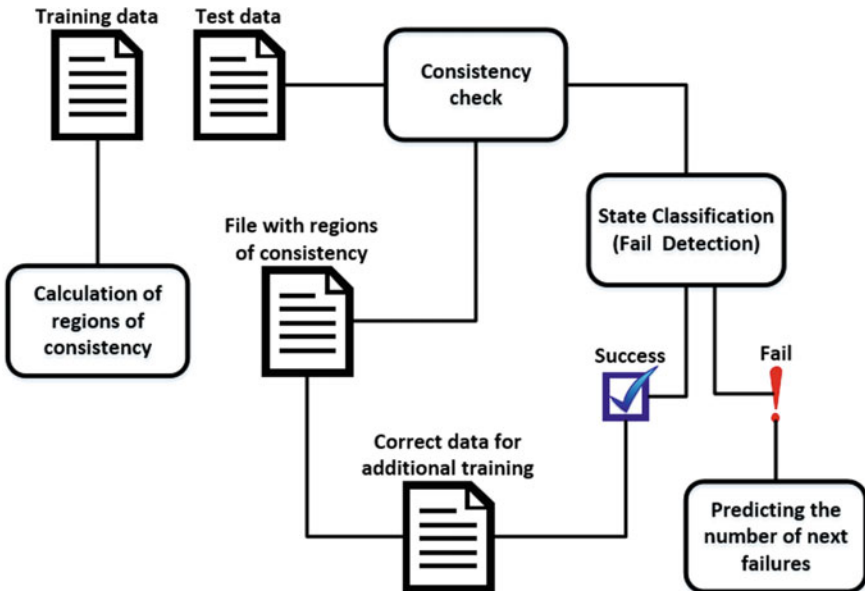


Fig. 1 Scheme of the data-driven failure analysis

The fault detection module is based on the modified  $k$ NN method and performs the classification of the sensor indicators coming from the devices. The result is a list of sensor states, based on which further prediction of failures is made. The fault prediction module is based on a method that uses information about the past number of failures. The result of this method is a number of failures that will occur for a particular pair of devices. Failure detection and prediction consists of the following steps:

1. The module for calculating the regions of consistency receives data from  $N$  sensors from each device. The data is a sequence of sensor readings for a certain period of time  $T$ .
2. Based on the training data  $S_i(t, T)$  and  $S_j(t, T)$  taken up to a predetermined time  $T$ , for each pair of counters that are neighbors, a region of consistency of spatial correlations is calculated.
3. The regions of consistency are constructed for each device, and these data are stored in a database.
4. To detect failures, the  $k$ NN method is extended with the DS theory. Before applying this method, the incoming data is also split in pairs for each pair of devices and checked for consistency using saved regions of consistency. If the data is consistent, they are further classified, otherwise the data is not processed.
5. The result of the classification is a sequence of assessments of the state of each pair of devices (“Normal\_state” or “Fault\_state”).
6. Next, prediction of failures occurs based on the number of “Fault\_state” ratings for each pair of devices using the method based on the Weibull distribution.

### 3.2 Module for Calculating the Regions of Consistency

In this module, correlation patterns are built for each pair of devices located in one correlation area, which is determined manually. There are several devices in the correlation area, while devices may belong to several different areas. The area of correlation, the Spatial-Correlation Consistency Regions (SCCR), ensures that the devices in it are spatially and temporally correlated, which is required to solve the task.

The formula for the center of the SCCR ellipse is (1):

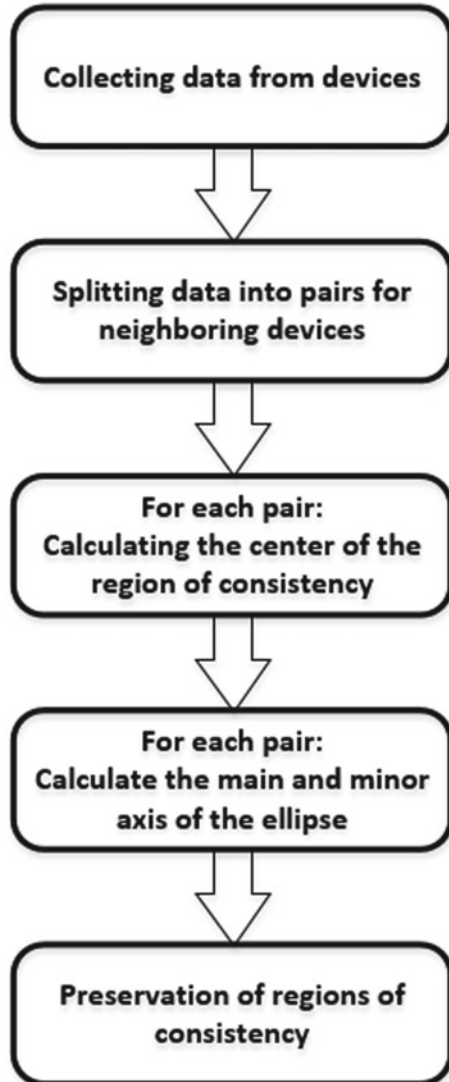
$$EWMA = a * p_t + (1 - a) * EWMA_{t-1}, \quad (1)$$

where EWMA is the value of the exponentially weighted moving average at the point  $t$  (the last value, in the case of a time series);  $EWMA_{t-1}$  is the value of the exponentially weighted moving average at the point  $t - 1$  (the previous value in the case of a time series);  $p_t$ —the value of the original function at time  $t$  (the last value, in the case of a time series);  $a$ —the coefficient characterizing the rate of decrease in weights, takes a value from 0 to 1, the smaller its value, the greater the influence of the previous values on the current average value.

Main and minor axes of the SCCR ellipse are calculated using principal component analysis (PCA). The orthogonal principal components  $\vec{a}$  and  $b$  determine the angle of rotation of the main axis to the axis  $\theta$ , and the lengths of these axes are set as three deviations  $3\sigma_a$  and  $3\sigma_b$ . In general, the multidimensional case, the process of isolating the main components occurs as follows (Fig. 2):

1. The center of the data cloud is searched, and a new origin is transferred there—this is the zero principal component (PC0).

Fig. 2 Algorithm for SCCR calculation





2. The direction of the maximum data change is selected—this is the first main component (PC1).
3. If data is not fully described (noise is large), then another direction (PC2) is chosen—a perpendicular to the first one, so as to describe the remaining change in data, etc.

### 3.3 Fault Detection Module

To classify the state of the device, an improved “k” method of nearest neighbors is used applying the Dempster–Shafer theory.

BPA is a basic probability assignment. For one neighbor it is calculated as  $m_{s,i}(\{C_q\}) = \alpha_0 \phi_q(d^{s,i})$ , where  $\alpha_0 = 0.95$ ,  $d^{s,i}$  is a distance between the state being classified and the nearest neighbor. The value of  $\phi_q$  is calculated as  $\phi_q(d^{s,i}) = e^{-\gamma_q d^\beta}$ , where  $\beta = 2$ ,  $\gamma_q = 1/d_q^\beta$ , and  $d_q$  is the average distance between the closest distances to the class  $C_q$ .

The total BPA for the set of all neighbors of the class  $C_q$  is calculated as  $m_q^s(\{C_q\}) = 1 - \prod_{x_i \in \phi_q^s} (1 - \alpha_0 \phi_q(d^{s,i}))$ .

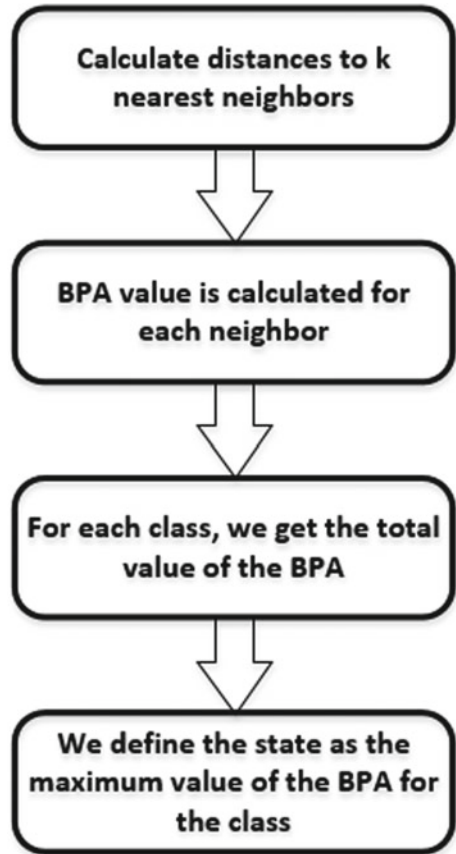
Figure 3 presents fault detection algorithm. Figure 4 shows a sample of the SCCR ellipse for two devices (the axes of coordinates are the values of sensor readings (voltage, temperature, etc.) at a certain time; sensors’ data are marked with dark green dots).

### 3.4 Failure Prediction Module

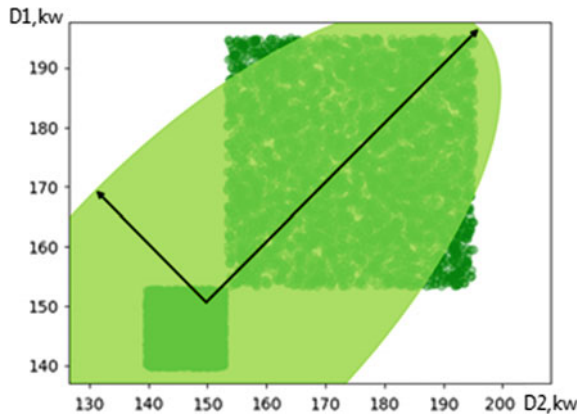
To predict failures, the method described in [21] was used. This method is based on the use of the Weibull distribution to assess the reliability of various systems. Suppose that a set of classified states of a device of size  $N$  that was received in the period from 0 to  $t_c$  contains  $X$  failures. You need to know how many failures will be in the time interval  $[t_c, t_w]$ .

It is desirable to have one prediction  $\hat{Y}$  for the number of failures  $Y$  in the future interval  $[t_c, t_w]$ . Considering the observed (non-zero) number of failures  $X$  for the period  $t_c$ , the predicted number of failures is calculated as  $\hat{Y} = N * \hat{q}$ , where  $\hat{q}$  is calculated as  $\hat{q} = [1 - (\frac{X}{N})] - [1 - (\frac{X}{N})]^{(\frac{t_w}{t_c})^\beta}$ , where  $t_w$ —the upper limit of the time interval,  $t_c$ —the lower limit of the time interval,  $\beta$ —form parameter (depending on the failure rate).

**Fig. 3** Fault detection algorithm



**Fig. 4** An example of SCCR



### 4 Experimental Results

For the experiment, three different situations of occurrence of failures in the cyber physical environment are selected [15]: random failures, periodic failures, noise. A dataset of 5000 values of indicators of various sensors for training and a test dataset of 1000 values were used. The experiment was conducted for 20 devices. Figures 5 and 6 present the outputs of the suggested method for 20 devices for random and periodic failures.

The figures show the average error of the developed method for the number of neighbors (1...40). The traditional  $k$ NN method copes with the task of classification worst of all (with a maximum number of neighbors  $k = 40$ , its accuracy is around 78%). The results of the  $k$ NN + DS method are better. The accuracy of 99.997% is reached by the suggested  $k$ NN + DS + SCCR method.

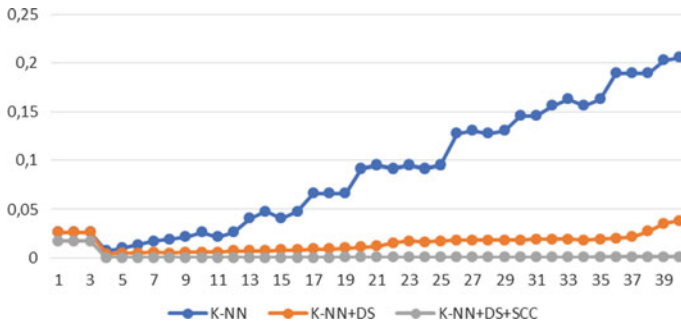


Fig. 5 Classification accuracy of random failures

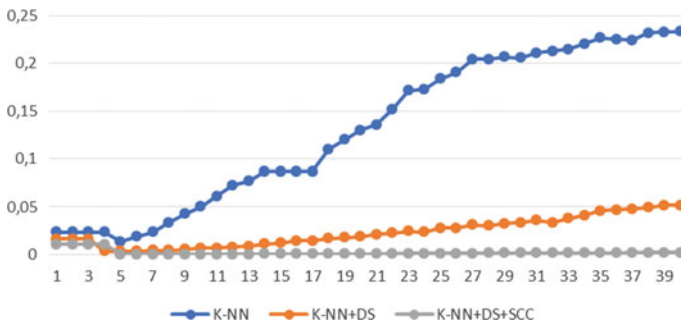


Fig. 6 Classification accuracy of periodic failures

## 5 Conclusion

The potential of IoT, its applications, and prospects is growing rapidly. Nevertheless, there are several open issues for IoT security and safety research varying from protection against cyber attacks [2, 4, 6, 22] to failures detection [23, 24]. Many important topics in the field of protection against cyber threats and disruptions are currently at the stage of intensive discussion.

The proposed data-driven method assembles the abilities of traditional  $k$ NN approach [14], Dempster–Shafer calculations [15] and data consistency checking. The obtained results have shown a high failure detection accuracy, above 99%.

Our further work is targeted for resources estimation and implementation of the suggested solution for the area of safety management systems.

**Acknowledgements** Great appreciation to the Branch office of LG Electronics Inc. (Korea) in St. Petersburg (Russia) for assistance and supporting this work in synthetic modeling of the smart building IoT environment.

**Funding** The research is funded by LG Electronics Inc.




## References

1. Lavrova, D., Pechenkin, A., Gluhov, V.: Applying correlation analysis methods to control flow violation detection in the internet of things. *Autom. Control Comput. Sci.* **49**(8), 735–740 (2015)
2. Sisinni, E., et al.: Industrial internet of things: challenges, opportunities, and directions. *IEEE Trans. Ind. Inf.* **14**(11), 4724–4734 (2018)
3. Parker, D., Stojanovic, M., Yu, C.: Exploiting temporal and spatial correlation in wireless sensor networks. In: *Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, pp. 442–446 (2013)
4. Singh, R., Nand, P.: Literature review of routing attacks in MANET. In: *International Conference on Computing, Communication and Automation (ICCCA)*, Noida, pp. 525–530 (2016)
5. Belenko, V., Chernenko, V., Kalinin, M., Krundyshev, V.: Evaluation of GAN applicability for intrusion detection in self-organizing networks of cyber physical systems. In: *International Russian Automation Conference (RusAutoCon)*, Sochi, pp. 1–7 (2018)
6. Protecting Smart Buildings from Cyber Attacks: <https://medium.com/iot-security-institute/protecting-smart-buildings-from-cyber-attacks-b6a1ad2f4cd/>. Last accessed 2019/09/05
7. IoT Interview Series: 5 Questions with Rob van Kranenburg of the Internet of Things Council. <https://www.postscapes.com/iot-voices/interviews/iot-interview-series-5-questions-rob-van-kranenburg-internet-things-council/>. Last accessed 2019/09/05
8. Chen, H., Chou, P., Duri, S., Lei, H., Reason, J.: The design and implementation of a smart building control system. In: *IEEE International Conference on e-Business Engineering*, Macau, pp. 255–262 (2009)
9. Crooks, A., Schechtner, K., Dey, A., Hudson-Smith, A.: Creating smart buildings and cities. *IEEE Pervasive Comput.* **16**(2), 23–25 (2017)
10. Belenko, V., Krundyshev, V., Kalinin, M.: Synthetic datasets generation for intrusion detection in VANET. In: *Proceedings of the 11th International Conference on Security of Information and Networks* (2018)

11. Kim, W., Katipamula, S.: A review of fault detection and diagnostics methods for building systems. *Sci. Technol. Built Environ.* (2017). <https://doi.org/10.1080/23744731.2017.1318008>
12. Luo, R., Su, K., Tsai, K.: Fire detection and isolation for intelligent building system using adaptive sensory fusion method. In: *IEEE International Conference on Robotics and Automation* (Cat. No.02CH37292), Washington, DC, USA, vol. 2, pp. 1777–1781 (2002)
13. Kim, Y., Sharifi, R., Cha, Y., Langari, R.: *Sensor fault diagnosis of smart buildings* (2010)
14. Yu, F., Liu, J., Liu, D.: An approach for fault diagnosis based on an improved k-nearest neighbor algorithm. In: *35th Chinese Control Conference (CCC)*, Chengdu, pp. 6521–6525 (2016)
15. Denoeux, T.: A k-nearest neighbor classification rule based on Dempster-Shafer theory. *IEEE Trans. Syst. Man Cybern.* **25**(5), 804–813 (1995)
16. Andre, A., Beltrame, E., Wainer, J.: A combination of support vector machine and k-nearest neighbors for machine fault detection (2013). <https://doi.org/10.1080/08839514.2013.747370>
17. Bangalore, P., Tjernberg, L.: An artificial neural network approach for early fault detection of gearbox bearings. *IEEE Trans. Smart Grid* **6**(2), 980–987 (2015)
18. Rajakarunakaran, S., et al.: Artificial neural network approach for fault detection in rotary system. *Appl. Soft Comput.* **8**, 740–748 (2008)
19. Chen, K., Chen, L., Chen, M., Lee, C.: Using SVM based method for equipment fault detection in a thermal power plant. *Comput. Ind.* **62**, 42–50 (2011)
20. Han, H., Li, Y., Qiao, J.: A fuzzy neural network approach for online fault detection in waste water treatment process. *Comput. Electr. Eng.* **40**, 2216–2226 (2014)
21. Nordman, D., Meeker, W.: Weibull prediction intervals for a future number of failures. *Technometrics* **44** (2000). <https://doi.org/10.1198/004017002753398191>
22. Lavrova, D.: An approach to developing the SIEM system for the internet of things. *Autom. Control Comput. Sci.* **50**(8), 673–681 (2016)
23. Lavrova, D., Poltavtseva, M., Shtyrkina, A., Zegzhda, P.: Detection of cyber threats to network infrastructure of digital production based on the methods of big data and multifractal analysis of traffic. In: *SHS Web of Conferences*, vol. 44, p. 00007, CC-TEESC2018WoS (2018)
24. Pavlenko, E., Zegzhda, D.: Sustainability of cyber-physical systems in the context of targeted destructive influences. In: *IEEE Industrial Cyber-Physical Systems (ICPS)*, pp. 830–834 (2018)

# Development of an Algorithm for Determining the Railway Tracks on Video Image



Ivan Deylid , Sergej Molodyakov , and Boris Tyutin 

**Abstract** The subject of the work is in the field of creating autopilot railway transport. One of the problems related to railway safety is considered. This paper presents a developed algorithm for determining the railway track from a video image. It is the first step in determining the free path or obstacles in front of the locomotive. The proposed algorithm is a combination of several machine learning algorithms that are applied sequentially (boosting). The first stage of the algorithm is the extraction and classification of features from the image. In this stage, the speeded up robust features or SURF-method is used. At the output of the SURF-stage, we obtain information in the form coordinates of key points and their descriptors. In the second stage, the selected key points are classified. Combinations of two classification methods are used: the  $K$ -nearest neighbors or KNN-method and the support vector machines or SVM-method. The final step is the compilation of a railway track mask. For this, the nearest neighbor graph method is used. For practical use of the found mask, the inverse perspective transformation is performed. The efficiency of the developed algorithm is shown experimentally. It can be considered as one of the ways of image segmentation. The main advantages of the algorithm are associated with minimal preparation of the training sample and the ability to analyze its work for further improvement. The results of processing real video images obtained from a video camera mounted on a locomotive are presented.

**Keywords** Computer vision · Railway track · Machine learning · Algorithm ·  $K$ -means · SURF · SVM

---

I. Deylid · S. Molodyakov (✉)  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [molodyakov\\_sa@spbstu.ru](mailto:molodyakov_sa@spbstu.ru)

B. Tyutin  
Opera Software, Gothenburg, Sweden

## 1 Introduction

Today, the application field of computer vision is continuously expanding. Hardware and software development follows the demand for better tools in this area all-new algorithms for image processing and recognition are being created. More powerful, high-speed multi-core video processors start to get used for computations. Better cameras, lidars, radars, and other types of sensors gather more information from the environment. Integration of all of this data allows creating systems capable of reliable object recognition. They are robust enough to operate in different weather conditions.

Transport is an area where the usage of computer vision gives great results. A lot of modern cars already include autopilot elements based on video processing. Commercial success of the Tesla autopilot [1] showcased the great potential of computer vision in modern industries. Right now active work is being carried out to introduce it in railway transport [2–5]. The usage of computer vision makes it possible to take the next step in the construction of autopilot trains. The use of autopilot trains increases traffic safety, expands our ability to monitor train telemetry and, in general, improves the automation of rail transportation.

Great research in this area was done at the University of Salzburg (Austria). There an experimental locomotive was equipped with sensors, and several computer vision algorithms were developed [6, 7]. There is also ongoing work in the Research and Design Institute of Informatization, Automation, and Communication in Railway Transport (JSC NIIAS). It focuses on automation of the railway transport movement, creation of autonomous trains, and usage of computer vision systems. Currently, one of the shunting locomotives is equipped with two cameras for near and far observation areas, a lidar and a radar. Data from cameras can be sent real-time to the NIIAS laboratory in St. Petersburg [8].

Figure 1 shows the original image frame obtained from the 2K resolution camera. One of the main tasks of image processing is to determine the railway track in the path of locomotive motion using a video camera. The main difficulties here come from the presence of railway switches, the real-time nature of all the computations, and the need to mitigate the influence of various weather phenomena (snow, rain, etc.).

Several algorithms can be used to determine the railway tracks. One of them is connected with using some of the methods based on segmented neural networks (SegNet, Unet, FCN, etc.) [9–11]. Main disadvantages of these methods are large computing costs per frame and difficulties in interpreting the results of a deep neural network. Thus, it is reasonable to use several algorithm together to reduce the probability of mistakes.

In this chapter, an algorithm for determining the railway track on the way of the motion using the analysis of the features will be considered. The work of the algorithm in the tasks of classification, removal of noise, and construction of the individual features of the railway track mask will be shown in stages. The represented algorithm can be considered as one of the algorithms of image segmentation.

**Fig. 1** The original image of the railway tracks



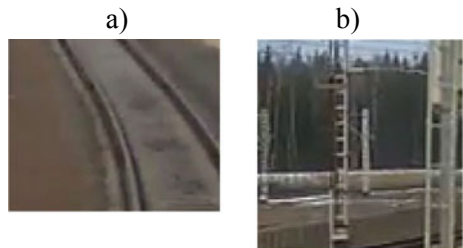
## 2 Algorithm

### 2.1 Frame Extraction

The first stage of the algorithm is the extraction and classification of features from the image. For this, the speeded up robust features (SURF) method is used, which is a fast method for extracting features from an image [12, 13]. At the output of the SURF-stage, we obtain information in the form coordinates of key points and their descriptors. This information is further used to classify points.

It was previously selected set of data consisting of the image of the railway track—a positive example, and background images—negative example (Fig. 2). During the training, features were extracted from each image, and their descriptors are marked with a class label (depending on which data set belongs to the image).

**Fig. 2** The examples of positive (a) and negative (b) images from the dataset





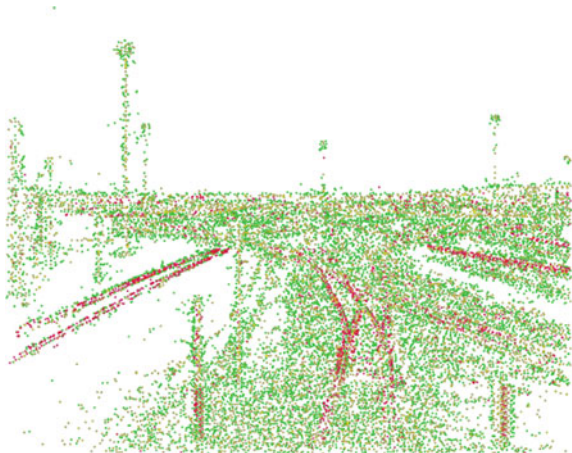
## 2.2 Key Point Classification

In the second stage, the selected key points are classified. A combination of two fairly simple and popular classification methods is used: the  $K$ -nearest neighbors (KNN) [14] and the support vector machines (SVM) [15, 16]. It is important to say about features classifier training: for SVM it is recommended to select the tuning parameters in such a way as to avoid overfitting, assigns, results in one set of data may, with a high probability be in another. For the  $K$ -nearest neighbor method, optimization is more complicated: after creating training samples, clustering is performed on it by the  $K$ -Means algorithm [17], after which the label of each centroid is determined by the overwhelming number of signs (positive or negative). Then new training samples for the algorithm are created, consisting of fewer examples.

Two main advantages of this approach are that the amount of training data for the nearest neighbor method (in this example 500 instead of 7000) is significantly reduced, which will increase the speed of the classifier, secondly, thus avoiding retraining, and thirdly, the selection task number of neighbors more easily now. All points on the screen are classified, in the case that both classifiers gave a positive result—this point is considered as a railway's point.

Figure 3 shows the result of the aggregation of classifiers based on the support vector method and the  $K$ -nearest neighbor method: red points are rails determined by two classifiers; yellow points defined as rails only one of the classifiers; green points are background.

**Fig. 3** The result of classification of key points



### 2.3 Noise Removing

No machine learning algorithm can always produce a correct result. So the next step in determining a railway is to remove noise or false-positive classifier results. It is possible to skip points defined as not related to rails, but being them (it is enough to know some points of the contours of rails and approximate the rest), we will ignore the classifier error data. Many more problems can happen with points falsely defined as relating to rails, because in the next stage when they participate in drawing up the mask, it is possible to get distortions. To remove them, we use the compactness property of the data. It is logical to assume that most of the rail track points will be near to each other because on the frame they also are near to each other, so the right way is to remove all those points, among which the nearest  $N$  neighbors are defined as not rails.

### 2.4 Masking

The final step is the compilation of a rail track mask. For this, the NNG (Nearest neighbor graph) method [18] is used. Let us compose the graph on the image, according to the following principle: connect each point with  $K$  closest to it. Thus, the contours of these figures cover image areas related to the rails. By varying the parameter  $K$ , one can avoid distortions and errors in the construction of the mask (by reducing the parameter  $K$ ), or reduce the probability of rupture of the track sections (by increasing the parameter  $K$ ). In Fig. 4 it is a graph composed of several neighbors equal to 6. To create a mask from the obtained graphs, we repeatedly do a morphological closure operation on the resulting graph (Fig. 5). In Fig. 6 you could see that the mask does not cover all the railway tracks.

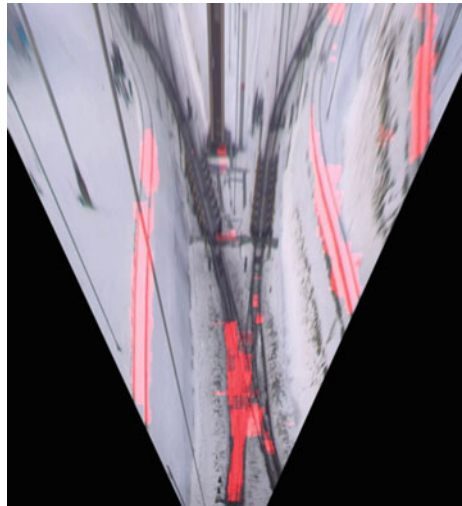
**Fig. 4** The graph of key points connections with several neighbors equal to 6



**Fig. 5** Image frame with a mask obtained after execution of the morphological closure



**Fig. 6** An image of railway track with a mask over successive 1 frame

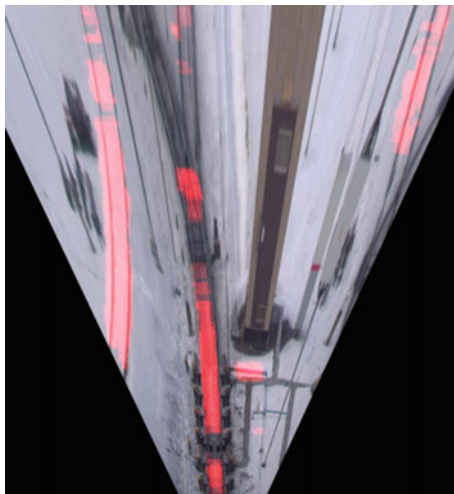


For getting more complete coverage of the railway track with a mask, it is suggested to accumulate data obtained from previous frames, averaging their masks. This allows more stable determination of the track.

For practical usage of the found path, it is necessary to compare the data about it with other different sensors and methods of semantic segmentation of the image. Therefore, an inverse perspective transformation (Bird View [19, 20] projection) is performed above the frame, which further allows the results obtained in the complex description of the surrounding locomotive environment.

Figure 6 shows the railway map with a mask calculated in 1 frame. Figure 7 shows an image of the railway track, on which a mask is applied, averaged over successive 5 frames.

**Fig. 7** Another image of railway track with a mask over successive 5 frames



### 3 Results

The accuracy of the classifiers was determined: 83% for the nearest neighbor method and 64% for the method of support vectors machine. It is determined that the combination of these methods, as well as additional operations on the obtained set of results, allowed to obtain an acceptable approximate railway tracks.

### 4 Conclusion

Thus, a method for searching the track gauge has been developed as a combination of machine learning and computer vision algorithms. To increase the accuracy of the classifier, the procedure of sequential application of the algorithms of  $K$ -nearest neighbors, support vectors, and frame averaging is used. The working capacity of the developed algorithm is shown experimentally. The method can be considered as one of the ways of image segmentation. The main advantages of the algorithm that implements the method are associated with minimal preparation of the training sample and the ability to analyze its work for further improvement. The described method of segmentation is an element of solving the main tasks of machine vision: positioning and identifying obstacles in the way of the train. Further development of the system is associated with the complexing data obtained from lidar and radar scanners and other algorithms of semantic image segmentation.

## References

1. Shantanu, I., Phute, M.: Tesla autopilot: semi autonomous driving, anuptick for future autonomy. *Int. Res. J. Eng. Technol.* **3**(9), 369–372 (2016)
2. Ukai, M., Nassu, B.T., Nagamine, N., Watanabe, M., Inaba, T.: Obstacle detection on railway track by fusing radar and image sensor. In: *World Congress on Railway Research (WCRR) 2011*, Lille, France (2011)
3. Lisanti, G., Karaman, S., Pezzatini, D., Del Bimbo, A.: A multi-camera image processing and visualization system for train safety assessment. *Multimedia Tools Appl.* **77**(2), 1583–1604 (2018)
4. Mukojima, H., Deguchi, D., Kawanishi, Y., Ide, I., Murase, H., Ukai, M., Nagamine, N., Nakasone, R.: Moving camera background-subtraction for obstacle detection on railway tracks. In: *IEEE International Conference on Image Processing (ICIP) 2016*, pp. 3967–3971. IEEE Press, New York (2016). <https://doi.org/10.1109/ICIP.2016.7533104>
5. Gleichauf, J., Vollet, J., Pfitzner, C., Koch, P., May, S.: sensor fusion approach for an autonomous shunting locomotive. In: *International Conference on Informatics in Control, Automation and Robotics ICINCO 2017: Informatics in Control, Automation and Robotics*, pp. 603–624 (2017)
6. Gebauer, O., Pree, W., Stadlmann, B.: Autonomously driving trains on open tracks - concepts, system architecture and implementation aspects. *Inf. Technol.* **54**(6), 266–278 (2012). <https://doi.org/10.1524/itit.2012.0689>
7. Weichselbaum, J., Zinner, C., Gebauer, O., Pree, W.: Accurate 3D-vision-based obstacle detection for an autonomous train. *Comput. Ind.* **64**(9), 1209–1220 (2013). <https://doi.org/10.1016/j.compind.2013.03.015>
8. Gavrilova, N.M., Dailid, I.A., Molodyakov, S.A., Boltenkova, E.O., Korolev, I.N., Popov P.A.: Application of computer vision algorithms in the problem of coupling of the locomotive with railcars. In: Vavilov, D. (eds.) *IEEE International Symposium on Consumer Technologies (ISCT)*, pp. 1–4. IEEE Press, New York (2018). <https://doi.org/10.1109/ISCE.2018.8408904>
9. Long, J., Shelhamer, E., Darrell T.: Fully Convolutional Networks for Semantic Segmentation. *arXiv: 1411.4038v2* (2015).
10. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(12), 2481–2495 (2017). <https://doi.org/10.1109/TPAMI.2016.2644615>
11. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. *arXiv.org* (2015). <https://arxiv.org/abs/1505.04597>
12. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: speeded up robust features. *Comput. Vis. Image Underst. (CVIU)* **110**(3), 346–359 (2008). <https://doi.org/10.1016/j.cviu.2007.09.014>
13. Makarov, A., Bolsunovskaya, M., Zhigunova, O.: Comparative analysis of methods for keypoint detection in images with different illumination level. In: Abramov, A.D., Murgul, V. (eds.) *MATEC Web of Conferences. Siberian Transport Forum*, 239(3), 01028 (2018). <https://doi.org/10.1051/mateconf/201823901028>
14. Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K.: KNN model-based approach in classification. In: Meersman, R., Tari, Z., Schmidt, D.C. (eds.) *On the Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE. OTM 2003. Lecture Notes in Computer Science*, vol. 2888. Springer, Berlin (2003). [https://doi.org/10.1007/978-3-540-39964-3\\_62](https://doi.org/10.1007/978-3-540-39964-3_62)
15. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition. IEEE Press, New York* (2014) <https://doi.org/10.1109/CVPR.2014.81>
16. Ben-Hur, A., Weston, J.: A user’s guide to support vector machines. In: Carugo, O., Eisenhaber, F. (eds.) *Data Mining Techniques for the Life Sciences. Methods in Molecular Biology*, vol. 609. Humana Press, Totowa (2010)
17. Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C., Silverman, R., Wu, A.Y.: An efficient k-means clustering algorithm: analysis and implementation. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**(7), 881–892 (2002). <https://doi.org/10.1109/TPAMI.2002.1017616>

18. Harwood, B., Drummond, T.: FANNG: fast approximate nearest neighbour graphs. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, pp. 5713–5722. IEEE Press, New York (2016). <https://doi.org/10.1109/CVPR.2016.616>
19. Makarov, A.S., Bolsunovskaya, M.V.: The method of obtaining orthographic top view for around view system at vehicles. In: IEEE Russia Section Young Researchers in Electrical and Electronic Engineering Conference, ElConRus 2017, pp. 697–699. IEEE Press, New York (2017). <https://doi.org/10.1109/ElConRus.2017.7910652>
20. Bruls, T., Porav, H., Kunze, L., Newman, P.: The right (angled) perspective: improving the understanding of road scenes using boosted inverse perspective mapping. In: 2019 IEEE Intelligent Vehicles Symposium (IV). IEEE Press, New York (2019). <https://doi.org/10.1109/IVS.2019.8814056>

# Hardware and Software System for Collection, Storage and Visualization Meteorological Data from a Weather Stand



Pavel Pankov , Igor Nikiforov , and Yufeng Zhang 

**Abstract** In the modern world, the use of measuring sensors and microcontrollers is growing every year. Their application ranges from household appliances to industrial automation. The article describes the process of developing and creating a software and hardware system designed for collection, storage and visualization meteorological data. The system includes a stand with weather sensors, a microcontroller and a software. It allows user to receive data from the stand with a microcontroller and to interact with the MongoDB database. As a result of the research and the work, a prototype of a hardware and software system was developed, which provides the tracking of atmospheric environmental indicators, their processing, storage and visualization. In addition, the possibilities for further project development are described.

**Keywords** Big data · Microcontroller · Weather · Sensors · esp32 · MongoDB · Data storage · Data processing

## 1 Introduction

Devices, which include sensors and microcontrollers, have gained ground over the world. People are surrounded everywhere by measuring sensors, such as climate sensors, space sensors, lights and microcontrollers. These components are used both in the simplest household appliances, such as air conditioning, and in the most complex control systems of assembly workshops in industrial enterprises [1], or it can be used in medical application [2].

In everyday life, it is important for people to take into account the weather conditions, however, weather forecasts on TV or on the Internet are not always correct,

---

P. Pankov (✉) · I. Nikiforov  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [pankov.pavel.a@gmail.com](mailto:pankov.pavel.a@gmail.com)

Y. Zhang  
Baidu Inc., Beijing, China

and the location may not be accurate. Therefore, there is a high need for software and hardware to obtain information about the current weather outside.

The hardware and software system must collect relevant information from sensors located indoors or outdoors, and be able to visualize the results of work on a mobile application or a stationary workstation.

In addition, such system can be used in places where it is useful to know the meteorological indicators in the garage [3], in the country, for construction work, or to have similar stands in industrial premises to monitor and comply with technological processes [4].

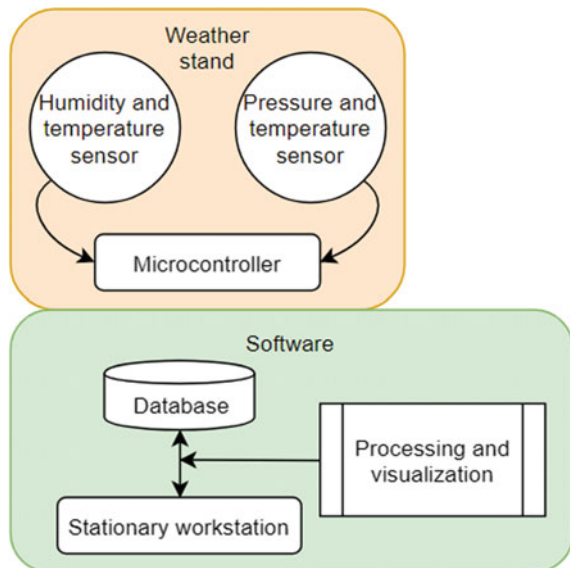
## 2 Architecture of Hardware and Software System

The article proposes the following structure of the software and hardware system, shown in Fig. 1, which includes a stand with meteorological sensors and a microcontroller that collects and processes data from sensors and software to work with these data for a local or a mobile workstation.

We can define the following basic modules:

- environmental sensors: temperature sensor, atmospheric pressure sensor and humidity sensor;
- microcontroller that is responsible for polling sensors, collecting information and sending it over the network to a computer;

**Fig. 1** Structure of the hardware and software system





- software system that provides data reception from the controller, processing of received data, its recording in the database and visualization.

In turn, the software system consists of:

- database;
- module for receiving and recognition data;
- module for data processing and data visualization.

The database stores information obtained from the microcontroller.

The module for receiving and writing data receives data packets from the microcontroller over the network, processes them and writes data to the database.

The visualization module provides a graphical representation of received and processed data from the database as necessary.

The main difference between this project and existing meteorological stations on the market, such as the Oregon Scientific BA900 [5] is the data storage. That means that the sensors data is not simply collected and displayed. The entire amount of data is stored in the database for a certain time so that they can be further analyzed and visualized statistics [6].

### 3 Analog Review

At the moment on the Internet, there are many projects of personal weather stations. Analogs<sup>1,2,3</sup> for comparison were considered and their main idea, similarities, advantages and disadvantages were highlighted.

Most projects use Arduino as a microcontroller. One of the advantages of this microcontroller is its low price. Arduino provides a simple project development platform, and many examples can be found on the Internet.

All similar projects have the following similar features:

- small dimensions;
- low cost of software development;
- available hardware components;
- low cost of the final product;
- the possibility of use in everyday life;
- mobility of the application.

One common disadvantage was identified when reviewing projects.

Data collection is carried out for a single display on the display. Some projects allow you to write data to the SD memory card. However, this method has significant disadvantage:

---

<sup>1</sup><https://create.arduino.cc/projecthub/ragingradish/improved-weatherstation-20x4-18dd89>.

<sup>2</sup><https://create.arduino.cc/projecthub/GilettaStefano/aws-arduino-weather-station-9e5a21>.

<sup>3</sup>[https://create.arduino.cc/projecthub/Arduino\\_Genuino/mkr-zero-weather-data-logger-574190](https://create.arduino.cc/projecthub/Arduino_Genuino/mkr-zero-weather-data-logger-574190).

- one SD card can work with only one device;
- it is need to pull out the SD card to save data one the computer;
- SD card is not a reliable storage device;
- small amount of memory;
- there is no possibility of collecting data from different devices.

The main idea of the project being developed is the possibility of remote collection and storage of information.

## 4 Hardware Selection

### 4.1 Microcontroller Selection

One of the problems solved in the work is the choice of the element that will collect data from the sensors and send them to the local workstation.

Several variants of microcontrollers [7] were considered: Arduino [8, 9], STM 32 [10], ESP 32 [11], Raspberry Pi [12–14].

Let us introduce the main selection criteria for microcontrollers:

- price (K1). The price is taken not of the microcontroller itself, but of the development board, which include the microcontroller and the necessary components for its full operation;
- clock frequency (K2). The clock frequency is the number of processor cycles (operations) per second;
- support of information transfer channels the Wi-Fi (K3). Having access to an Ethernet network using Wi-Fi is mandatory for wireless data exchange;
- open-source libraries for working with selected sensors (K4). This criterion is rather important, since availability of open source libraries for work with a particular sensor significantly reduces development time. The libraries can be used as a basis for the development, it is possible to optimize it and leave only the necessary functions.

Table 1 shows the characteristics of the considered microcontrollers, where “+” means the presence of what is described in the criterion and “–” means its absence.

**Table 1** Microcontrollers characteristics

Microcontroller	K1, ₺	K2, MHz	K3	K4
Arduino	100–1000	8–48	Add board	+
STM 32	1000–6200	24–216	Add. board	–
ESP 32	400–600	160/240	+	+
Raspberry Pi	2900–3800	1400	+	+

As a result of the comparative analysis, the ESP 32 microcontroller was selected, since its capacity is sufficient to complete the task, and the price is one of the lowest.

## ***4.2 Sensors Selection***

The selection of sensors is carried out to obtain the following data in operation:

- ambient temperature;
- atmosphere pressure;
- humidity.

The following sensors are selected:

- BMP180 GY-68. It combines an atmospheric pressure sensor and a temperature sensor. I2C sensor interface. Accuracy of pressure determination is 0.02 hPa. Accuracy of temperature determination is 0.01 °C;
- SI7021 HTU21. It combines a humidity sensor and a temperature sensor. I2C sensor interface. Accuracy of humidity determination is 3% RH. The accuracy of temperature determination is 0.4 °C.

## **5 System for Collection, Storing and Visualization of Meteorological Data from the Weather Stand**

### ***5.1 Work with Database***

A document-oriented database MongoDB is chosen for data storage [15, 16].

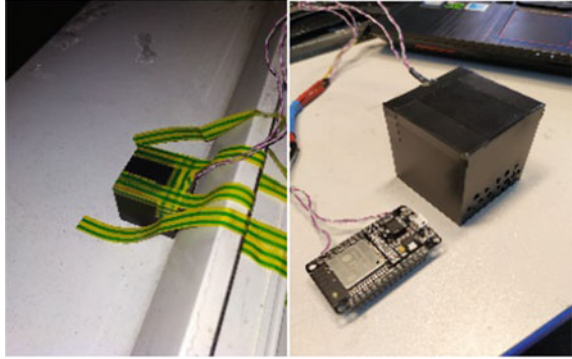
Due to the fact that MongoDB does not allow working directly with the ESP 32 microcontroller, and nothing was found from the considered libraries to solve this problem, it was decided to write data to the local MongoDB database from a computer and transfer data over the network to exclude use wired connection.

### ***5.2 System Development Stages***

The development process included the following stages:

- design and assembly of a stand with weather sensors and an ESP 32 microcontroller;
- writing firmware for the ESP 32 microcontroller, which collects data from sensors and sends data over the network using a home Wi-Fi network;

**Fig. 2** Assembled stand with sensors in the case and outdoors case with sensor



- writing software to connect a computer with a microcontroller and write data to the database;
- writing software for reading data from the database, its processing, analysis and visualization.

### ***5.3 Design of the Stand***

As mentioned earlier, the stand includes an ESP 32 microcontroller, an atmospheric pressure and temperature sensor BMP180 GY-68 and a humidity sensor SI7021.

The communication interface with the sensors is an I2C [17, 18]. The ESP 32 microcontroller has one data bus with an I2C interface; however, this interface allows using many sensors, and its address is used to select the required one.

Due to the fact that the sensors need to be placed outdoors, their use without the case is impossible, since precipitation can lead to short circuits or other situations that will lead to negative consequences.

On the other hand, it is also impossible to enclose the sensors in a completely sealed case, since the indication of the sensors will differ from those outside the case.

As a result, it was decided to place the sensors in a case that fixed outdoors, and the microcontroller with power indoors, which provided quick access to the microcontroller for rebooting and flashing it. The assembled stand is shown in Fig. 2.

### ***5.4 Algorithm for Data Collecting and Sending***

The microcontroller firmware had the following requirements:

- it requires constant interrogation of sensors and collection of actual data;
- sending data to a computer over the network should be implemented when receiving a command from a computer, meaning that it is ready to receive data.

First of all, libraries for interaction with sensors were implemented.

At the beginning, the microcontroller performs test requests to the sensors to determine their presence and correct operation.

After client has connected to the server, the microcontroller expects a command from the client, while continuing to poll the sensors for the latest up-to-date data.

After receiving the command from the client, the microcontroller sends four packets with the current sensor indications.

The data packet contains the packet prefix, the command identification number, the number of “useful bytes” which contains the data from the sensors, the data itself and the end of the packet.

After sending the data, the microcontroller waits for the command from the client again, while updating the data from the sensors.

## ***5.5 Receiving and Writing to a Database on a Computer***

The script for receiving data from the microcontroller and writing data to the database was written in Python [19, 20].

After launching the script, it connects to the local MongoDB database, which is operated using the PyMongo module.

After connecting to the database, a connection is made to the server using the IP address that ESP 32 receives after connecting to the Wi-Fi network. Then the script sends a command to receive data and waits for a response from the microcontroller.

If the waiting time for new data exceeds the specified time, the script reports an error and continues its work.

After receiving the message, the received packet is parsed in accordance with the described protocol for receiving and transmitting data.

If the received data packet was correctly parsed and identified, the received data is recorded in the database. Otherwise, an error packet message or non-existing identifier is displayed.

Below there is an example of the format for writing data to the database in JSON format:

```
{
  sensor: "HTU Temperature"
  data: "-0.16"
  date: "2018_12_12"
  ime: "23:33:12:565"
}
```

## 5.6 Reading and Visualizing Data from the Database

The following software was written to read and visualize data from the MongoDB database:

- Python script that uses standard PyMongo<sup>4</sup> module methods to connect to the database;
- Python script that builds a graph of changes in sensor readings based on the latest actual data from the database in real time;
- a program written in C# which shows the latest actual data from the sensors in the form of a simple widget with three fields and graphs of changes in the indicators also in real time.

The first Python script demonstrates how to use the standard PyMongo module. Script launch is possible with parameters:

- the name of the sensor which data is required;
- the date for which data is required. This parameter can also specify “today” or “yesterday”;
- if the third parameter is the date, then it is the end of the data collection period, and the second parameter is its beginning. This parameter also provides for the option “today.”

If the script is launched without specifying any parameters, all data on the default sensor will be requested.

After parsing transmitted parameters, the script receives from the database the data for the specified sensor and period of time and builds a graph for them, also indicating the average level.

The second Python script also accepts a sensor name as a parameter. After launching, the script builds a graph of the change in sensor data in real time. Connection to the database is carried out in the same way as in the second example, using the PyMongo module.

A program written in C# consists of two widgets in the form of two semitransparent windows. The first window displays up-to-date information on the three sensors requested from the base. The second window contains three graphs of changes in sensor readings since the launch of the program.

After starting the program and establishing a connection with the database, a timer is started, requesting data from the database and displaying this data on the forms.

A GitHub repository was created for the project, which contains all the source code. Repository is available by reference.<sup>5</sup>

---

<sup>4</sup><https://api.mongodb.com/python/current/>.

<sup>5</sup>[https://github.com/DucklingDark/Meteo\\_Station](https://github.com/DucklingDark/Meteo_Station).

## 6 Example of System Work

### 6.1 Python Scripts

An example of data visualization using Python tools is presented in Figs. 3 and 4. Figure 3 shows the collected values from the BMP180 pressure sensor for 9 days. Figure 4 shows the collected values from the HTU21D for 9 h. The horizontal line indicates the average value for displayed period of time.

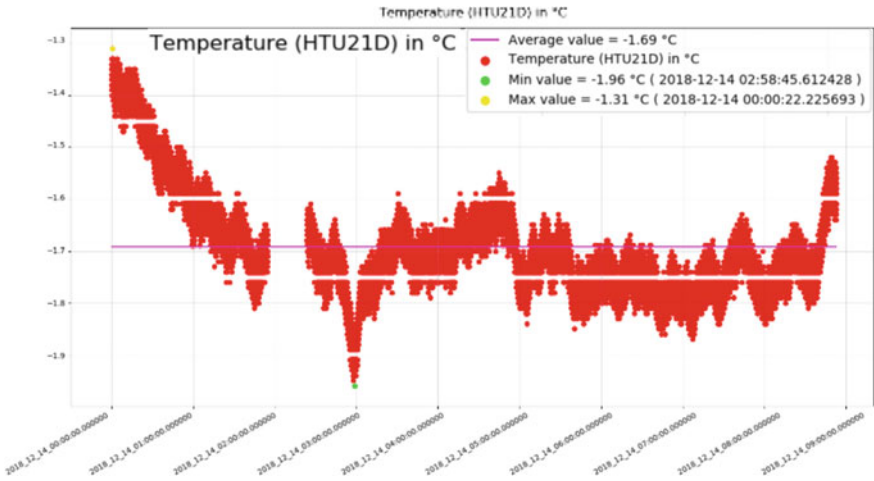


Fig. 3 Pressure values for 9 days

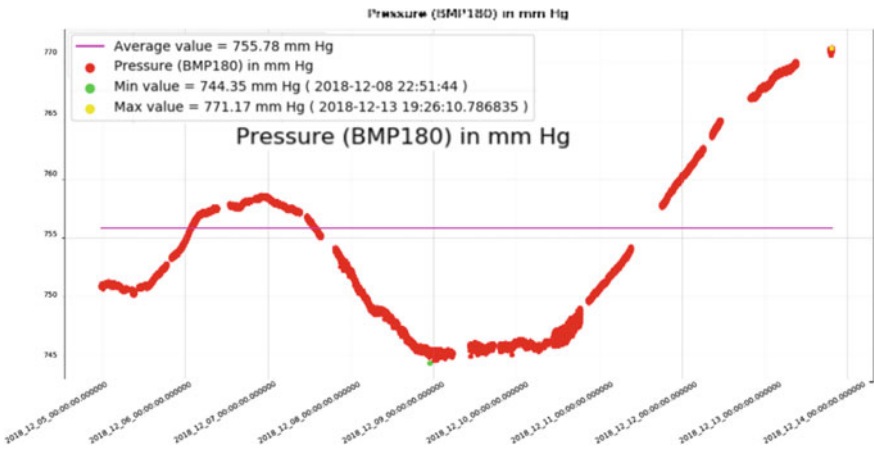
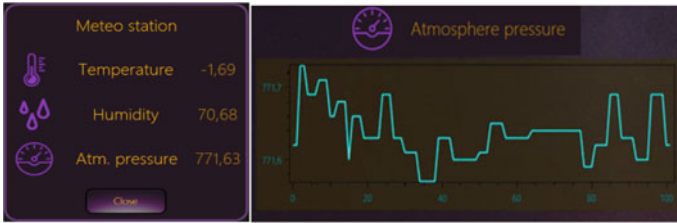


Fig. 4 Temperature values for 9 h



**Fig. 5** Window with the current digital values of the sensors and the graph of changes in atmospheric pressure for 10 s

Legend describes the chart. It contains the name of the sensor, the average value of the collected data, the maximum and minimum values with time stamps. There are also units of measurement in the legend.

There are data gaps on the charts. The stand was switched off and data collection was not performed during these periods of time.

## 6.2 C# Program

After starting the program, two windows open. Both windows are made in the form of semitransparent widgets with the ability to move around the screen:

- the first window contains three fields with the name of the indicators and the actual value that is currently relevant;
- the second window contains three graphs that shows changes in sensor readings.

If necessary, each of the windows can be closed, leaving the second. The program will stop its work only after closing both windows.

An example of the program is shown in Fig. 5.

## 7 Conclusion

As a result of the work, a hardware and software system was created. It can collect meteorological data, store, visualize and process it. The paper described comparative analyzes of microcontrollers and sensors, based on which ESP 32 and BMP 180 and SI7021 HTU 21D sensors were selected.

The main difference between the developed project and its analogs is the ability to store a large amount of data for further analysis of the current state of the environment in the room.



For the first stages of development, Raspberry Pi microcomputers were not considered due to their high cost, however, their use can simplify and improve the system, contributing to its development.

In addition, for the completeness of indications of the stand and the system, there is a need to add a speed and direction of wind, but these sensors are expensive and made in most cases for use in industrial areas.

The next stages of development can be the transfer of the database to cloud services and use the Raspberry Pi instead of ESP 32. In addition, there is a need to add the coordinates of the stand or a short name of the room to the database. On the other hand, user can use these stands in different rooms where user needs to know such indicators, for example, for a system for maintaining temperature and humidity. This will allow user to combine several systems into one centralized.

Such development of the system will allow access to data not only from the computer on which the database is located, but from any location with access to the Internet.

Also, it is necessary to send error messages to the database and to track their occurrence in order to fix stand and system problems in time, and to predict them if it is possible [21].

## References

1. Kapustin, N.M.: Automation of production processes in mechanical engineering. In: Kapustina, N.M. (ed.) Proceedings for Technical Colleges, 415 p (2004)
2. Vitabile, S., Marks, M., Stojanovic, D., Pillana, S., Molina, J.M., Krzysztan, M., Sikora, A., Jarynowski, A., Hosseinpour, F., Jakobik, A., Ilic, A.S., Respicio, A., Moldovan, D., Pop, C., Salomie, I.: Medical data processing and analysis for remote health and activities monitoring. [https://link.springer.com/chapter/10.1007/978-3-030-16272-6\\_7](https://link.springer.com/chapter/10.1007/978-3-030-16272-6_7)
3. Kasatkin, I., Egorov, M., Kotov, E., Zakhlebaev, E.: Cooling of a battery pack of a car, working on renewable energy. In: MATEC Web Conference, vol. 245, Article number 15003 (2018). <https://doi.org/10.1051/mateconf/201824515003>
4. Youssef, H.A., El-Hofy, H.A., Ahmed, M.H.: Manufacturing Technology: Material, Processes, and Equipment, 948 p (2011)
5. BA900 Weather Station. URL: <https://mobile-review.com/articles/2007/oregon-scientific-ba900.shtml>
6. Laboshin, L.U., Lukashin, A.A., Zaborovsky, V.S.: The big data approach to collecting and analyzing traffic data in large scale networks. *Procedia Comput. Sci.* **103**, 536–542 (2017). <https://doi.org/10.1016/j.procs.2017.01.048>
7. Knoll, M., Breitegger, P., Bergmann, A.: Low-power wide-area technologies as building block for smart sensors in air quality measurements. <https://link.springer.com/article/10.1007/s00502-018-0639-y>
8. Purdem, J.: Beginning C for Arduino, 2nd edn: Learn C Programming for the Arduino, 276 p (2012)
9. Blum, J.: Exploring Arduino, 1st edn, 387 p (2017)
10. Noviello, C.: Mastering the STM 32 Microcontroller, 782 p (2016)
11. Kolban, N.: Kolban's Book on ESP 32, 1228 p (2018)
12. Monk, S.: Raspberry Pi Cookbook, 412 p (2013)
13. Monk, S.: Programming the Raspberry Pi: Getting Started with Python, 192 p (2012)

14. Harrington, W.: Learning Raspbian, 154 p (2015)
15. Chodorow, K.: MongoDB: The Definitive Guide: Powerful and Scalable Data Storage, 432 p (2013)
16. Grolinger, K., Higashino, W.A., Tiwari, A., Capretz, M.A.M.: Data management in cloud environments: NoSQL and NewSQL data stores. <https://link.springer.com/article/10.1186/2192-113X-2-22>
17. Valdez, K., Becker, J.: Understanding the I2C Bus, 8p (2015)
18. Himpe, V.: Mastering the I2C Bus, 247 p (2011)
19. McKinney, W.: Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 466 p (2012)
20. Jones, B., Beazley, D.: Python Cookbook, 3rd edn. Recipes for Mastering Python, 706 p (2013)
21. Utkin, L.V., Zaborovskii, V.S., Popov, S.G.: Detection of anomalous behavior in a robot system based on deep learning elements. *Autom. Control Comput. Sci.* **50**(8), 726–733 (2016). <https://doi.org/10.3103/S0146411616080319>

# Plant Disease Recognition Based on Multi-dimensional Features of Leaf RGB Images



Basim Al-Windi  and Vladimir Tutygin 

**Abstract** A system based on the representation of sets of texture parameters of Haralik as multi-dimensional vectors, including the stages of normalizing leaf images, training, testing, recognition. The training stage includes geometric and parametric normalization of the source photo images, calculation of GLCM adjacency matrices, matrix of estimates of mathematical expectations and confidence intervals of the scatter of the Haralik parameters of the components R, G, B, RG, RB, GB of the source RGB photo images; the testing phase, performed by modeling statistical tests in order to determine the required quantity averaging parameters of the diagnosed photo images; recognition stage, based on a correlation comparison of the column vectors of the matrix of estimates of mathematical expectations with the column vector of the averaged values of the parameters of the diagnosed photo images.

**Keywords** Normalization of photo images · GLCM—matrix · Haralik texture parameters · Membership function

## 1 Introduction

Most plant diseases cause changes in the appearance of leaves in the visible spectrum.

To solve the problem of highlighting features on images for the purpose of classifying them (diagnosing diseases), various methods of forming a set of features are used to uniquely identify images, i.e., assign them to a particular class.

Most often, fuzzy logic methods and neural networks used to solve the problem of highlighting features in images of plant leaves in order to classify the type of plant

---

B. Al-Windi (✉)  
University of Diyala, Baqubah, Iraq  
e-mail: [gohn\\_smith2002@yahoo.com](mailto:gohn_smith2002@yahoo.com)

V. Tutygin  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

disease, and diagnostics performed both directly from color RGB or HSV images of leaves, and according to their texture descriptions.

### Related Work

Ananthi and Vishnu Varthini [1] in their project (detection and classification of plant leaf diseases) used RGB image as resource to their work, then used HSI color transformation; after the transformation, he take the intensity as a base for disease detection and masking the healthy region on pictures by doing threshold, his threshold idea was that he identify the mostly green colored pixels. After that, based on specified threshold value that is computed for these pixels, the mostly green pixels, if the green component of the pixel intensity is less than the recomputed threshold value, the red, green, and blue components of the this pixel are assigned to a value of zero. This is done in sense that the green-colored pixels mostly represent the healthy areas of the leaf, they do not add any valuable weight to disease identification, and furthermore, this significantly reduces the processing time. Removing the masked cells: the pixels with zeros red, green, blue components were completely removed. This is helpful as it gives more accurate disease classification and significantly reduces the processing time. This idea of deleting picture background is useful but it takes a long time because he make a threshold to get the result, which take more time. He used four parameters from GLCM (Contrast, Energy, Local homogeneity, Cluster shade, Cluster Prominence) as feature extractions.

His extension of this work will be developing algorithm for classification features.

Kaushal and Bala [2] in their work (GLCM and KNN-based Algorithm for Plant Disease Detection) used GLCM.

Algorithm to extracts textural features from the image, then they used the k-mean clustering algorithm for the segmentation of input images based on their properties and divide then to several regions. Their idea of defining the distance of nearest centroid is Euclidean distance. They did that by calculated consecutively for each data point and the data point having the minimum distance assigned to the cluster. These minimum points are summed up to get a centroid. After that, the SVM classifier is applied in the existing algorithm, which will classify the input image into two classes. They improve result performance of existing algorithm by replacing the SVM classifier with KNN classification. This action leads to improve the accuracy of disease detection because the KNN can classify more than two types of classes that are why I think they replaced the SVM.

Dhaware and Wanjale [3] in their project (A Modern Approach for Plant Leaf Disease Classification, which depends on Leaf Image Processing). Their work was consisting of preprocessing, segmentation, feature extraction, and different image classification techniques. His project dataset depends of 120 healthy and infected leaves. He resized the image to  $512 \times 512$  in order to keep up the consistency as far as size of the images because he depends on mobile picture, which the minimum size of it is 2 megapixels. Then he covert RGB image into HSV, for background subtraction he used cluster based and color based, in color based he did that by using R, G, B element and selecting G element because it has pixels more that R&B so he removed them. In cluster-based, the connected elements in the image are discovered out and

the immense part of the image is kept and other part is removed. He makes two classes of feature extractions first healthy image features and infected leaf features as a data set in segmentation part he used 21 methods like.

Number of the strategies are principal component analysis (PCA), fuzzy logic,  $K$ -nearest neighbor (KNN), support vector machine (SVM), artificial neural network (ANN), neuro-fuzzy interference system, etc.

KNN method classifies images with using nearest distance between trained dataset and testing dataset. Choosing the appropriate value for  $k$  is major drawback of KNN. The ANN applies estimation functions that depend on lots of inputs given to the system, which are known. The disadvantage of this method is over fitting problem.

In proposed paper, support vector machine (SVM) technique is used for the classification of images. SVM is the supervised learning method, which usually applied for pattern recognition and classification. In his project, he used many classification methods but he didn't mention the differences of accuracy which given by the project for each classification method he used.

Mahajan and Dhumale [4] focused on an image processing technique been used for detection of plant diseases. They used 57 images of leaves for test. For segmentation, they used Otsu's method. In feature extraction, parameters used are mean, standard deviation, and entropy, extract standard deviation, and extract kurtosis, skewness. The system gives higher accuracy as compared to the techniques used in the past. By using fuzzy logic, it gives an accuracy of 88% for the detection of leaf diseases. Reaching 88% its fear enough for improving their strategy but using 57 images of different leafs are little to back up their accuracy reality speed and accuracy are the main characteristics of disease detection [5].

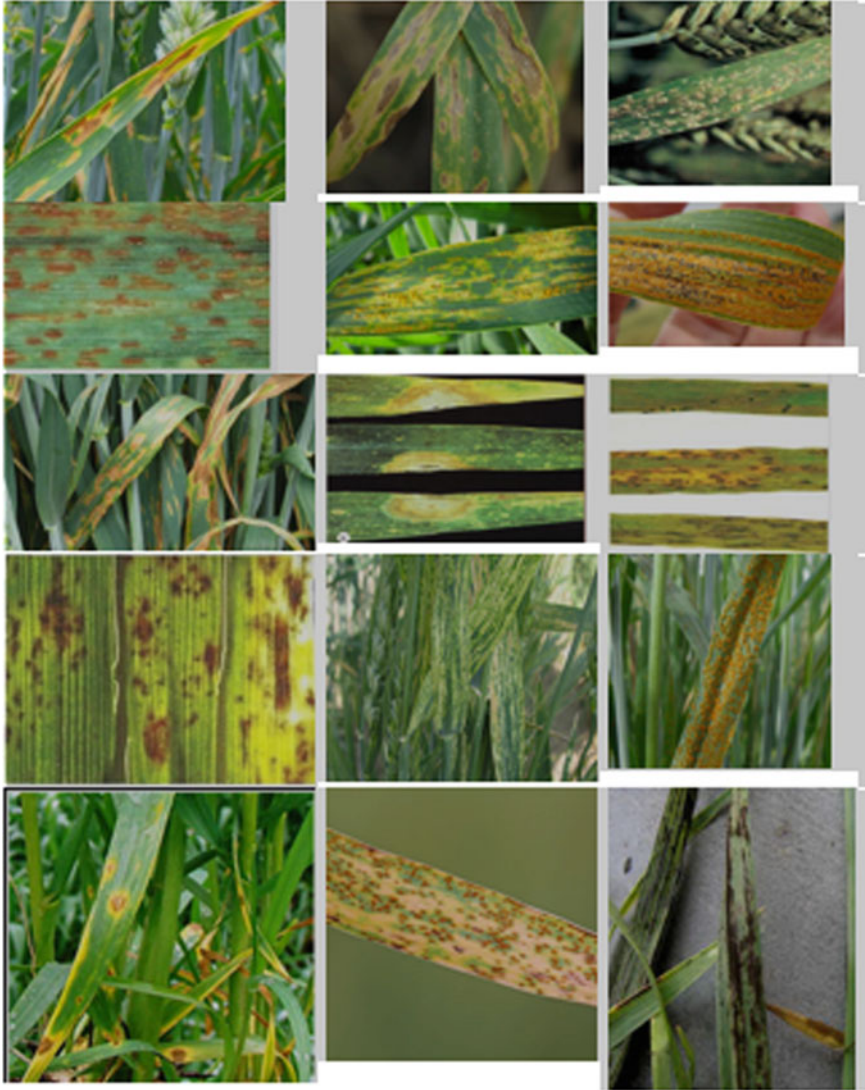
Sahaya Merlin and Sree Thayanandeswari. A Novel Approach to Detect and Classify Leaf Diseases Based on Image Processing [6] in their method for detection and classification of leaf diseases is implemented. The segmentation of the diseased part is done using K-means segmentation. Then, GLCM texture features are extracted, and classification is done using SVM. The method is tested for detection of diseases in citrus leaves.

In Jose et al. Haritham: a plant disease identification system [7] many recognition systems for identification plant diseases.

The described systems make it possible to recognize the type of plant disease if the number of diseases is not more than 5.

## 2 Our Proposed Recognition System

Our proposed work consists of identifying 15 diseases as minimum to cover major wheat leaf diseases. However, uniquely determine the type of plant diseases (in particular, wheat or soy) based on the metrics described impossible because the number of types of diseases in crops is sufficiently large (e.g., the number of major diseases of wheat or soybeans more than 15 [8]). Example wheat leaf images with different diseases is shown in Fig. 1.



**Fig. 1** Still images leaves of wheat in various diseases. 1—Septoria leaf spot (*Septoria*), 2—pirenophoroza (*Pyrenophora tritici-repentis*), 3—powdery mildew (*Erysiphe graminis*), 4—brown rust (*Puccinia recondita*), 5, 6—yellow rust (*Puccinia striiformis*), 7—leaves Septoria leaf spot (*Septoria tritici*), 8—snow mold (*Fusarium nivale*), 9—blotch (*Helminthosporium sativum*) 10—root rot, 11—stripe mosaic (wheat streak mosaic virus), 12—brown (sheet), rust [fungal diseases (*Puccinia triticina*)], 13—blotch (*P. tritici-repentis*), 14—linear (stem) rust diseases (*P. triticina*), 13—blotch (*P. tritici-repentis*), 14—linear (stem) rust (*Puccinia graminis*), 15—head smut (*Ustilago tritici*)

A characteristic feature of the distribution function of generalized indicators of facsimiles—considerable overlap, which eliminates the possibility of forming a detection threshold according to the criteria of an ideal observer or the Neyman–Pearson and unambiguous identification of the type of disease. To overcome this drawback, we propose the calculation of 6 sets of generalized histogram indicators for components R, G, B, RG, RB, GB, calculating a set of 15 distribution functions for each of the six sets of generalized indicators histograms R, G, B, RG, RB, GB and a final decision on the image belonging to one of 15 possible types by majority voting.

The known approach for solving the problem is based on a comparison of texture images based on the adjacency matrix (GLCM matrix [9, 10]). In this case, the object of analysis is not an image matrix, the adjacency matrix and R, G, B, RG, RB, GB, which are calculated based on the basic parameters: Contrast, Correlation, Energy, Entropy, Homogeneity. The direct use of these parameters for identifying the type of the disease does not lead to uniquely correct recognition results.

Our approach is to form generalized indicators based on indicators: Contrast, Correlation, Energy, and Homogeneity [9]:

1. contrast:

$$CN = \frac{1}{(G - 1)^2} \sum_{u=0}^{G-1} \sum_{v=0}^{G-1} |u - v|^2 p(u, v);$$

2. correlation:

$$CR = \frac{1}{2} \sum_{u=0}^{G-1} \sum_{v=0}^{G-1} \frac{(u - \mu_u)(u - \mu_v)}{\sigma_u^2 \sigma_v^2} p(u, v) + 1;$$

3. energy:

$$EN = \sum_{u=0}^{G-1} \sum_{v=0}^{G-1} p(u, v)^2;$$

4. homogeneity:

$$HM = \sum_{u=0}^{G-1} \sum_{v=0}^{G-1} \frac{p(u, v)}{1 + |u - v|},$$

where  $u, v$ —coordinate of the adjacency matrix,  $G$ —number of gray levels,  $\mu_u, \mu_v, \sigma_u,$  and  $\sigma_v$ —mean values and standard deviations of row  $u$  and column  $v$  of the matrix matches, respectively. The above definition will ensure that all functions have the range [1].

### 2.1 Plant Disease Recognition System on Fuzzy Logic

The formation of the results of recognition applies fuzzy logic [11]. The expediency of its use in solving the problem of plant disease diagnosis by images of leaves was considered in [12]. A distinctive feature of our proposed solution to this problem lies in the fact that it involves the calculation of the functions belonging to the standard descriptions of each of the 6 sets of R, G, B, RG, RB, GB, binarization results and a final decision on the image belonging to one of 15 possible types by majority voting.

Figure 2 shows the proposed structure of the plant disease diagnostic system from the images of leaves.

The proposed method of diagnosing the type of images of leaves of plant diseases is based on the adjacency matrix of parameters and majority voting and consists of two stages.

The first step is calculation Contrast, Correlation, Energy, Homogeneity parameters, and comparison with standard descriptions in the form of parameters ranges of values for all 15 diseases and binarization diagnostic results (a value of the binarized result of the comparison is 1—if parameter value included in the reference description of the range for a given disease, and 0—if not included). Table 1 shows results of simulation of the diagnostic process for the case when analyzed adjacency matrix for the red component image, and the model parameter values of Contrast, Correlation,

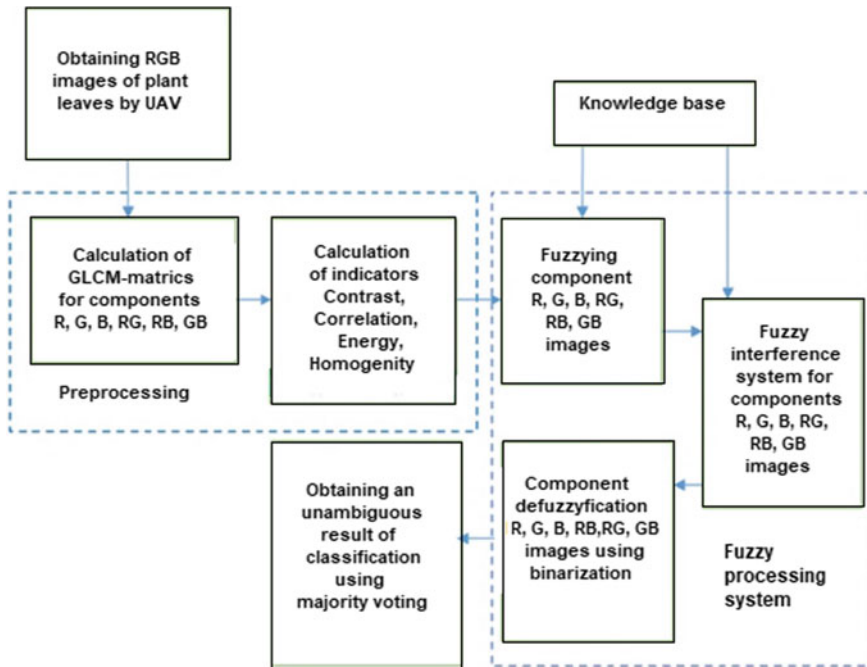


Fig. 2 Structure of the plant disease diagnostic system from the images of leaves





intervals DCN, DEN, DCR, DHM values for these parameters of the distribution functions. In conducting the experiment model expectations CN indicators, EN, CR, HM were taken equal to the parameter values obtained by image processing for all 15 diseases and confidence intervals DCN, DEN, DCR, DHM were taken equal to 0.016. Statistical evaluation of the diagnostic results by the proposed method was carried out for each of 15 diseases. The result of model experiments conducted 15,000 (1000 experiments for each of the 15 disease) proper diagnosis was 93.6–96%.

### Description of the Algorithm

Initial data for the diagnostic algorithm is the standard description of leaf images for all diseases: the expectations CN0 indicators, EN0, CR0, HM0 values and confidence intervals DCN, DEN, DCR, DHM for the values of the parameters of the distribution functions.

#### Sequencing

1. Calculation of parameters values: Contrast, Correlation, Energy, Homogeneity for all image components: R, G, B, RG, RB, GB (component number = 1.0.6) and calculating the comparison source image with the reference indices sheet descriptions for all components and all diseases: KCN ( $i, j$ ), KEN ( $i, j$ ), KCR ( $i, j$ ), KHM ( $i, j$ ), ( $j = 1.0.6, i = 1.0.15$ ):

$$\begin{aligned} KCN(i, j) &= \frac{|CN(i, j) - CN0(i, j)|}{DCN(i, j)}; \\ KEN(i, j) &= \frac{|EN(i, j) - EN0(i, j)|}{DEN(i, j)}; \\ KCR(i, j) &= \frac{|CR(i, j) - CR0(i, j)|}{DCR(i, j)}; \\ KHM(i, j) &= \frac{|HM(i, j) - HM0(i, j)|}{DHM(i, j)}. \end{aligned}$$

2. Primary binarization parameters:

$$\begin{aligned} KCN_B(i, j) &= 1, \text{ if } KCN(i, j) \leq 1, \text{ or } 0, \text{ if } KCN > 1; \\ KEN_B(i, j) &= 1, \text{ if } KEN(i, j) \leq 1, \text{ or } 0, \text{ if } KEN > 1; \\ KCR_B(i, j) &= 1, \text{ if } KCR(i, j) \leq 1, \text{ or } 0, \text{ if } KCR > 1; \\ KHM_B(i, j) &= 1 \text{ if } KHM(i, j) \leq 1, \text{ or } 0 \text{ if } KHM > 1. \end{aligned}$$

3. Primary majority voting and secondary binarization results of the voting:

$$\begin{aligned} KB(i, j) &= KCN_B(i, j) + KEN_B(i, j) + KCR_B(i, j) + KHM_B(i, j); \\ KBB(i, j) &= \text{round}\left(\frac{KB(i, j) - 0.1}{4}\right). \end{aligned}$$

4. Majority voting and the final binarization:

$$KBG(i) = \sum_{j=1}^6 KBB(i, j);$$

$$K(i) = \text{round}\left(\frac{(KBG(i) - 0.1)}{6}\right).$$

## 2.2 Plant Disease Recognition System Using Neural Network

To solve the problem of identification of the disease on the leaves of plants, the still images we studied the possibility of using deep convolutional neural network (SNS) [12, 13]. Advanced soft computing methods like deep learning have proven its success in image recognition and detection tasks in vast areas [8]. Neural networks are tolerant to noisy inputs. We chose the convolutional neural network precisely because it allows us to recognize specific visual signs, in our case, the foci of the disease, regardless of where in the leaf of the plant they are. For this purpose, the input of the neural network is supplied images which are not in the form of a one-dimensional array, as it would be with other types of neural networks, and as a matrix of pixels. The same matrix will be obtained and after convolutional layer: They will continue to be a spatial structure corresponding to the original image.

Convolution—a linear transformation of the input data a special kind. If the sign card (template image) in the layer with number  $l$ , the result is a two-dimensional convolution with a kernel size of  $2d + 1$  and the weight matrix  $W$  size  $(2d + 1) \times (2d + 1)$  for the next layer will be the

$$y_{i,j}^l = \sum_{-d \leq a, b \leq d} (W_{a,b} * x_{i+a, j+b}^l)$$

wherein  $y_{i,j}^l$ —the convolution result to level  $l$ , and  $x_{i,j}^l$ —its input, i.e., the output of the entire previous layer. In other words, to obtain component  $(i, j)$  of the next level, we apply a linear transformation to the square of the previous level window, that is, scalar multiply pixels of the window by the vector convolution. Consider the application of a convolution matrix with the matrix size of  $5 \times 5 \times 3 \times 3$  Weights:

$$\begin{pmatrix} 0 & 1 & 2 & 1 & 0 \\ 4 & 1 & 0 & 1 & 0 \\ 2 & 0 & 1 & 1 & 1 \\ 1 & 2 & 3 & 1 & 0 \\ 0 & 4 & 3 & 2 & 0 \end{pmatrix} * \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 2 & 1 & 0 \end{pmatrix} = \begin{pmatrix} 9 & 5 & 4 \\ 8 & 8 & 10 \\ 8 & 15 & 12 \end{pmatrix}$$

After convolution in a neural network layer should nonlinearity. As the function  $h$  using ReLU—linear rectification unit that computes a function, in fact, is an operation clipping the negative portion of a scalar value which will be applied to each element of the obtained matrix.

$$z_{i,j}^l = h(y_{i,j}^l) f(x) = \max(0, x)$$

However, that’s not all, in addition to linear convolution and the accompanying nonlinearity should be one more operation—down sampling (pooling; in Russia, it is called more subsample operation). The meaning of subsampling is simple: The convolutional network is much more presence or absence of a particular trait than its exact coordinates. Therefore allowed “generalization” allocated signs of losing part of the information on their location, but reducing the dimension. Typically, the subsampling to each local group of neurons is applicable maximum taking operation—max-pooling. Thus, max-pooling represented as, where  $d$ —a subsampling window size.

$$x_{i,j}^{l+1} = \max_{-d \leq a \leq d, -d \leq b \leq d} z_{i+a, j+b}^l$$

Thus, standard convolution network layer consists of three components:

1. Convolution in the form of a linear map which distinguishes local features;
2. Nonlinear function of the applied component-wise to the results of the convolution;
3. Down sampling.

The important part is the training. Let us assume that we optimize some error function  $E$ , and already know its value at the outputs of our convolutional layer. To perform iteration of training, you need to understand how they are expressed through values of the gradient of the error on the weights.

Let us go for convolution layer. After taking the maximum error function passes without change, making layer subsampling passing through the graph calculating gradients sparse because all of the partial derivative elements of the window relate to only one—the maximum, the rest will be zero gradient. Pass through a nonlinearity: And on the level of convolution appear weight, which must be able to teach. The difficulty is that all weights are divided and each participates in all the outputs, so the amount is large enough

$$\begin{aligned} z_{i,j}^l \frac{\partial E}{\partial x_{i,j}^{l+1}} \frac{\partial E}{\partial y_{i,j}^l} &= \frac{\partial E}{\partial z_{i,j}^l} * \frac{\partial z_{i,j}^l}{\partial y_{i,j}^l} = \frac{\partial E}{\partial z_{i,j}^l} * h'(y_{i,j}^l) \\ \frac{\partial E}{\partial w_{a,b}^l} &= \sum_i \sum_j \frac{\partial E}{\partial y_{i,j}^l} * \frac{\partial y_{i,j}^l}{\partial w_{a,b}^l} = \sum_i \sum_j \frac{\partial E}{\partial z_{i+a, j+b}^{l-1}} \end{aligned}$$

where  $i$  and  $j$  run over all elements of the image on the intermediate layer  $y_{i,j}^l$ .

It remains only to miss the gradients of the previous layer:

$$\frac{\partial E}{\partial x_{j,j}^l} = \sum_a \sum_b \frac{\partial E}{\partial y_{i-a,j-b}^l} * \frac{\partial y_{i-a,j-b}^l}{\partial x_{i,j}^l} = \sum_i \sum_j \frac{\partial E}{\partial y_{i-a,j-b}^l} * w_{a,b}$$

This is the procedure for back-propagation in the convolution layer.

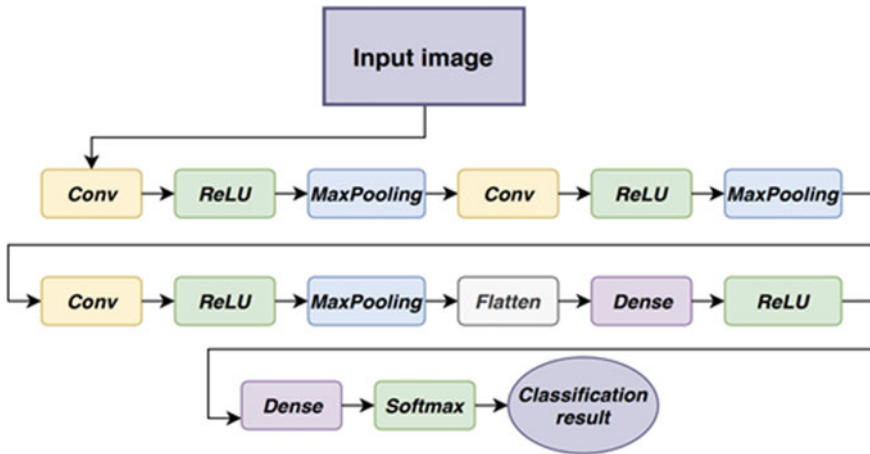
However, a convolution layer that cannot express the relationship between the pixels arranged far from each other, which convolution layers are repeated several times to solve this problem, is built to deep convolutional network. It would seem that the more layers, the better, but this is only partially true. By increasing the layers is a significant complication of the model, which usually does not give a strong growth in its efficiency. Yield of the next layer is to be used as input for the next, and it is clear that due to subsampling layer resolution will decrease. After the last layer of convolution using the Flatten the layer, is responsible for reformatting the input data in a full mesh layer, which was already to be a layer of Dense classification with the number of neurons equal to the number of classes that need to identify + Activation function. As activation, function was chosen Softmax. Softmax—a generalization of the logic function in the multi-dimensional case. Function converts a vector of dimension N in a vector of the same dimension, wherein each coordinate of the resulting vector is represented by a real number in the interval [1], and the sum of the coordinates is equal to 1. Therefore, the index of the maximum element in the resultant vector is a class index to which the neural network is carried preview.

To solve the problem of the identification of the image sheet plant diseases, deep convolutional neural network consisting of three convolutional layers was realized with window convolution  $3 \times 3$ , the number of neurons in each layer 32, followed by transformation layer into a fully connected list and then a classification layer for 3 types of disease. The structure of the neural network is shown in Fig. 3.

We were only investigated 300 of wheat leaf images (100 healthy, 100 patients, and 100 Septoria—brown rust). The whole sample was divided as follows: 70%—education, 20%—in the training, and validation of 10%—testing. A result of convolutional neural network has been created, and the proportion of correct diagnosis of disease was 83.33%.

### 2.3 Main Results of Statistical Modeling of Fuzzy Logic and Neural Network System

1. To diagnose the type of plant diseases on RGB—images of leaves with a significant number of possible diseases reliable results are obtained by calculation of 24 indicators Contrast, Correlation, Energy, Homogeneity; GLCM matrix components for R, G, B, RG, RB, GB images of leaves, the use of fuzzy logic in the step of defuzzification—binarization signs and majority voting. As a result



**Fig. 3** Structure of the neural network

of 15,000 model experiments for all 15 true diseases share diagnosis of disease was about 95%.

2. The reliability of diagnostic results can be improved when used for the diagnosis diseases of plants instead indicators Contrast, Correlation, Energy, Homogeneity and for the single sheet average values of these same parameters for several leaves. For a software implementation of the proposed algorithms of diagnostics of diseases, it is advisable to specify the parameters of the distribution functions of the key parameters Contrast components, Correlation, Energy, Homogeneity standard descriptions in the operation and, thus, improve the accuracy of diagnosis.
3. For the diagnosis of the type of plant diseases by RGB—leaves the still images, we proposed the use of a convolutional neural network. We conducted using 300 wheat leaf images (100 healthy, 100 patients and 100 Septoria leaf rust) studies have shown that using our proposed convolutional neural network for 3 types of diseases the proportion of correct diagnosis of the disease was approximately 83%.

The greatest application in solving problems of recognizing plant diseases from leaf images was found by texture features using adjacency matrices (GLCM matrix for grayscale images), features based on spatial frequency measurements, features using statistical characteristics of images (average, energy, variation, uniformity, contrast, correlation coefficient, entropy, differential dispersion), signs based on the description of structural elements [14–17].

### 3 Plant Disease Recognition System on Multi-dimensional Descriptions of Leaf RGB Images

#### 3.1 Normalization

The source photographs of the leaves of the plants require preprocessing before performing the diagnosis of diseases, regardless of how the diagnosis will be performed. A typical example of the original image is shown in Fig. 4.

When using a neural network for diagnostics, it is possible not to carry out preprocessing (normalization); however, this requires a multiple increase in the source images at the training stage and reduces the percentage of correct recognition results. In any case, performing image normalization increases the likelihood of correct recognition. In normalization, however, this requires a multiple increase in the source images at the performing image normalization increases the likelihood of correct recognition. In addition, images of leaves of plants affected by the same disease can visually differ significantly, especially taking into account the phase of the disease [8].

Normalization should solve the problems:

1. Removal of non-informative parts of images (it is necessary to leave only images of informative parts of leaves with signs of disease);
2. Standardization of the shape, size, and orientation of the informative parts of the leaves. Processing of 1300 images showed that the optimal format for the informative parts of wheat leaf images is  $300 * 100$  pixels;
3. Unification of parameters of informative parts of images (brightness, contrast). The need for this is caused by unequal lighting conditions for the leaves as a whole or their individual parts when shooting.

As a result, we get a normalized RGB image (see Fig. 5).

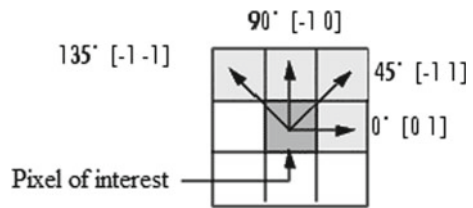
**Fig. 4** The original image of a leaf of wheat sick with Septoria





**Fig. 5** Examples of normalized images of wheat leaves, healthy and diseased. 1—Healthy plants; 2—brown rust; 3—yellow rust; 4—mildew; 5—pyrenophora; 6—striped mosaic; 7—Septoria; 8—dark brown spotting

**Fig. 6** Adjacency matrices directions



### 3.2 Using an Adjacency Matrix (GLCM Matrix)

Based on the RGB—normalized image of the plant leaf, 6 GLCM matrices can be obtained for the R, G, B, RG, RB, GB components, for each of the GLCM normalized matrices, the main characteristics of the texture Contrast, Correlation, Energy, Homogeneity, called the Haralik parameters [9, 10]:

Typically, the number of gray-level gradations in the original image is preliminarily reduced to eight, so the size of the GLCM matrices is 8 \* 8.

Adjacency matrices can be calculated for four possible directions (Fig. 6). The total number of Haralik parameters is 14. Thus, the maximum number of texture parameters of Haralik will be 14 \* 6 \* 4 = 336.

### 3.3 Reference description

Reference description may be presented in the form of a matrix of estimates of mathematical expectations for the scatter of the Haralik parameters for each color component R, G, B, RG, RB, GB of the source RGB photo images for all N diseases (see Table 3).



**Table 3** Reference description matrix

<i>i</i>	<i>j</i>					
	1	2	3	4	...	<i>n</i>
1	$A_{11}$	$A_{12}$	$A_{13}$	$A_{14}$	...	$A_{1n}$
2	$A_{21}$	$A_{22}$	$A_{23}$	$A_{24}$	...	$A_{2n}$
3	$A_{31}$	$A_{32}$	$A_{33}$	$A_{34}$	...	$A_{3n}$
4	$A_{41}$	$A_{42}$	$A_{43}$	$A_{44}$	...	$A_{4n}$
5	$A_{51}$	$A_{52}$	$A_{53}$	$A_{54}$	...	$A_{5n}$
6	$A_{61}$	$A_{62}$	$A_{63}$	$A_{64}$	...	$A_{6n}$
7	$A_{71}$	$A_{72}$	$A_{73}$	$A_{74}$	...	$A_{7n}$
8	$A_{81}$	$A_{82}$	$A_{83}$	$A_{84}$	...	$A_{8n}$
9	$A_{91}$	$A_{92}$	$A_{93}$	$A_{94}$	...	$A_{9n}$
...	...	...	...	...	...	...
<i>m</i>	$A_{m1}$	$A_{m2}$	$A_{m3}$	$A_{m4}$		$A_{mn}$

**Table 4** Target image options

<i>i</i>	
1	$B_1$
2	$B_2$
3	$B_3$
4	$B_4$
5	$B_5$
6	$B_6$
7	$B_7$
8	$B_8$
9	$B_9$
...	...
<i>m</i>	$B_m$

Here *j* is the number of the disease, *i* is the number of the Haralik parameter. Only one generalized matrix of estimates of mathematical expectations can be constructed, in which the row number determines not only the number of the Haralik parameter, but also the number of the color component, for example, *i* = 1 will correspond to the color component—R, the parameter—Contrast, *i* = 2—the color component—R, parameter—Correlation, etc.

Diagnostic image parameter set (Target image options) is a column vector (Table 4).

### 3.4 Target Image Description

The probability of a correct diagnosis of plant diseases depends on the degree of proximity of the calculated Haralik parameters of the target image from the reference description, i.e., from the mathematical expectations of the Haralik parameters for this disease. The permissible deviation of the Haralik parameters of the target image from the reference description at which the confidence probability of correct recognition will be a given value, for example, 0.95, can be set in the form of a confidence interval of the permissible deviation  $\beta$ . The value of  $\beta$  can be obtained by simulation.

In the event that the confidence intervals for the deviation of the Haralik parameters of the target image from the reference description are greater than this value, we propose to diagnose using the average values of the Contrast, Correlation, Energy, Homogeneity parameters for several ( $M$ ) analyzed leaf images (this is possible, because in the area of the disease affliction, there are always several plants), and the number of averaged parameter values should be selected from the condition that the confidence interval, taking into account the averaging of the parameter values Contrast, Correlation, Energy, Homogeneity was no more than  $\beta$ . If the confidence interval of values of a parameter without taking into account averaging is  $\beta_1$ ,  $\beta_1 > \beta$ , then, assuming that the distribution law for the values of this parameter is close to normal, the required quantity averaging  $M$  can be calculated based on the well-known expression for the confidence interval of mathematical expectation as

$$M = \frac{\beta_1^2}{\beta^2}$$

### 3.5 Diagnostic Algorithm

An algorithm for diagnosing plant diseases using multi-dimensional reference descriptions could be built based on fuzzy logic [18, 19]. A simpler solution is the use of a correlation comparison of the Haralik parameters of the target image and the reference description. The decision on whether the target image belongs to a particular class is made on the condition of maximum membership function.

1. calculation of membership function:

$$MF(j) = \frac{\sum_{i=1}^m (A(j, i) - \overline{A(j, i)})(B(i) - \overline{B(i)})}{\sqrt{\sum_{i=1}^m (A(j, i) - \overline{A(j, i)})^2 \sum_{i=1}^m (B(i) - \overline{B(i)})^2}}$$

where

$A(j, i)$  element of column  $j$ , row  $i$  of the matrix of the reference description;

- $B(i)$   $i$ th element of the vector—column of averaged Haralik parameters of diagnosed plant leaf samples;
- $m$  the total number of Haralik parameters for all components of R, G, B, RG, RB, GB of normalized leaf images.
2. calculation of the number  $k$  of the most probable disease according to the maximum membership function:

$$k = j, \text{ if } MF(j) = \max(MF(j)).$$

## 4 Conclusion

1. A necessary condition for correct recognition of plant diseases with a confidence probability of more than 0.95 with a significant number (up to 15) of diseases is the use of Haralik parameters: Contrast, Correlation, Energy, Homogeneity of the GLCM matrix for components R, G, B, RG, RB, GB of leaf images, image normalization when creating reference descriptions, normalizing, and averaging the parameters of target images, correlation comparison of the set of averaged Haralik parameters of the target normalized leaf images with the mathematical expectations of the Haralik parameters for each of the diseases.
2. As a result of 1000 model experiments for each of the 15 image classes when the confidence interval of the spread of the averaged values of the Haralik parameters of the target normalized photo images of the leaves is 0.016, the share of correct diagnosis was 97%.
3. The reliability of the diagnostic results increases if you expanding the list of Haralik parameters of the reference description or increase the number of averaged values of the Haralik parameters of the target normalized photo images of the leaves.

## References

1. Ananthi, S., Vishnu Varthini, S.: Detection and Classification of Plant Leaf Diseases (2012)
2. Kaushal, G., Bala, R.: Certificate of state registration of a computer program. Int. J. Adv. Res. Electr. Electron. Instrum. Eng. (2017)
3. Dhaware, C.G., Wanjale, K.H.: A modern approach for plant leaf disease classification which depends on leaf image processing. In: International Conference on Computer Communication and Informatics, Jan 2017, Coimbatore, India (2017)
4. Mahajan, V., Dhumale, N.R.: Leaf disease detection using fuzzy logic. Int. J. Innov. Res. Sci. Eng. Technol. 7(6) (2018)
5. Ashish, P., Tanuja, P.: Survey on detection and classification of plant leaf disease in agriculture environment. Int. Adv. Res. J. Sci. Eng. Technol. (IARJSET) 4(4) (2017)
6. Sahaya Merlin, M., Sree Thayanandeswari, C.S.: A novel approach to detect and classify leaf diseases based on image processing. Int. J. Sci. Res. Eng. Dev. 2(2) (2019)

7. Jose, J., Jayachandran, H., George, A.S., Jiya, S., Pratap, A.: Haritham. A plant disease identification system. *Int. J. Inf. Syst. Comput. Sci.* **8**(2) (2019)
8. Koyshibaev, M.: *Wheat Diseases*. Food and Agriculture Organization of the United Nations (FAO), Ankara (2018)
9. Stancheva, Y.: *Atlas of Crop Diseases. T.3., Diseases of Field Crops*. Sofia—M. PENSOFT (2003)
10. Haralick, R.M.: Statistical and structural approaches to texture. *Proc. IEEE* **67**(5), 768–804 (1979)
11. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **3**, 610–621 (1973)
12. Shtovba, S.D.: *Vvedenie v teoriju nechetkih mnozhestv i nechetkiju logiku [Introduction to the Theory of Fuzzy Sets and Fuzzy Logic]*. URL: [matlab.exponenta.ru/fuzzylogic/index.php](http://matlab.exponenta.ru/fuzzylogic/index.php)
13. Aung, Ch.H., Tant, Z.P., Fedorov, A.R., Fedorov, P.A.: *Razrabotka algoritmov obrabotki izobrazhenij intellektual'nymi mobil'nymi robotami na osnove nechjotkoj logiki i neironnyh setej. Jelektronnyj zhurnal «Sovremennye problemy nauki i obrazovanija»*. № 6 (2014)
14. Tutygin, V.S., Al'-Windi Basim, K.M.A., Ryabtsev, I.A.: *Sistema raspoznavanija boleznjej rastenij na osnove nechjotkoj logiki I neironnyh setej. Sovremennaja nauka: aktual'nye problemy teorii i praktiki. Serija «Estestvennye i tehniczeskie nauki»*. M.: Nauchnye tehnologii. № 3, pp. 107–115 (2019)
15. Denisjuk, V.S.: *Algoritmy vydelenija osobennostej na izobrazhenijah s cel'ju klassifikacii zabojevanij rastenij*. URL: [iis.nsk.su/files/articles/sbor\\_kas\\_16\\_denisjuk.pdf](http://iis.nsk.su/files/articles/sbor_kas_16_denisjuk.pdf)
16. Xu, K.-M.: *Using the Bootstrap Method for a Statistical Significance Test of Differences Between Summary Histograms*. NASA Langley Research Center, Hampton, VA. URL: [ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20080015431.pdf](http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20080015431.pdf)
17. Barbedo, J.G.A.: *Digital Image Processing Techniques for Detecting, Quantifying and Classifying Plant Diseases*. Barbedo Springer Plus. URL: [springerplus.com/content/2/1/660/](http://springerplus.com/content/2/1/660/) (2013)
18. Bianconi, F., Harvey, R., Southam, P., Fernandez, A.: *Theoretical and Experimental Comparison of Different Approaches for Color Texture Classification*. URL: [pdfs.semanticscholar.org/31a0/cf98ca459ab6e4676ac45700cc2485358347.pdf](http://pdfs.semanticscholar.org/31a0/cf98ca459ab6e4676ac45700cc2485358347.pdf)
19. Tutygin, V.S., Al-Windi Basim, K.M.A.: *Sistema klassifikacii teksturnyh izobrazhenij na osnove nechjotkoj logiki. Sovremennaja nauka: aktual'nye problemy teorii i praktiki. Serija «Estestvennye i tehniczeskie nauki»*. M.: Nauchnye tehnologii. No. 3, pp. 99–106 (2019)
20. Tutygin, V.S., Al-Windi Basim, K.M.A., Leliuhin, D.O.: *The use of an extended set of key texture features Haralick in the diagnosis of plant diseases on leaf images. Vibroeng. Procedia* **25**, 122–127. In: 39th International JVE Conference in St. Petersburg, Russia, 25–26 June (2019)

# Methodology of Service Development with a Single Application Programming Interface



Vitaly Monastyrev , Pavel Drobintsev , and Petar Kochovski 

**Abstract** Users of the services can interact with the application using a browser or using mobile devices. The most popular mobile platforms today are iOS and Android. Development of any service includes backend (application logic, database) and frontend (interface) part. Development of frontend part for web, iOS and Android parts is carried out separately, but you can use a single API, which is implemented in the backend part, instead of implementing different backend parts for each platform. In this article, we will consider the architecture of a single API and describe the methodology of its development, which allows you to save resources when creating a service.

**Keywords** Mobile development · Web development · Application programming interface · Backend architecture

## 1 Introduction

When developing any service, it is necessary to try to reach the largest possible audience of users. If 10 years ago the service was usually implemented only in the form of a web application, today it is also necessary to develop for mobile platforms. The most popular are Android and iOS. Development for several platforms requires significant resources, so you need to try to optimize this process as much as possible [1]. One way is to implement a backend part that will support a single Application Programming Interface (API) for all platforms.

We will consider creating a single API based on the Representational State Transfer (REST) API architecture. By a single API, we will mean a system built on the REST API architecture, which does not contain duplicate methods for different

---

V. Monastyrev (✉) · P. Drobintsev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [vit34-95@mail.ru](mailto:vit34-95@mail.ru)

P. Kochovski  
University of Ljubljana, Ljubljana, Slovenia

platforms, but implements only one method used by all platforms. The REST API architecture implies that server-side methods are implemented that perform certain actions, for example, return values from the database, change values in the database, load the sent file on the server, and so on. The main problem here is that the implemented methods should work equally well for different clients (Android, iOS, web). For example, different devices have different speed of connection to the Internet, and if for one device ten photos in high resolution from the server are downloaded quickly, then for another device 10 photos from the server need to be shipped in lower quality. This problem could be solved by implementing two different methods, but it is not effective. In the following chapters, we'll look at the methodology of the process of creating a single API that allows you to not implement different methods for different devices.

At the moment, there are no standards for how a single API should be implemented, although there are various studies in the field [20–22]. Existing studies consider either REST API capabilities, or only a single item—documentation, or a specific tool. In this article, we will consider the methodology for using REST API regardless of the framework used. Usually, each company decides for itself how it will implement its API. Also, the team chooses in what language it will do it. This may depend on the goals of the service (C++ is better for high-load systems, NodeJS is better for fast development, etc.), the preferences of the team, the experience of creating previous solutions, the existing code base, and so on. As examples of a single API, it is possible to consider Twitter, Vkontakte, Yandex, and Google. API of these companies can be used for authorization on other platforms, obtaining information from the account, etc. It is also a source of income for some companies. For example, Yandex [2] and Google [3] provide access to the API of their maps for a certain price.

However, many companies often do not develop a single API that would be optimized for all platforms and release their applications only on one of the platforms. For example, Prisma [4] and Face App [5] are available for iOS and Android, but are not available in the web version, which could attract new users.

When developing a single API, you need to consider many factors to avoid rewriting it in the future, adding duplicate queries, system failures [6], and so on. Since at the moment, there is no single methodology for developing single APIs and different authors offer different approaches [7, 8], in this article we have tried to build a methodology on the example of our own development. The main purpose of this article is a general methodology for developing a single API, taking into account the problems that may arise. We offer the following methodology points that will help to solve most of the problems that will arise in the development:

1. Application architecture. The most important part, as it defines the architecture of the entire project.
2. User registration and authorization. Usually, to be able to use all the functions of the service, the user needs to register in it, so it is necessary to provide a mechanism for registration and authorization of the user.

3. Using the same GET/POST requests for different clients. You need to understand what methods are needed for each platform and how they can be combined so as not to implement duplicate or redundant code.
4. Documenting the implemented API. You need to make sure that the developers of the frontend part have documentation and understand why one or the other method is implemented.
5. Backward compatible versions. When updating the API, in addition to taking care of backward compatibility of older versions of the application, you also need to take care of compatibility of all platforms.

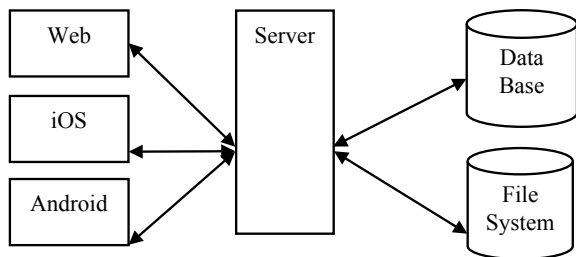
Thus, our development methodology is represented by five points. It is worth noting that we recommend that the development follow the above sequence of actions. The first step is to determine the structure of the project, then it is convenient to do authorization and registration, as it affects how the three points will be implemented. After implementing step 3, you need to document the API. You can then consider backward compatibility for future releases.

## 2 Application Architecture

Consider a single system architecture that uses a single API (Fig. 1). This architecture is general and allows you to abstract from specific programming languages and frameworks. The architecture of the application discussed in this article is shown in Fig. 1. It is worth noting that this architecture can be used in the development of almost any service, regardless of its purpose.

Let's consider the presented architecture in more detail. The interaction of the client parts of the application with the server takes place via the REST Protocol using GET/POST requests. The information is transmitted in JSON [9] format. All GET/POST requests implemented on the backend side are the API of our service.

**Fig. 1** Architecture of the service



## 2.1 Frontend

The first three modules (web, iOS, and Android) are the frontend part of the app. Web—to view and interact with the service through any browser, iOS—for Apple smartphones and Android for smartphones based on Android.

1. We propose to implement the web part using two submodules—frontend and backend. The frontend part is responsible for UI rendering and user interaction. The backend part will work with the single API of the service. This separation simplifies the development process. In addition, in this case it will be easier to correct requests to the service if the API changes. As an example of the framework—for the backend part you can use Laravel [10], and for the frontend part Angular [11].
2. iOS—a client that allows you to work with the service using a device running iOS. The main development languages for iOS are objective-c and swift. XCode was used as the development environment. To send GET/POST requests to the server, it is very convenient to use the free libraries SwiftyJSON [12] and Alamofire [13]. It is worth noting that all requests must be executed asynchronously so as not to block the main thread.
3. Android—a client that allows you to work with the service using a device running Android. The most popular development languages are Kotlin and JSON. It is convenient to use gradle [14] as a build system. Requests are sent to the server via REST-Protocol, data format—JSON.

## 2.2 Backend

As mentioned earlier, the team decides in what language backend will be implemented. It can be C++, Python, NodeJS, Java, etc. Later in the article, we will give examples implemented in Java + Spring, but similar technologies are available for other programming languages. Any database can also be used as a database—Oracle, Postgresql, MySQL, mongoDB. This depends on the amount and type of data the service is working with. Our team used for their purposes and for the MySQL [15] server connection was used SpringJPA [16]. To store media files, you should use the file system and store only links to these resources in the database.

Because any service has users, you must create a user table in the database. This table has an id, user name, and mail field that is unique. It also stores the user's password as a hash for greater security. It is worth noting that the table does not store an access token, which is described in more detail in the next section.



### 3 User Registration and Authorization

User registration must be supported on any of the three clients described above. Also, users must be able to log in through any of the three clients and access the functionality of the application. To solve this problem, you can use JSON Web Token (JWT) [17].

Also, in our case we used Spring filters to filter requests with and without tokens, but almost all other languages have similar frameworks. Filters in the Spring framework filter out all requests that come to the server and do not contain a token. Without a token, only two methods are available—registration of a new user and getting token. Moreover, (new user registration and token acquisition) are the same for all frontend clients.

Filters in Spring are implemented using Spring Security. In our case, we created a special class `WebSecurity` extends `WebSecurityConfigurerAdapter` and added an annotation `@EnableWebSecurity`. After that, we redefined the method `configure` as follows:

The authentication filter is used to register a new user, and authorization to access the service of an already registered user. For the new user registration method, this is achieved by sending JSON to a controller whose content looks like this:

```
{
    "username": "someUsername",
    "password": "somePassword",
    "email": "someEmail@email.com"
}
```

The JSON body is sent from the client to the server via https, which allows traffic to be encrypted. Library to create POST request and use JSON in web applications, iOS and Android. Therefore, we use only one method in the server-side controller that allows you to register a new user using a template defined by the JSON body. Upon successful registration, an access token is returned to the user in the response, which may or may not have a lifetime. This token is generated on the server side and encodes inside the user name, which can be decrypted only using a special key stored on the server. This allows you to further accept requests containing an access token and, on its basis, to determine from which user the request came. The token encryption algorithm is chosen by the developer. In this case, we used HMAC512.

Typically, the token is stored after it is received and used when sending requests. For example, to the web client a token is stored in the web session, and for the iOS client token is stored using `SwiftKeychainWrapper`.

If the user is already registered, he can request his token by sending it to the JSON server with the following content:

```
{
    "username": "someUsername",
```

```
"password": "somePassword"  
}
```

This information is also transmitted via https. The ability to obtain a token is necessary, since the user can register through one client and save the token on one device, and then try to log in from another device.

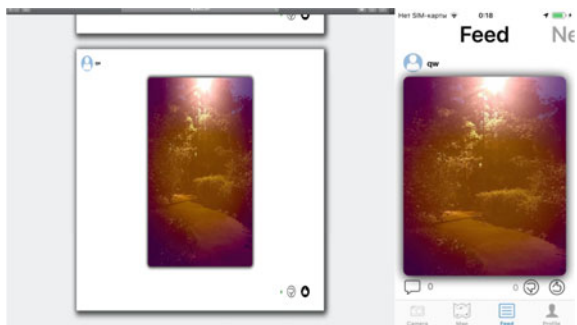
When sending any other request needed in the request header, add the following key/value pair—Authorization/“bearer someAccessToken”. Thanks to the token, it is possible to accurately determine which user sent the request, regardless of the client used by him. As a result, the problem of registration and authorization of users from different clients is solved by only two methods on the server side and the use of JWT tokens for authorization.

It is worth noting that we do not store tokens in the database, as this is not necessary. The user can enter the application from any number of devices and all of them will be issued a token. The most important thing is that we create a token ourselves using a secret word that is stored on the server and the user name is encrypted inside the token, which allows us not to store it in the database.

#### 4 Using the Same GET/POST Requests for Different Clients

The next problem that occurs is the use of the same GET/POST requests by different clients. Thus, before developing it is necessary to determine in advance what functionality the service should perform and strive to ensure that the web, Android application and iOS application work according to the same logic and provide the same functionality. In addition, it should be noted that it is also worth striving to ensure that all clients have the most similar interface with the necessary adaptations for specific platforms. This allows users to seamlessly switch from one client to another. Below are examples of the interface of our application. Figure 2 shows the interface on web and the interface on the iOS.

Fig. 2 Web and iOS interface



When working out the GET/POST requests necessary for the functionality of the application, the following factors should be taken into account: as a rule, the speed of the station Internet is higher than the speed of the mobile Internet, so it can request more information from the server for one request. Another factor is the amount of RAM on different mobile devices that is available to the app. On older devices, there may be a situation that when requesting information from the server there is not enough memory to store the result in the cache (e.g., if the server requests photos, videos or other files).

One of the most single problems in development is loading information by pages. Pagination should work equally well on iOS, Android, and the web. As part of our methodology, we offer the following universal approach. The information with the request comes to the server in the form of JSON and has the following form:

```
{
  "page": "1",
  "id": "23",
  "pageSize": "10"
}
```

“Page” is the page number that you want to return. “Id” is the number of the record from which the page is counted. This is done to ensure that when the user has reached the end of the list and the server needs to upload the next batch of data does not load those that have already been uploaded, because during this period, new records may appear and then the page will shift. This is the number of records that are contained on 1 page. “PageSize” is set depending on the client used. For example, both iOS and Android allow you to get the size of the available RAM. Based on the average amount of memory occupied by one record, the developer can set the desired value.

Thus, the creation of the single requests to upload information using pagination solves the problem of app usage across different customers and different mobile device. We have only one upload method on the server side instead of a bunch of different methods.

## 5 Documenting the Implemented API

Documenting the API is one of the most important parts in development. Clear documentation gives developers an understanding of how to use the provided API. Documentation can reduce or even completely eliminate the time that the development team spends discussing the frontend part and the backend part. In addition, documentation is necessary if you develop only the backend part of the service, and the client part can be developed by third-party developers. This practice is used for example in Telegram [18], V Kontakte [19], etc.

For the possibility of further API support it is recommended to implement methods responsible for different entities in different controllers. So, it will be easier to Orient backend developers and will not cause confusion in the project.

Each team chooses its own method and place for documentation. In our case, the documentation for the single API is based on the following points:

1. Request address. Typically, this is the data transfer Protocol (http or https), server IP address, port, and request name. For example:

<https://127.0.0.1:8080/getBestPhotoArray>

2. Request type. For example, it can be GET, POST, PUT, PATCH, etc.
3. Header fields. For example, it can be Authorization, Content-Type, Accept, Cookie, etc.
4. Request JSON fields. These are the fields and the field type of the outgoing request that the backend command has identified as required. It is also worth noting which fields are required and which are not.
5. Response JSON fields. These are the fields and type that will be returned from the server to the client. Mandatory fields are also specified.
6. A description of the purpose of the method. Describes what the query does, what it was created for, and when to use it.

The API storage location is selected depending on who will use the API. If this is an internal command, it is recommended to use some internal portal, for example, in this project we used confluence. If you expect the API to be used by third-party developers, it makes sense to place it in a separate section of your website or put the documentation on git.

## 6 Backward Compatible Versions

When developing a single API service, you need to be very careful when modifying old methods. Any renaming of JSON fields can result in users with older versions of clients simply not being able to connect to the service. There are also several possible solutions to the problem.

First you need to understand who uses the service. If these are firm-specific services, you can modify existing methods to ensure that users are delivered a timely version and that all frontend development teams are ready to incorporate the changes into their client application implementations by the due date. But most often we are dealing with services that work with third-party users who can both update their application and not update.

If we work with users who do not always update their application, it is necessary to provide the possibility of backward compatibility with older versions. For example, if in the new version we need to return an additional field to render it to the UI, then we can just add another JSON field to the existing JSON method. This will also allow the frontend development teams to start using the new field when needed.

If the structure of the entire query changes completely, you should implement the new API method and mark the old method as obsolete. You can finally remove the deprecated method after the frontend development teams replace the old method with the new one, and most users update their application. For users who have not updated their application (there should be a small percentage), you can display a message on the download screen that their client version is outdated and no longer supported, they need to update their application.

## 7 Conclusion

As a result of the work, a methodology for developing a single API was proposed, which provides for five points: development of the application architecture, the use of JWT tokens for authorization and registration, optimization of GET/POST requests for different clients by adding additional parameters to requests, maintaining structured documentation and working out the possibilities of backward compatibility of the API.

The development methodology was used by our development team and allowed us to reduce the cost of implementing the server part. Instead of three backend API developers for different platforms, we only had one backend developer. The project size is about 3500 lines of code. It took us about four months to write the backend. Thus, we get a development speed of about 30 lines of code per day. Even if we assume that optimization for different platforms would take a thousand lines of code, and the speed would be, for example, 50 lines of code per day, since we do not need to negotiate with all the teams about the architecture, we will get that three separate parts would be implemented in 50 days, i.e., a total of 5 months instead of 4 (if 1 programmer). Or should we keep three developers instead of 1.

It is worth noting that this metric is not entirely objective. To write a method that will work equally well on all platforms, a person needs more time than a simple method. But in any case, the implementation of one method, instead of, for example, three is more profitable in terms of further support of the code. For example, if in the future we want the method that returns photos to the client to return more and text records, then it will be enough for us to edit and test one method, not three.

Future plans—continue to develop the service with a single API, search for new methods to improve and optimize the development process. Collect feedback from the frontend and backend development teams.

## References

1. Volkova, V.N., Loginova, A.V., Shirokova, S.V., Kozlovskay, E.A.: Development of the innovative IT-project and managing project human resources. Paper presented at the Proceedings of the 19th International Conference on Soft Computing and Measurements, SCM 2016, pp. 470–473 (2016). <https://doi.org/10.1109/SCM.2016.7519816>
2. API Yandex Maps. URL: <https://tech.yandex.ru/maps%20/>
3. Google Maps Platform Documentation. URL: <https://developers.google.com/maps/documentation/>
4. Prisma. URL: <https://prisma-ai.com>
5. Face App. URL: <https://www.faceapp.com>
6. Ziniakov, V.Y., Gorodetskiy, A.E., Tarasova, I.L.: System Failure Probability Modelling (2016). [https://doi.org/10.1007/978-3-319-27547-5\\_19](https://doi.org/10.1007/978-3-319-27547-5_19)
7. Raman, R.C.S.P., Dewailly, L.: Building RESTful Web Services with Spring 5, p. 228 (2018). ISBN 1788475895
8. Kalin, M.: Java Web Services: Up and Running, p. 320 (2009). ISBN 1449365116
9. Marrs: JSON at Work: Practical Data Integration for the Web, p. 320 (2017). ISBN 1449358322
10. Stauffer, M.: Laravel: Up & Running, 2nd edn, p. 544 (2019). ISBN 1492041211
11. Wilken, J.: Angular in Action, p. 320 (2019). ISBN 1617293318
12. SwiftyJSON. URL: <https://github.com/SwiftyJSON/SwiftyJSON>
13. Alamofire. URL: <https://github.com/Alamofire/Alamofire>
14. Darwin, I.F.: Android Cookbook: Problems and Solutions for Android Developers Android, p. 774 (2017). ISBN 1449374433
15. Murach, J.: Murach's MySQL, p. 612 (2012). ISBN 1890774685
16. Keith, M., Schincariol, M.: Pro JPA 2 (Expert's Voice in Java), p. 508 (2013). ISBN 1430249269
17. JSON Web Tokens. URL: <https://jwt.io>
18. Telegram APIs. URL: <https://core.telegram.org/>
19. VK Learning API. URL: [https://vk.com/dev/first\\_guide](https://vk.com/dev/first_guide)
20. Adamczyk, P.: REST and web services: in theory and in practice. In: REST: From Research to Practice, pp. 35–57 (2011). ISBN 978-1-4419-8302-2
21. Verborgh, R.: Survey of semantic description of REST APIs. In: REST: Advanced Research Topics and Practical Applications, pp. 69–89 (2013). ISBN 978-1-4614-9298-6
22. Inzunza, S.: API documentation. In: WorldCIST'18 2018: Trends and Advances in Information Systems and Technologies, pp. 229–239 (2018). ISBN 978-3-319-77711-5

# Using Symbolic Computing to Find Stochastic Process Duration Distribution Laws



Georgiy Zhemelev  and Alexandr Sidnev 

**Abstract** Stochastic processes describe the dynamic behavior of systems modeled by formalisms such as queuing networks, network planning models, semi-Markov processes, and some other. Flowgraph models provide an analytical approach to the problem of stochastic process duration distribution law (SPDDL) finding. The problem is solved in two stages. Initially, the moment generating function (MGF) of the process duration is to be obtained using the graphical evaluation and review technique (GERT) or the flowgraph algebra. This stage is straightforward in contrast to the second one—the analytical transition from the MGF to the process duration distribution law in terms of probability distribution function (cumulative or non-cumulative). The transition is nontrivial and is implemented in this study using MATLAB Symbolic Math Toolbox along with various examples of finding SPDDLs when the processes are represented as ordered activity sets. Also, the capabilities of the statistical flowgraph methodology can be extended over the case when flowgraphs have parallel branches. The results of symbolic calculations are validated via simulation using GPSS World software. This study opens up real possibilities of replacing simulation with symbolic mathematics when searching for duration distribution laws of stochastic processes spawned by flowgraph models.

**Keywords** Flowgraph models · Moment generating function · Padé approximation · Laplace transform

## 1 Introduction

A significant number of system models for solving problems of finding a process duration can be turned into stochastic flowgraphs. For instance, in queuing networks, we have the process of moving a customer from the start node to the end node. In network planning models, Markov and semi-Markov processes with absorbing states, there is a process of transition from the initial to the final state. The topic of applying

---

G. Zhemelev (✉) · A. Sidnev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [www.dev@gmail.com](mailto:www.dev@gmail.com)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021  
N. Voinov et al. (eds.), *Proceedings of International Scientific Conference on Telecommunications, Computing and Control*, Smart Innovation, Systems and Technologies 220, [https://doi.org/10.1007/978-981-33-6632-9\\_7](https://doi.org/10.1007/978-981-33-6632-9_7)

the methodology of stochastic flowgraphs to various mathematical models is being actively developed at present [1–4]. Both the GERT [5] and the flowgraph algebra [6] allow finding moment generating functions for random variables characterizing the duration of various processes that are adequately described by these graphs. Despite the fact that the MGF uniquely determines the distribution and is better suited for finding moments, it is often necessary to obtain a description of the process duration in the form of a probability density, a distribution function, or a survival function.

In practice, the transition from a random variable MGF to an appropriate distribution law is a nontrivial task and requires complex symbolic calculations [7, 8]. Such a transition was implemented in the MATLAB environment with several examples of stochastic processes represented as ordered sets of activities with different distributions of their durations. At the same time, if it is possible to abandon simulation in favor of some analytical method for obtaining the process duration distribution law, then the accuracy of system behavior prediction can be significantly improved, especially for rare events. That is why we investigate symbolic computations and do not content ourselves with estimation of empirical distributions from flowgraph simulation results.

## 2 Materials and Methods

### 2.1 Analytical Method of Obtaining the Process Duration Distribution Law

The process duration distribution law extraction from the MGF is carried out by different methods [6]. We evaluate here the MATLAB implementation of one of them, suggesting the following sequence of steps.

1. Obtain the MGF for a given process flowgraph.
2. Find the Padé approximant of the MGF.
3. Perform an integral transform of the approximated MGF to obtain an approximated distribution density.
4. Obtain the cumulative distribution function and the survival function by integrating the resulting probability density distribution.

The last step is trivial, so let us consider in more detail each of the first three stages.

**Construction of a Flowgraph MGF.** Having data on activity scheduling and parameters of their duration distributions, we form a stochastic process flowgraph, in accordance with the GERT. Here, the graph nodes are the process states: initial, final, and intermediate ones. An arc  $ij$  of the graph symbolizes an individual activity from the set. Each arc is marked by its  $W$ -function:

$$W_{ij} = p_{ij}M(Y_{ij}), \quad (1)$$



where  $p_{ij}$  is the probability of performing the activity  $ij$  when the state  $i$  is reached, and  $M(Y_{ij})$  is the MGF of the random variable  $Y_{ij}$ , i.e., the activity  $ij$  duration.

It is well known that MGF can be acquired via Laplace transform of the corresponding probability density function:

$$M(Y_{ij}) = L[f](-s) = \int_0^{\infty} f(Y_{ij})e^{sY_{ij}} dY_{ij}, \quad (2)$$

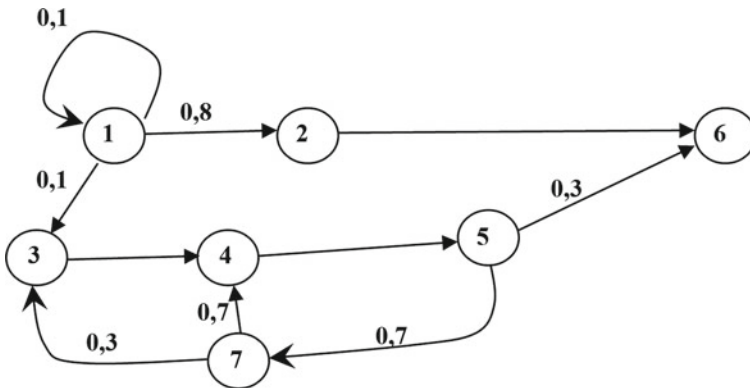
where  $L[f](s)$  is the Laplace transform of the function  $f(\cdot)$ .

It is important that for a significant number of distribution laws, the MGF can be obtained in an analytical form, which opens up the possibility to analytically construct the MGF  $M(Y_{0N})$  of a reduced stochastic flowgraph containing only the initial and final nodes ( $Y_{0N}$  is the duration of the process from the initial state 0 to the final state  $N$ ).

The GERT [5] and the flowgraph algebra [6] imply a significant restriction on the structure of stochastic graphs—concurrent activities are not allowed. This means banning parallel arcs in a flowgraph. In other words, at any given time, only one activity is in progress. Formally, this means that for each node of the stochastic flowgraph the following condition must be satisfied:

$$\sum_j p_{ij} = 1, \quad (3)$$

that is, the sum of the probabilities of the arcs coming from the node  $i$  is always equal to 1. This condition is satisfied for all nodes of the sample process graph analyzed below (see Fig. 1). Different methods of obtaining  $M(Y_{0N})$  from a stochastic flowgraph are discussed in detail in [5–7]. They give same results, can be programmed, and are beyond the scope of this paper.



**Fig. 1** An example of a stochastic process flowgraph (Graph #1)

**Padé Approximation of the MGF.** One can question the point regarding MGF approximation—why not right away transform the exact MGF to the corresponding distribution density? The fact is such a transformation is barely possible for many flowgraphs that appear in practice, especially the ones that have loops and non-exponential distributions on their arcs, since they cause very complex analytical expressions and hence symbolic computation software fails to perform the required integral transform [7, 8] (a similar problem arises in other applications that deal with inversion of Laplace transform [9–11]). Consequently, there is a need to approximate MGF so the transform becomes always possible and feasible for a symbolic computation engine.

Padé approximation (the choice of which will be justified in the Discussion section) enables us to represent an arbitrary function as a rational function of a given order:

$$M(Y_{0N}) \approx \frac{A(s)}{B(s)}, \quad (4)$$

where  $A(s)$  and  $B(s)$  are polynomials of orders  $a$  and  $b$ , respectively, and thus the approximation has the order of  $[a/b]$ . Under this technique, the approximant's power series agrees with the power series of the function it is approximating up to the  $(a + b)$ th derivative. The choice of the numerator and denominator orders of the Padé approximation in this paper is performed according to [6].

**Transition from MGF to Probability Density Function.** The integral transformation applied to transition from the approximated MGF to the corresponding approximation of the distribution density can be performed using the Heaviside expansion formula that gives a transparent inverse Laplace transform of a rational  $M(Y_{0N})$ :

$$f(Y_{0N}) = \sum_{k=1}^K \frac{A(r_k)}{B'(r_k)} e^{r_k Y_{0N}}, \quad (5)$$

where  $f(Y_{0N})$  is the probability density function that describes the sought-for SPDDL and  $r_k$  are (possibly complex valued) solutions of the  $B(s) = 0$  equation. For programmatic implementation of the Heaviside expansion formula, we suggest the use of the method of grouping of complex conjugate terms described in [12] so as to avoid complex exponents that cannot be simplified by a symbolic computation engine.

**Validation of the Suggested Method.** In order to validate the results of computations described above, we used the approach of stochastic flowgraph simulations and statistical hypotheses checking.

Simulation in GPSS World software [13] allows to obtain a sample of values of the random process duration on the basis of multiple execution of the entire sequence of activities in accordance with the process flowgraph. The resulting sample is used to construct a histogram and empirical distribution function, as well as to estimate the expectation and other moments of a random variable.

Having empirical and theoretical descriptions of the distribution law under consideration, it is possible to put forward a statistical hypothesis  $H_0$  (null hypothesis) of the given sample belonging to this law and to check it using a goodness-of-fit test. Based on the results of testing and comparing the moments, we can draw conclusions about the applicability of the suggested analytical method to obtain characteristics of distribution laws of stochastic process duration.

### 2.2 Stochastic Process Simulation with GPSS World

The GPSS World simulation system has all necessary expressive power to simulate various stochastic processes that can be interpreted as the processes of performing ordered sets of activities. A set of executable instructions (blocks) inherent in the GPSS World allows a user to create a single, but customizable template of a universal program suitable for modeling processes of arbitrary structure with different distributions of each individual activity duration. The algorithm of forming such a template of a GPSS program is given in [14]. It allows creation of a full simulation model for a set combining at most 60 activities with GPSS World Student Version limited to 180 blocks. The Student Version is provided free of charge. In accordance with the proposed algorithm, Dmitry Korenev, a graduate student of Peter the Great St. Petersburg Polytechnic University, developed a program editor of process diagrams named Violet-BP [14], which provides the ability to generate code of GPSS programs for simulation of described processes in order to estimate their time costs. The Violet-BP software is developed using the Java programming language and is based on Violet [15], an open-source framework for diagram editors. The screenshot of the main application window for Graph #1 (see Fig. 1) is shown in Fig. 2.

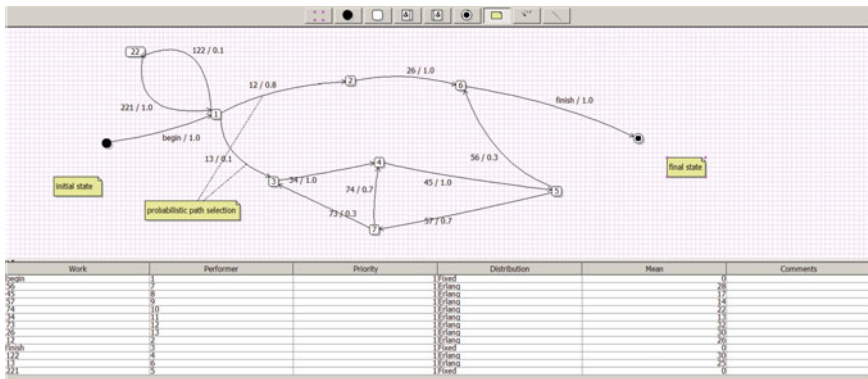


Fig. 2 Graph #1 in the main window of Violet-BP application

### 2.3 *Software Implementation of the Solution in MATLAB*

The suggested method of SPDDLs finding was implemented in MATLAB R2016b. The main program acts as a full-fledged study of analytical transition from the MGF of a stochastic flowgraph to the distribution law of its duration in terms of probability distribution function (PDF) and corresponding cumulative distribution function (CDF). The software is comprised of a number of MATLAB classes, scripts, and functions and implements the following stages of the described method:

1. Construction of the MGF for a specified graph and activity distributions.
2. Selection of optimal orders of numerator and denominator for Padé approximation.
3. Finding the Padé approximant of the MGF with the selected orders.
4. Laplace transform with the  $-s$  substitution to obtain an approximated PDF from the approximated MGF using the Heaviside expansion formula and grouping of complex conjugate terms.
5. Obtaining an approximated CDF by integrating the approximated PDF.
6. Comparison of the moments (expectation and variance) obtained from the exact and the approximated MGF.
7. Extraction of simulation results from GPSS World as a sample of duration of the given stochastic process.
8. Comparison of the moments evaluated from the sample with the moments obtained from the exact and approximated MGFs.
9. Construction of the empirical CDF and histograms from the sample.
10. Checking the statistical hypothesis of the sample belonging to the found distribution law that was obtained from the MGF of the studied stochastic process duration.

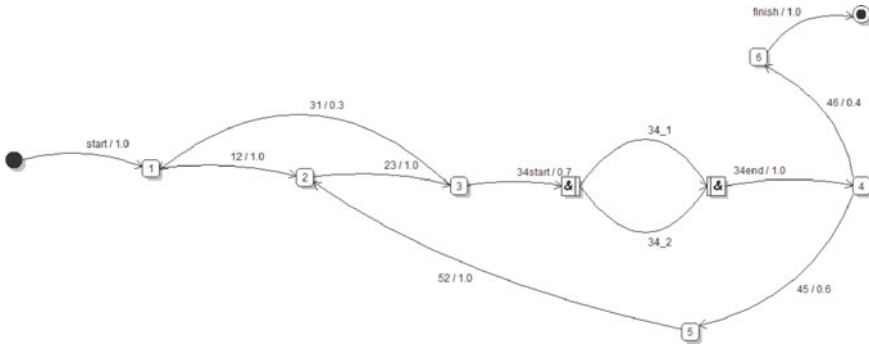
### 2.4 *Finding MGFs of Processes with Concurrent Activities*

Obtaining the MGF of a stochastic process duration with a flowgraph satisfying the “no concurrent activities” condition is a relatively easy task. However, if parallel branches are present, the problem becomes more complicated. In that case, it is necessary to obtain the MGF of each fragment of a flowgraph that contains parallel branches.

Consider the following process (Fig. 3).

Here, the transition from the state 3 to the state 4 can be done only when both activities “34\_1” and “34\_2” are complete that is reflected in the flowgraph by the AND gates.

As a result, the duration distribution of the activity “34” is expressed as the maximum of two random variables which are durations of activities “34\_1” and “34\_2”. The distribution density  $h(x)$  of the maximum of two independent random variables is known to be determined by the following formula:



**Fig. 3** An example of a stochastic process flowgraph with concurrent activities (Graph #2)

$$h(x) = g(x)F(x) + f(x)G(x), \tag{6}$$

where  $f(x)$  and  $g(x)$  are the probability distribution densities and  $F(x)$  and  $G(x)$  are their cumulative distribution functions, respectively.

It should be noted that the expression (6) does not have a general counterpart in the domain of moment generating functions; i.e., one cannot express the activity “34” MGF formula  $M(Y_{34})$  in terms of  $M(Y_{34_1})$  and  $M(Y_{34_2})$ . In order to find an MGF of a maximum of two random variables having only their MGFs, it is needed to acquire their CDFs and PDFs first and only then the MGF of the maximum can be computed. This makes finding the MGF of a stochastic process with concurrent activities a challenging task that requires powerful symbolic calculations in general case—when the distribution densities of random variables are not given and only their MGFs are known. Then, the task is to perform the inverse transform of the MGFs (to guarantee the possibility of inverse transformation, it is proposed to work with the Padé approximants of MGFs) to obtain the distribution densities  $f(t)$  and  $g(t)$ , to integrate these densities so as to find the functions  $F(t)$  and  $G(t)$ , then to apply the formula (6), and finally to compute the MGF of the graph fragment with parallel branches via integral transformation (2) of the  $h(x)$  function.

### 3 Results

The results of the conducted research are grouped in two studies: of stochastic processes with and without concurrent activities.

### 3.1 Study of Stochastic Processes Without Concurrent Activities

A process without concurrent activities is a classical object of study for the flowgraph methodology or the GERT. Consider the process whose graph is shown in Fig. 1. The  $W$ -function from the initial node 1 to the final node 6 of this flowgraph is as follows:

$$W_{16} = \frac{W_{12}W_{26}(1 - W_{34}W_{45}W_{57}W_{73} - W_{45}W_{57}W_{74}) + W_{13}W_{34}W_{45}W_{56}}{1 - W_{11} - W_{45}(W_{34}W_{57}W_{73} - W_{57}W_{74}) + W_{11}W_{34}W_{57}W_{73} + W_{11}W_{57}W_{74}}, \quad (7)$$

The expression of the corresponding MGF depends on the distribution of each activity duration. Let us consider 5 variants of these: the exponential, uniform, triangular, Erlang-3 (Erlang distribution with the shape parameter  $k = 3$ ), and normal. For simplicity, in all cases we take the same expectation  $m$  for the duration of a corresponding activity and the rest of parameters are set relative to this value:

- For the uniform and triangular distributions:  $a = 0.4 m, b = 1.6 m$ .
- For the exponential distribution:  $\lambda = 1/m$ .
- For the Erlang-3 distribution:  $\lambda = 3/m$ .
- For the normal distribution:  $\sigma = 0.2 m$ .

Without loss of generality, when selecting duration distribution laws for activities in a flowgraph let us use a single law for all activities. Thus, activities in a graph differ only in the value of  $m$ , and each graph is studied for 5 different distribution laws of its activities.

After importing simulation results (1000 iterations, unless otherwise stated) from GPSS World, a histogram and empirical cumulative distribution function are constructed and then compared with the corresponding approximations obtained analytically. The comparison makes use of Kolmogorov–Smirnov goodness-of-fit test at the significance level  $\alpha = 0.05$ . Calculations of  $p$ -values were conducted by the MATLAB built-in *kstest* function in accordance with paper [16].

**Results for the Exponential Distribution.** Here and further for results, we state the orders of the Padé approximant and give the resulting values of expectation and variance (or the mean and the unbiased variance estimate for simulation samples). Then, we say if the null hypothesis was accepted by the goodness-of-fit test. For the exponential distribution, the orders are  $a = 6, b = 7$ , and for the moment comparison see Table 1.

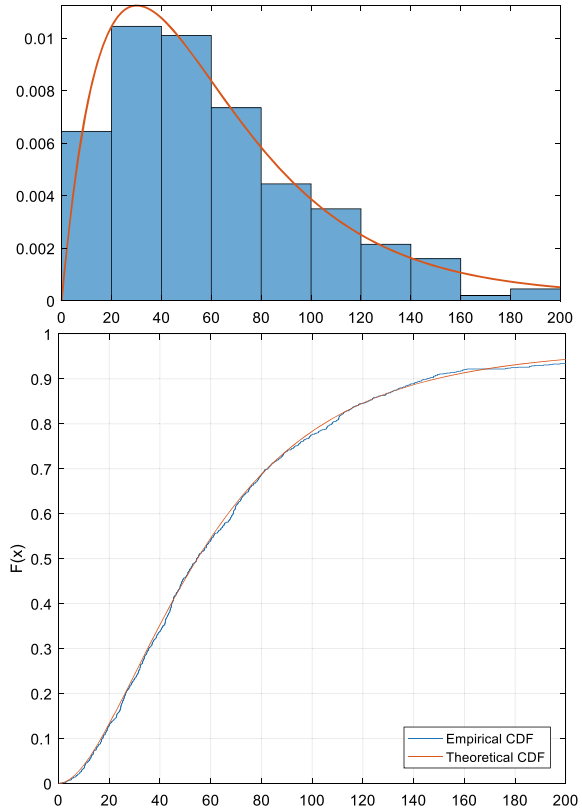
Figure 4 shows analytical PDF and CDF compared with the empirical ones constructed from the sample acquired via simulation in GPSS World. With  $p = 0.75$ , we failed to reject the  $H_0$  hypothesis.

**Results for the Normal Distribution.** The Padé approximant orders:  $a = 6, b = 7$ . Expectation and variance are presented in Table 2. For comparative charts, see Fig. 5.

**Table 1** Moment comparison for MGFs and simulation: the exponential distribution

	Obtained from exact MGF	Obtained from approximated MGF	Obtained from simulation
Expectation	77.863	77.863	79.659
Variance	8002.1	8002.1	8917.7

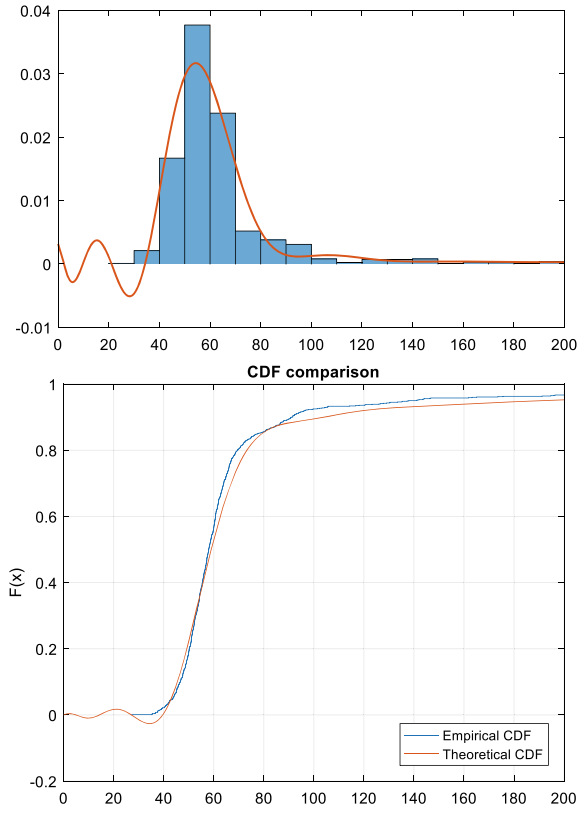
**Fig. 4** Analytical approach versus simulation for the exponential distribution



**Table 2** Moment comparison for MGFs and simulation: the normal distribution

	Obtained from exact MGF	Obtained from approximated MGF	Obtained from simulation
Expectation	77.863	77.863	71.421
Variance	6068.0	6068.0	3512.9

**Fig. 5** Analytical approach ( $a = 6, b = 7$ ) versus simulation for the normal distribution



The hypothesis  $H_0$  was rejected ( $p = 0.0001$ ) for the abovementioned approximant. Then, the experiment was continued with an increased order of Padé approximation to  $a = 12, b = 13$ . Plots of approximated PDF and CDF had become more adequate to the empirical ones (see Fig. 6), and finally we failed to reject the  $H_0$  hypothesis ( $p = 0.07$ ).

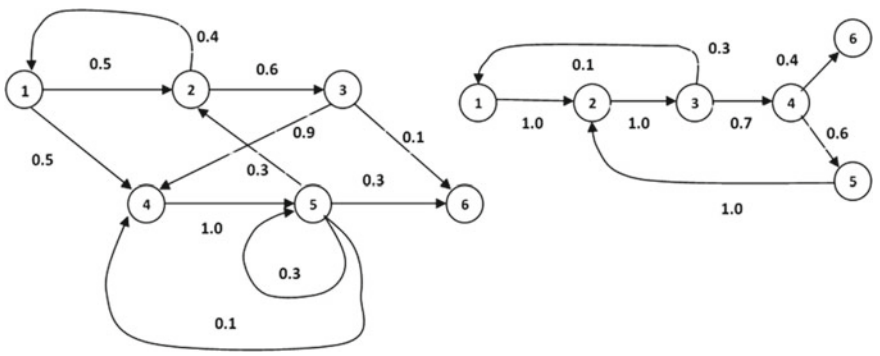
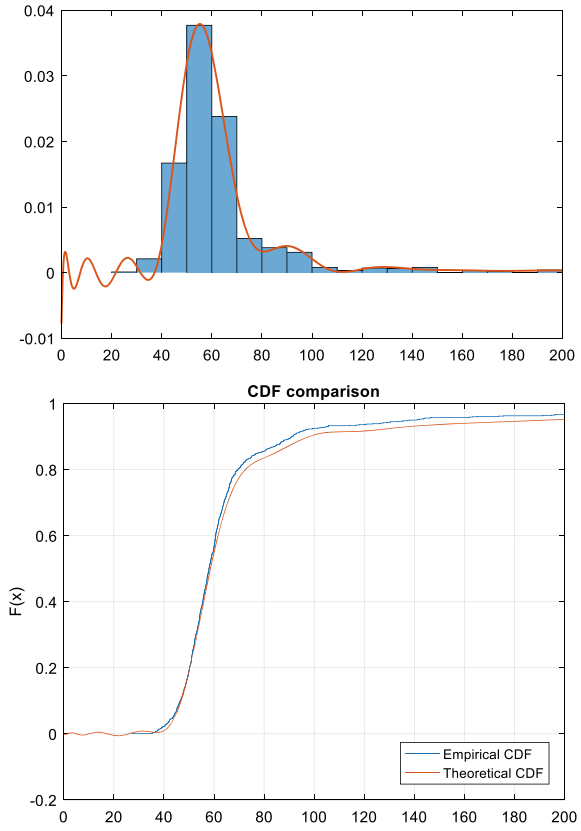
Thereby, increasing the order of the Padé approximant made it possible to achieve better results and match the theoretical functions with the experimental ones. At same time, the values of the moments (see Table 2) had not changed.

**Results for Other Distributions.** Analogous experiments were conducted for the same flowgraph and the other three distributions: uniform, triangular, and Erlang-3. The results of those experiments are fully presented in the Appendix. Here, we report that in all cases we failed to reject the  $H_0$  hypothesis with  $p = 0.99, 0.10,$  and  $0.16,$  respectively.

**Results for Other Flowgraphs.** Similar results were obtained by the authors for the other two flowgraphs that are presented in Fig. 7. In all cases, the comparison of



**Fig. 6** Analytical approach ( $a = 12, b = 13$ ) versus simulation for the normal distribution



**Fig. 7** Other flowgraphs without concurrent activities that were used in this study

approximated distributions and experimental data led us to failing to reject the  $H_0$  hypothesis (at the significance level  $\alpha = 0.05$ ).

### 3.2 Study of Stochastic Processes with Concurrent Activities

To show the applicability of the proposed method to flowgraphs with parallel branches, let us examine Graph #2 (see Fig. 3) which is similar to the flowgraph on the right in Fig. 7 where activity “34” is replaced with a block of two concurrent activities with same distributions and their parameters (let us call them twin activities).

We performed calculations for cases when CDFs and PDFs of the distributions are known in advance as well as the situation when only their MGFs are given. In the first case, expressions for MGFs of maxima can be calculated exactly and we still are able to conduct goodness-of-fit tests, whereas in the second case it is not possible (see Sect. 2.4) and we can assess the results only looking from the qualitative point of view.

#### Results for Concurrent Activities Distributed According to the Triangular Law.

Let each of the parallel arcs of activity “34” be distributed according to the triangular law with the same parameters, and all other distributions of activity durations are also triangular. When simulating in the GPSS World, 5000 iterations were carried out for greater accuracy, and we took  $a = 12$ ,  $b = 13$  for the Padé approximation.

As a result, we obtained correct values of expectation and variance (see Table 3), though calculation of the exact variance took a substantial amount of time (about 30 s on a computer with Intel Core i5-6500 processor). After performing the goodness-of-fit test, we failed to reject the  $H_0$  hypothesis with  $p = 0.27$ .

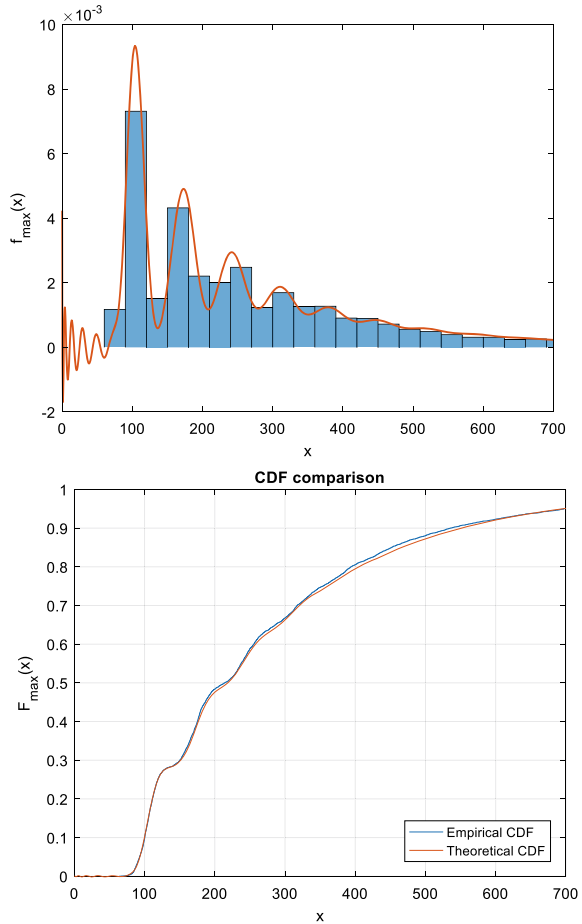
Figure 8 shows the analytical PDF and CDF compared with the empirical ones constructed from the simulation results.

The results of an analogous experiment with the exponential and uniform distribution are presented in the Appendix. Both experiments were successful as we failed to reject the  $H_0$  hypothesis with  $p = 0.53$  and  $p = 0.64$ , respectively. In case of the normal and Erlang-3 distributions, MATLAB was unable to finish all needed symbolic calculations in reasonable time (after 30 min of waiting we stopped the experiment).

**Table 3** Moment comparison for MGFs and simulation: concurrent triangular distributions

	Obtained from exact MGF	Obtained from approximated MGF	Obtained from simulation
Expectation	279.79	279.79	276.81
Variance	43,270	43,270	42,810

**Fig. 8** Analytical approach versus simulation for the triangular distribution (two concurrent twin activities)



**Results for Concurrent Activities with Only MGFs Given.** In this case, the flow-graph shown in Fig. 9 was examined. It has a pair of parallel branches, each of which comprises several activities (enclosed in ovals). The duration of all activities is distributed according to the exponential law, which is well suited to application of the sequence of transformations to construct an MGF of parallel branches that have only MGFs given.

Figure 10 shows the comparison of the approximated analytical PDF and the histogram for that stochastic process obtained from the results of simulation in GPSS World (10,000 iterations).

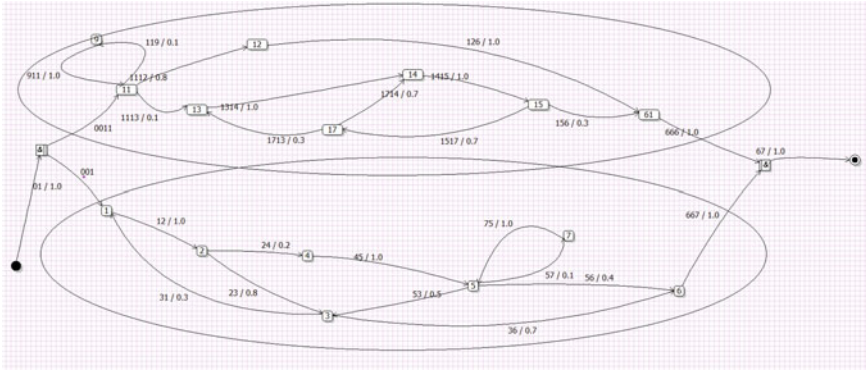


Fig. 9 A flowgraph of a stochastic process with complex parallel branches (Graph #3)

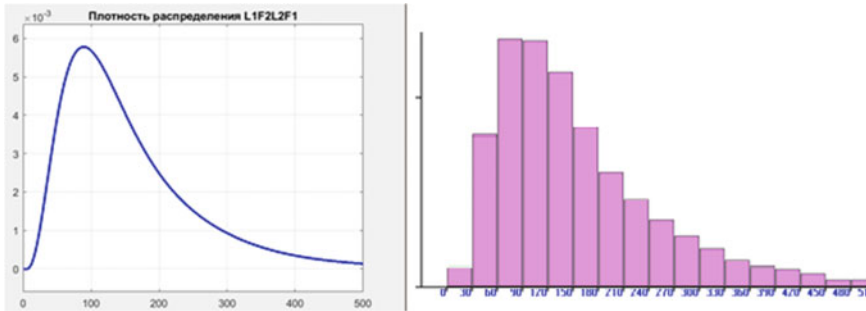


Fig. 10 Comparison of the analytically obtained PDF and the histogram for Graph #3

### 4 Discussion

The achieved results indicate that the suggested analytical method of SPDDLs finding works well for flowgraphs without concurrent activities for various distribution laws, and can also be applied to flowgraphs having parallel branches of different complexity. Experiments with concurrent activities have shown that the capabilities of MATLAB Symbolic Math Toolbox are not always enough for their full support. Nonetheless, the suggested method does not depend on the symbolic computation engine implementation and can be used in software other than MATLAB.

At the same time, during the conducted research it had become clear that the process of selection of optimal orders for Padé approximation requires a separate research because of its importance for the MGF approximant construction. Here, we were using the approach suggested in [6] which is based on MGF's Hankel matrix rank calculation, both in numeric and symbolic calculations that procedure can be imprecise especially for ill-conditioned matrices. Furthermore, although the rule that the numerator order should be equal to the denominator order minus one gives good

results, it was noticed that sometimes a lesser numerator order can lead to better approximations, especially in the aspect of PDF oscillations.

These oscillations are a side issue of approximation and do not affect the calculated moments of SPDDLs but can cause some complications for methods that require monotonicity of approximated CDF. This makes the oscillations undesirable, and further research is needed to develop methods that can solve that issue.

Certainly, apart from Padé approximation, there is a multitude of other methods that can facilitate the inverse Laplace transform of the exact MGF, and the reader can find a thorough review of them in [17, 18]. One of the most successful of these methods [19] even uses computer algebra software but only in the aspect of variable precision arithmetic, so the produced output of the algorithm is still a number for a given value of the independent variable and not an analytical (symbolic) expression that can give a much more profound base for further analysis of the underlying SPDDL. Another promising method that can approximate PDF directly from MGF is known as saddle point approximation and was successfully used for that purpose in [7, 8]. Despite its high accuracy, it has a drawback of computational complexity that limits the applicability of symbolic saddle point approximation to simple flowgraphs as mentioned in [7] (for complex flowgraphs R. Butler used numerical saddle point approximation). At the same time, methods based on Padé approximation were successfully used in [6, 20, 21]. One way or another, in this paper it was shown that Padé approximation gives good results in symbolic form for various flowgraphs and also makes further analysis of SPDDLs feasible for the MATLAB Symbolic Math Toolbox even in complex scenarios such as flowgraphs with concurrent activities.

## 5 Conclusion

Summarizing the results of the research, we can conclude that symbolic computing can be successfully applied to the problems that require finding stochastic process duration distribution laws. The suggested method enables obtaining solutions for such tasks in analytical form and can be used as an alternative to simulation. Moreover, the result that we achieved regarding concurrent activities is not known to be presented in the state of the art. It is clear that there is a potential for further research on the use of symbolic mathematics in the area of stochastic processes but such research should take into consideration the limitations of available software to be most efficient.

## Appendix

**Results for the Uniform Distribution.** The Padé approximant orders:  $a = 6$ ,  $b = 7$ . Expectation and variance are presented in Table 4. For comparative charts, see

**Table 4** Moment comparison for MGFs and simulation: the uniform distribution

	Obtained from exact MGF	Obtained from approximated MGF	Obtained from simulation
Expectation	77.863	77.863	79.102
Variance	6229.2	6229.2	7341.9

Fig. 11. As a result of the goodness-of-fit test, we failed to reject  $H_0$  hypothesis with  $p = 0.99$ .

**Results for the Triangular Distribution.** The Padé approximant orders:  $a = 6, b = 7$ . Moment comparison is given in Table 5. For charts of PDFs and CDFs, see Fig. 12. Here, we failed to reject the  $H_0$  hypothesis with  $p = 0.10$ .

**Results for the Erlang-3 Distribution.** The Padé approximant orders:  $a = 6, b = 7$ . Moment comparison is presented in Table 6. For comparative charts of PDFs and CDFs, see Fig. 13. We failed to reject the  $H_0$  hypothesis with  $p = 0.16$ .

**Results for Concurrent Activities Distributed According to the Uniform Law.** Let each of the parallel arcs of activity “34” from Sect. 3.2 be uniformly distributed with the same parameters, and all other distributions of activity durations are also uniform. When simulating in the GPSS World, 5000 iterations were carried out for greater accuracy, and we took  $a = 12, b = 13$  for the Padé approximation.

As a result, we obtained correct values of expectation and variance (see Table 7), and Fig. 14 shows the analytical PDF and CDF compared with the empirical ones constructed from the results of simulation. We failed to reject the  $H_0$  hypothesis ( $p = 0.64$ ).

**Results for Concurrent Activities Distributed According to the Exponential Law.** The Padé approximant orders:  $a = 12, b = 13$ . Moment comparison is presented in Table 8. For comparative charts of PDFs and CDFs, see Figs. 13 and 15. We failed to reject the  $H_0$  hypothesis with  $p = 0.53$ .

**Table 5** Moment comparison for MGFs and simulation: the triangular distribution

	Obtained from exact MGF	Obtained from approximated MGF	Obtained from simulation
Expectation	77.863	77.863	72.007
Variance	6108.3	6108.3	3607.2

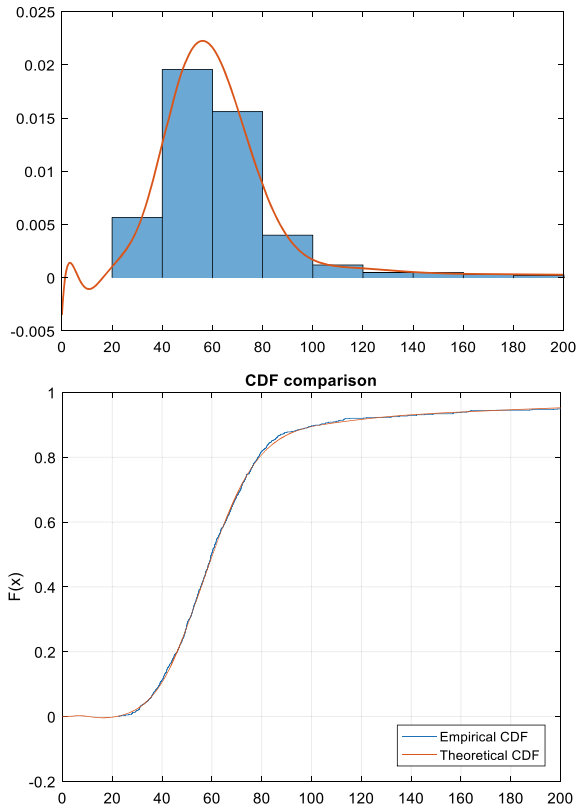
**Table 6** Moment comparison for MGFs and simulation: the Erlang-3 distribution

	Obtained from exact MGF	Obtained from approximated MGF	Obtained from simulation
Expectation	77.863	77.863	72.955
Variance	6659.0	6659.0	5383.1

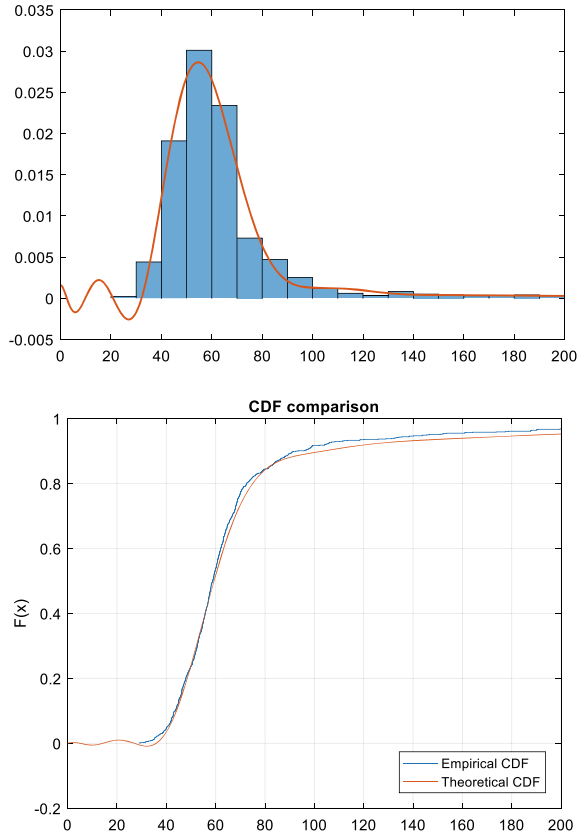
**Table 7** Moment comparison for MGFs and simulation: concurrent uniform distributions

	Obtained from exact MGF	Obtained from approximated MGF	Obtained from simulation
Expectation	284.29	284.29	282.73
Variance	44,953	44,953	45,672

**Fig. 11** Analytical approach versus simulation for the uniform distribution

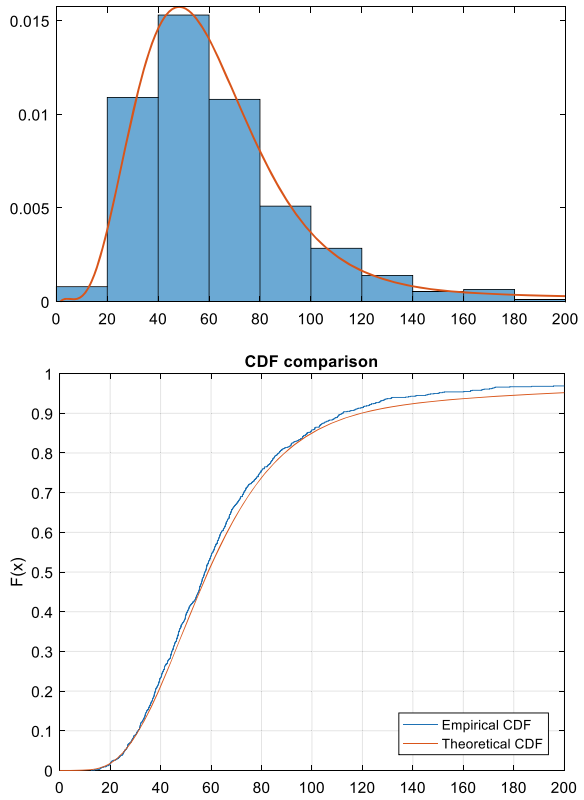


**Fig. 12** Analytical approach versus simulation for the triangular distribution



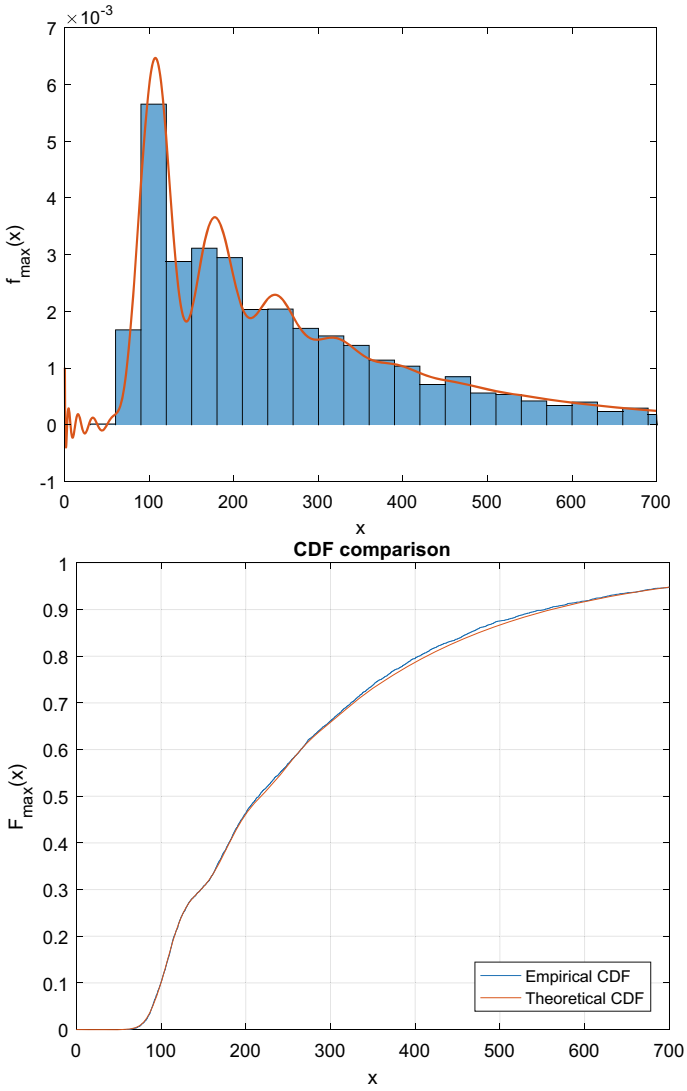


**Fig. 13** Analytical approach versus simulation for the Erlang-3 distribution

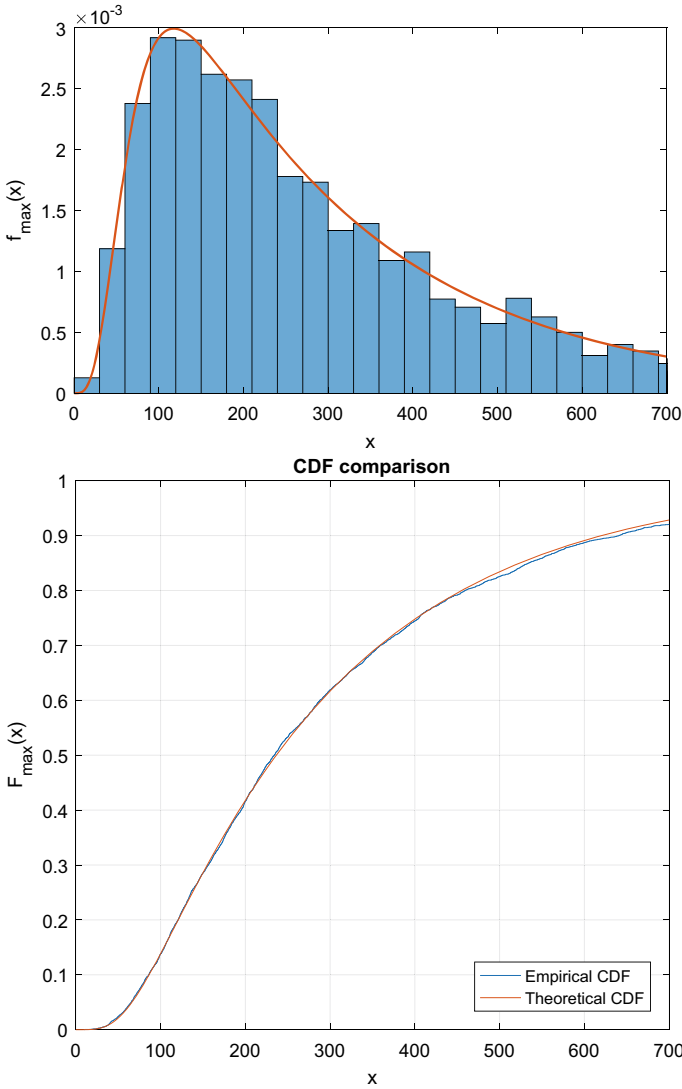


**Table 8** Moment comparison for MGFs and simulation: concurrent exponential distributions

	Obtained from exact MGF	Obtained from approximated MGF	Obtained from simulation
Expectation	306.79	306.79	309.98
Variance	58,175	58,175	60,527



**Fig. 14** Analytical approach versus simulation for the uniform distribution (two concurrent twin activities)



**Fig. 15** Analytical approach versus simulation for the exponential distribution (two concurrent twin activities)

## References

1. Collins, D.H., Warr, R.L., Huzurbazar, A.V.: An introduction to statistical flowgraph models for engineering systems. *J. Risk Reliab.* **227**, 461–470 (2013)
2. Budiana, M.F., Siddig, M.H.M.A.: Flowgraph models and analysis for Markov jump processes. In: 7th European Business Research Conference, Rome, Italy (2016)

3. Huzurbazar, A.V.: Flowgraph Models for Multistate Time-to-Event Data. Wiley, Blackwell (2004). <https://doi.org/10.1002/0471686565>
4. Rubio, G., García-Mora, B., Santamaría, C., Pontones, J.L.: A flowgraph model for bladder carcinoma. *Theor. Biol. Med. Model.* **11** (2014). <https://doi.org/10.1186/1742-4682-11-S1-S3>
5. Phillips, D.T., Garcia-Diaz, A.: *Fundamentals of Network Analysis*. Prentice-Hall, Englewood Cliffs, US (1981)
6. Ren, Y.: The methodology of flowgraph models. Ph.D. thesis, London School of Economics and Political Science, London, England (2011)
7. Butler, R.W.: Reliabilities for feedback systems and their saddlepoint approximation. *Stat. Sci.* **15**, 279–298 (2000). <https://doi.org/10.1214/ss/1009212818>
8. Huzurbazar, A.V.: Flowgraph models for generalized phase type distributions having non-exponential waiting times. *Scand. J. Stat.* **26**, 145–157 (1999). <https://doi.org/10.1111/1467-9469.00142>
9. Liang, J., Chen, Y.Q., Guo, B.Z.: A hybrid symbolic-numerical simulation method for some typical boundary control problems. *Simulation*. **80**, 635–643 (2004). <https://doi.org/10.1177/0037549704050183>
10. Fomicheva, S.G.: Linear complexity of recurrent sequences. *Radiotekhnika* **2**, 72–77 (1997)
11. Fatoorehchi, H., Abolghasemi, H.: An integration-free method for inversion of Laplace transforms: a useful tool for process control analysis and design. *Chem. Eng. Commun.* **203**, 822–830 (2016). <https://doi.org/10.1080/00986445.2015.1107722>
12. Cheever, E.: The Inverse Laplace Transform by Partial Fraction Expansion. <https://lpsa.swarthmore.edu/LaplaceXform/InvLaplace/InvLaplaceXformPFE.html>
13. GPSS World Simulation Environment. <https://www.minutemansoftware.com>
14. Sidnev, A.G.: Business process simulation GPSS-models construction (in Russian). In: Senichenkova, Y. (ed.) *Vychislitelnye, izmeritelnye i upravliaiushchie sistemy*, St. Petersburg, Russia, pp. 41–46 (2006)
15. Violet UML Editor. <https://violet.sourceforge.net>
16. Marsaglia, G., Tsang, W.W., Wang, J.: Evaluating Kolmogorov’s distribution. *J. Stat. Softw.* **8**, 1–4 (2003). <https://doi.org/10.18637/jss.v008.i18>
17. Abate, J., Choudhury, G.L., Whitt, W.: An introduction to numerical transform inversion and its application to probability models. In: Grassmann, W.K. (ed.) *Computational Probability*, pp. 257–323. Springer US, Boston, US (2000). [https://doi.org/10.1007/978-1-4757-4828-4\\_8](https://doi.org/10.1007/978-1-4757-4828-4_8)
18. Cohen, A.M.: *Numerical Methods for Laplace Transform Inversion*. Springer, US (2007)
19. Abate, J., Valkó, P.P.: Multi-precision Laplace transform inversion. *Int. J. Numer. Methods Eng.* **60**, 979–993 (2004). <https://doi.org/10.1002/nme.995>
20. Ismail, M.H., Matalgah, M.M.: On the use of Padé approximation for performance evaluation of maximal ratio combining diversity over Weibull fading channels. *Eurasip J. Wirel. Commun. Netw.* **2006**, 1–7 (2005). <https://doi.org/10.1155/WCN/2006/58501>
21. Boyd, J.P.: Padé approximant algorithm for solving nonlinear ordinary differential equation boundary value problems on an unbounded domain. *Comput. Phys.* **11**, 299 (1997). <https://doi.org/10.1063/1.168606>

# Comparison of the Shape of Digital Models of Pump Components



Evgeniy Ivanov , Aleksandr Zharkovskii , Igor Borshchev ,  
and Arsentiy Klyuyev 

**Abstract** The shape of a flow section determines hydraulic and performance characteristics of dynamic pumps. In order to create digital twins for pumping equipment, a comprehensive project—digital models of pumps—is being implemented in the Hydromechanical Engineering Laboratory of the St. Petersburg Polytechnic University (SPbPU). The engineering issues of the structural optimization are being addressed using a newly developed calculation and design method based on the comparison of component shape. The numerical algorithm is implemented in software codes and is being tested using elements of supercomputing and machine learning. 3D models of pump shafts and impellers with various specific speed rates were used as comparison objects.

**Keywords** Digital · Models · Pumps · Function · Shape · Comparison · Histogram

## 1 Introduction

The issues related to the recognition, comparison, and reconstruction of three-dimensional objects arise in laser scanning, in comparison with three-dimensional objects, at determining the accuracy of component manufacturing, and upon pattern recognition in photography and aerial imagery. In the course of three-dimensional laser scanning, a point cloud is formed, which is characterized by a large body of data. It is preferable to compare 3D objects pointwise using a regular grid. This can be achieved by means of interpolation. The grid should be fine enough to depict the surface properly. The number of points in such a grid is significantly larger as compared to the grid of the scanned surface shape. This results in unacceptable time consumption required to solve the problem. One of the approaches to recognize the obtained data is the use of graph spectra. A graph is described using an adjacency matrix, which for an undirected graph is a symmetric matrix with its elements being equal to the number of edges connecting the vertices of the graph. A spectrum is a

---

E. Ivanov (✉) · A. Zharkovskii · I. Borshchev · A. Klyuyev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [ivanov\\_ea@spbstu.ru](mailto:ivanov_ea@spbstu.ru)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021  
N. Voinov et al. (eds.), *Proceedings of International Scientific Conference  
on Telecommunications, Computing and Control*, Smart Innovation, Systems  
and Technologies 220, [https://doi.org/10.1007/978-981-33-6632-9\\_8](https://doi.org/10.1007/978-981-33-6632-9_8)

set of eigenvalues arranged in descending or ascending order [1]. The studies [2–4] have developed algorithms for the comparison and analysis of surfaces that do not require the recalculation of grids into a general regular one. The proposed approach is based on the calculation and fitting of piecewise linear models of the original surfaces on Delaunay triangulations, the labor efforts of which are  $O(N \cdot \lg(N))$ . When comparing three-dimensional areas, their shape descriptors are also being compared: These are obtained using shape functions that reflect connections and relations between points on a figure surface [5–11].

## 2 Algorithm for the Comparison of 3D Models of Pump Components Using Shape Functions

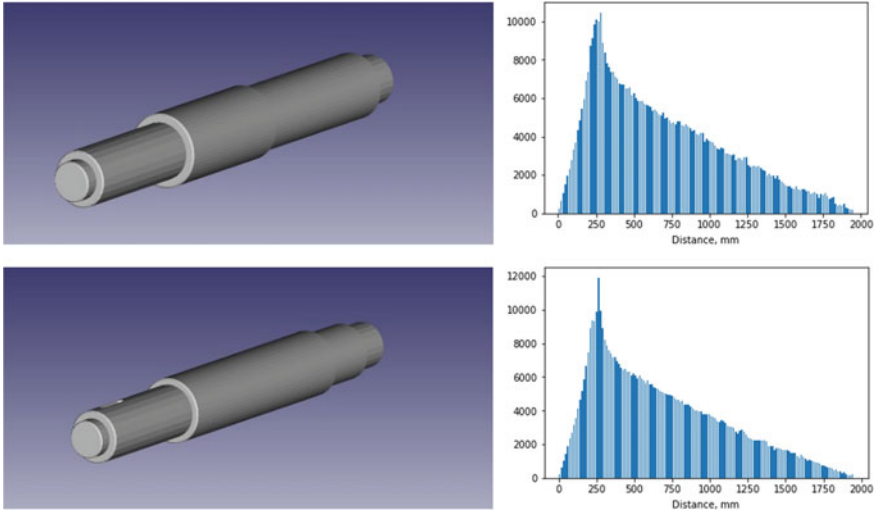
The 3D models are compared by representing the shape of an object using the probability distribution of the values of the so-called shape function, which measures the global geometric properties of the object. The following shape functions have been investigated and clarified in the literature [1–3]:

- A3: Measures the angle between three random points selected on the model surface.
- D1: Measures the distance from a fixed point to a randomly selected point on the model surface.
- D2: Measures the distance between two randomly selected points on the surface of 3D model.
- D3: The values of this shape function are the square root of the area of a triangle, the vertices of which are three randomly selected points on the surface of the 3D model.
- D4: The value of this function is the cubic root of the volume of a prism formed by four randomly selected points on the model surface.

The distribution of values of these functions can be designated as the shape distribution, which, for these functions, has the following properties:

- Invariance: Similar to shape functions, the distribution will not depend on the position of the three-dimensional model or its orientation relative to the absolute coordinate system.
- Robustness: The random selection of points ensures that the shape distribution is insensitive to small changes in the 3D model.
- Effectiveness: Shape distribution building is an operation that is performed once for each model and is generally quick and efficient. For example, the complexity of sampling of a point set equal to  $m$  for D2 will be  $O(m \cdot \lg(N))$ , where  $N$  is the number of triangles in the model.

The developed algorithm is based on the shape distribution D2, which according to the study [13] results are preferable in terms of stability and descriptive capability.



**Fig. 1** Examples of 3D models of shafts and their histograms

The essence of the proposed method is to bring the problem of comparing 3D models to the problem of comparing the probability distribution of a random variable. The random value is the distance between a pair of randomly selected points on the model surface belonging to the triangles that describe this model.

The shape distribution (distribution of the values of shape functions) is represented as a histogram (Fig. 1), which is the distribution of distances between pairs of points in the model showing how many distances between points fall in the intervals of fixed length.

### 3 Point Set Selection

The number of selected points is given by a fixed number  $m$ , sufficient for an accurate description of the model and, at the same time, requiring less calculations. The number of pairs under examination  $M$  is determined by the equation:

$$M = \frac{m!}{(m - 2)!2!}. \tag{1}$$

Algorithmic complexity of histogramming  $O(m^2)$  is a function of the number of points. From this perspective,  $m = 1024$  was chosen as a compromise between accuracy and speed.

This step of the algorithm determines the area of triangles, from which a point is selected through compiling an array of partial sums of areas of triangles (CA). The

area of the triangles is calculated using the Heron formula:

$$S_T = \sqrt{s(s-a)(s-b)(s-c)}, \quad (2)$$

where  $s$  is the semi-perimeter of the triangle, and  $a, b, c$  are the lateral lengths of triangle.

The indices of the CA array are the indices of triangles in the model, while the value is the sum of the areas of all previous triangles:

$$CA_i = CA_{i-1} + S_{T_i}; CA_0 = S_{T_0}, \quad (3)$$

where  $S_{T_i}$  is the area of the  $i$ -th triangle.

A random number is selected from the interval  $[0, S]$ , where  $S$  is the surface area of the model; the corresponding triangle is selected in the CA array using the binary search. The point  $p$  lying inside the triangle is selected using two random numbers,  $r_1, r_2 \in [0, 1]$ .

The coordinates of the point  $p$  are calculated by the formula:

$$p = (1 - \sqrt{r_1})A + \sqrt{r_1}(1 - r_2)B + \sqrt{r_1}r_2C, \quad (4)$$

where  $A, B, C$  are the coordinates of the triangle vertices.

It is a random choice of a point inside the triangle that provides the algorithm with the immunity to the model tessellation.

Thus, a set of points is formed for further comparison; the complexity of this algorithm is  $O(m \cdot \log(N))$ , where  $N$  is the number of triangles in the 3D model.

## 4 Histogram Generation

The distance between points is understood as the Euclidean metric [12]:

$$\rho(p_1, p_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}. \quad (5)$$

The resulting distances are combined into a histogram. The width of the histogram intervals  $h$  is chosen as the distance average divided by  $k$ :

$$h = \frac{\sum d_i}{M \cdot k}. \quad (6)$$

The divisor  $k$  determines the details of the histogram. Based on the research results, the value  $k = 50$  was chosen. Thus, the number of columns in the histogram  $n$  is determined by the formula:



$$n = \frac{\max(d_1, d_2, \dots) - \min(d_1, d_2, \dots)}{h}. \quad (7)$$

It is the very fact that the histogram interval spacing is determined relative to the average that enables the scale invariance.

## 5 Comparison of Histograms

The comparison metric is based on the Minkowski metric [13, 14]:

$$L_1(h, g) = \sum_{i=0}^n |h_i - g_i|, \quad (8)$$

where  $h_i, g_i$  are the heights of the  $i$ th column of the histogram of the first and second models, respectively.

Since the number of histogram columns can be different for each model, and when  $L_1$  is used two similar models with a large number of columns will have a greater distance than two similar models having a small number of histogram intervals, another metric is introduced [2]:

$$D(h, g) = \frac{L_1(h, g)}{n}, \quad (9)$$

where  $n$  is the number of histogram columns. Thus, this metric gives the difference between the numbers of distances included in the corresponding interval on average.

Figure 1 shows examples of models and histograms:

Metric  $D$  for the models presented:

$$D(h, g) = 394.48.$$

As we can see, this characteristic is not very indicative, so we will introduce another metric:

$$D_M(h, g) = (L_1(h, g)/M) \cdot 100. \quad (10)$$

The essence of this metric is the percentage difference between the shapes of models. For the models presented:

$$D_M(h, g) = 10.$$

Based on the developed algorithm, the software implementation of the method of shape comparison of three-dimensional models with complex geometry was performed.

## 6 Testing the Software Implementation of the 3D Model Comparison Algorithm

The algorithm can be used for comparison of elements of different types of pumps in cases, when it is necessary to determine the exact degree of their coincidence or difference [15, 16]. A database of three-dimensional models of pump components was created to assess the performance of the developed algorithm for the comparison of three-dimensional models. The total number of 3D models at the time of testing was 50.

The models were classified into the following types:

- shafts (Fig. 1);
- axial pump impellers (Table 1);
- centrifugal pump impellers (Fig. 2).

Verification of the developed software implementation of the algorithm was carried out by comparing models with different degrees of similarity. In order to assess the software ability to select models with minor geometry differences, the database also includes models that slightly differ from each other by several geometric features.

As a result of the software application for the AP-3 model from the first group, named axial pump impellers [17–19], the models similar thereto were arranged in the order as shown in Table 1.

Other models significantly differ from the AP-3 model, which is indicated by the developed algorithm.

The second group (Fig. 2) is centrifugal pump impellers [20], with a specific speed in the range of  $n_s$  from 105 to 324.

The result of the software application for comparison of models from the database against the model with  $n_s = 114$  is represented in Table 2.

Table 2 indicates the algorithm outputs: The models with  $n_s = 124$  and  $n_s = 105$  turned out to be the closest in geometric shape to the model with  $n_s = 114$ .

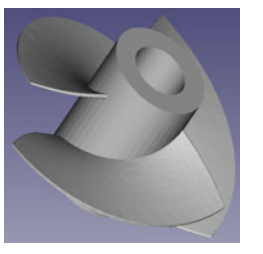
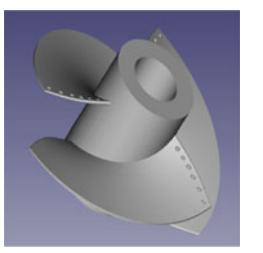
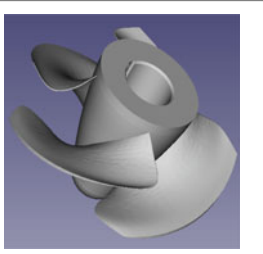
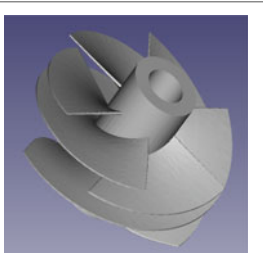

The models belong to groups designated as follows:

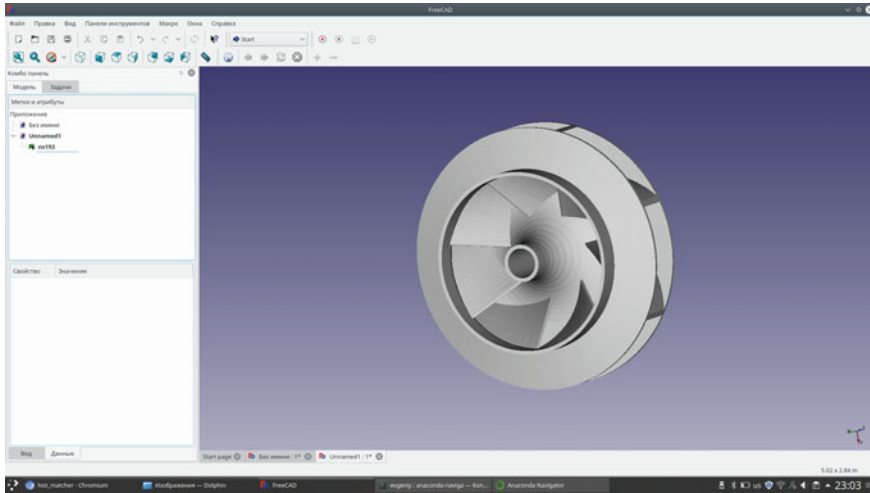
1. axial impellers;
2. centrifugal impellers.

## 7 Conclusions

1. A method and algorithm for the comparison of 3D models of pump components with complex geometric shape have been developed.
2. The developed algorithm has been implemented in software.
3. 3D models of pump components have been initially uploaded into the database.

**Table 1** Results of the AP-3 model comparison with models from the database

3D model	Visualization	DM(h, g)
AP-3 with three blades		-
AP-3 with perforated blades		1.96
AP-4 with 4 blades, tapered bushing, modified blade shape		4.25
AP-3 with 3 additional blades of the second tier		4.92
AP-5 with 5 blades, increased hub-tip ratio		5.77



**Fig. 2** Impeller of a centrifugal pump

**Table 2** Result of comparing the model with  $n_s = 114$  with other models from the database

№	Model group	Model name	Degree of histogram difference
1	2	$n_s = 114$	–
2	2	$n_s = 124$	1.91
3	2	$n_s = 105$	2.29
4	2	$n_s = 136$	2.59
5	2	$n_s = 150$	3.10
6	2	$n_s = 168$	3.83
7	2	$n_s = 180$	5.09
8	2	$n_s = 193$	5.75
9	2	$n_s = 208$	6.24
10	2	$n_s = 228$	8.22
11	2	$n_s = 252$	8.75
12	1	AP OD-2	11.66
13	2	$n_s = 283$	11.77
14	2	$n_s = 324$	13.17
15	1	AP-4	15.93
16	1	AP-3(6)	20.03

4. The results of the software implementation testing have testified that the developed algorithm allows the detection of discrepancies in 3D models, including those having minor shape variations.
5. The developed shape evaluation method makes it possible to create a database of the best samples of pump flow sections and an intelligent matching system to determine the similarity of 3D models.

## References

1. Tuzhilin, A.Yu.: Recognition and reconstruction of 3D objects by satellite images based on graph spectra. *Fundamental Res.* **2–17**, 3727–3732 (2015)
2. Dyshkant, N.F., Mestetsky, L.M.: Comparison of 3D portraits upon face recognition. In: All-Russian Conference Reports: Mathematical Methods of Pattern Recognition—13, pp. 314–316. MAKS Press, Moscow (2007)
3. Dyshkant, N.F.: Comparison of surfaces defined on unstructured grids and grids with different density. In: Reports of the 8th International Conference: Intellectualization of Information Processing (IOI-2010), pp. 339–342. MAKS Press, Moscow (2010)
4. Dyshkant, N.F.: Efficient Algorithms for the Comparison of Surfaces Defined by Point Clouds. Ph.D. Thesis in Physical and Mathematical Sciences Moscow: M. V. Lomonosov Moscow State University, 139 pages (2007)
5. Osada, R.: Shape Distributions. *ACM Trans. Graph.* **214**, 807–832 (2002) (R. Osada, T. Funkhouser, B. Chazelle)
6. Ohbuchi, R.: Shape-similarity search of 3D models by using enhanced shape functions. *Int. J. Comput. Appl. Technol. (IJCAT)* **23**(2/3/4/), 70–85 (2005) (Ohbuchi R., T. Minamitani, T. Takei)
7. Osada, R.: Matching 3D Models with Shape Distributions. R. Osada, T. Funkhouser, B. Chazelle, D. Dobkin. *SMI 2001 International Conference on*, pp. 154–166 (2001)
8. Rubner, Y.: Empirical evaluation of dissimilarity measures for color and texture. In: Rubner, Y., Puzicha, J., Tomasi, C., Buchman, J. (eds.) *IEEE International Conference on Computer Vision*, pp. 1165–1173 (1999)
9. Grudinin, S.N.: Comparison of tree-dimensional objects. Criteria for assessing similarities. *Young Sci.* **1**(5) (28), 42–44 (2011)
10. Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D.: Matching 3D models with shape distributions. In: *International Conference on Shape Modeling and Applications. ACM SIGGRAPH, the Computer Graphics Society and EUROGRAPHICS*, IEEE Computer Society Press, Genova, Italy, May 7–11, pp. 154–166 (2001)
11. Lapadat, D., Sieger, L., Regli, W.: Using shape distributions to compare solid models. In: *7-th ACM/SIGGRAPH Symposium on Solid Modeling and Applications*, pp. 273–280 (2002)
12. Srivastava, A., Klassen, E., Joshi, S.H., Jermyn, I.H.: Shape analysis of elastic curves in Euclidean spaces. *IEEE J. Sel. Areas Commun.* **10**, 391–400 (1992)
13. Subramaniam, J., Yagnanarayanan, K., Natraj, I., Karthik, R.: Developing an engineering shape benchmark for CAD models. *Comput. Aided Des.* **38**, 939–953 (2006)
14. Vaidya, P.M.: An  $O(n \log n)$  algorithm for the all-nearest-neighbors. *Problem Discrete Comput. Geometry* **4**, 101–115 (1989)
15. Cordeiro De Amorim, R., Mirkin, B.: Minkowski metric, feature weighting and anomalous cluster initializing in K-Means clustering. *Pattern Recogn.* **45**, 1061–1075 (2012)
16. Kondyurin, Yu., Shcherba, V.E., Shalai, V.V., Noskov, A.S., Khait, A.V.: Analysis and optimization of basic geometric parameters of annular slot seal made in the form of hydrodiode. *Chem. Petroleum Eng.* **52**(4), 280–289 (2016)

17. Pochyly, M., Haluza, S., Fialova, L., Dobšakova, A.V., Volkov, A.G., Parygin, A.V., Naumov, A.A., Vihlyancev, A.A.: Druzhinin. *Primenenie geterogennyh lopastnyh sistem - put' k povysheniyu energoeffektivnosti centrobeznyh energeticheskikh nasosov* «Teploenergetika (11), 13–22 (2017)
18. Svoboda, D.G., Zharkovskii, A.A., Ivanov, E.A., Shchutskii, SYu., Dyagilev, PYu.: High-efficiency axial pumps for reactor use. *Russ. Eng. Res.* **39**, 556–560 (2019)
19. Svoboda, D., Ivanov, E., Zharkovskii, A., Borshchev, I.: In *E3S Web of Conferences (EDP Sciences, 2019)*, vol. 121
20. Zharkovskij, A.A., Zimmitskij, A.V., Shkarbul', S.N.: Experimental and theoretical investigations of flow in blade casings of centrifugal pumps. *Gidrotekhnicheskoe Stroitel'stvo*, 28–29 (1994)

# Parametric Oscillations of Viscoelastic Orthotropic Rectangular Plates of Variable Thickness



Rustamkhan Abdikarimov , Bakhodir Normuminov ,  
Dadakhon Khodzhaev , and Davron Yulchiyev 

**Abstract** A mathematical model of the problem of parametric vibrations of viscoelastic rectangular orthotropic plates of variable thickness under periodic load is given in the paper on the basis of the Kirchhoff–Love hypothesis in a geometrically nonlinear statement. The mathematical model of this problem is constructed taking into account the propagation of elastic waves. Using the Bubnov–Galerkin method, based on a polynomial approximation of deflection and displacements, the problem is reduced to solving systems of nonlinear integro-differential equations with variable coefficients. The effects of viscoelastic properties of the material and changes in thickness on the oscillation process are studied.

**Keywords** Rectangular plate · Variable thickness · Viscoelasticity · Orthotropy · Parametric vibrations · Mathematical model · Relaxation kernel · Integro-differential equation · Numerical method

## 1 Introduction

Plates and shells of variable thickness are widely introduced in various fields of technology. This is primarily due to the requirements for strength, durability, and design of thin-walled elements of modern structures. Along with thin-walled structural elements from traditional metal materials, structures made of composite materials are widely used; this leads to the need to consider structures with homogeneous and inhomogeneous material properties. The study of problems for plates and shells of variable thickness is a very difficult task and sometimes faces insurmountable difficulties. On the one hand, this is connected with the solution of rather cumbersome equations, which are obtained in mathematical modeling, to reflect the real

---

R. Abdikarimov  
60A A.Temur Str., Tashkent Institute of Finance, Tashkent 100000, Uzbekistan

B. Normuminov · D. Khodzhaev (✉) · D. Yulchiyev  
Tashkent Institute of Irrigation and Agricultural Mechanization Engineers, 39 Kari Niyazov str.,  
Tashkent 100000, Uzbekistan  
e-mail: [dhodjaev@mail.ru](mailto:dhodjaev@mail.ru)

mechanical essence of the process of this problem. And on the other hand, it is connected with certain computational difficulties, i.e., the lack of suitable universal numerical methods for solving the obtained equations, and as a result, unified computational algorithms. The widespread use of personal computers and software products for solving similar problems of the theory of plates and shells of variable rigidity contributes to the increasing use of numerical analysis methods.

A number of papers [1, 2] are devoted to studying the behavior of plates and shells of constant thickness under dynamic loads in an elastic statement, and there a detailed review of the results of these studies can be found.

In [3] derives accurately, for the first time, the nonlinear damping from a fractional viscoelastic standard solid model by introducing geometric nonlinearity in it.

Theoretical and experimental nonlinear vibrations of thin rectangular plates and curved panels subjected to out-of-plane harmonic excitation are investigated in [4]. Experiments have been performed on isotropic and laminated sandwich plates and panels with supported and free boundary conditions.

Nonlinear vibrations of viscoelastic thin rectangular plates subjected to normal harmonic excitation in the spectral neighborhood of the lowest resonances are investigated in [5].

A review of publications devoted to the study of the behavior of plates and shells of smoothly variable thickness shows that at present, the behavior of such structural elements is insufficiently studied taking into account all the noted significant factors [6–11].

Studies of parametric vibrations of thin-walled structures have become a separate area of research in the mechanics of a deformable rigid body. They are widely applied to various mechanical systems, in particular to plates and shells.

In [12], a numerical–analytical method was proposed for studying parametric oscillations of plates under the action of static and periodic loads.

In [13–16], the results of a study of dynamic stability of various types of plates subjected to harmonic loading with and without nonlinearity are presented.

An analysis of the available literature showed [17–19] that there are almost no publications devoted to the study of nonlinear vibrations and dynamic stability of thin-walled structures such as viscoelastic plates and shells of variable thickness. In this paper, nonlinear parametric oscillations of viscoelastic orthotropic rectangular plates of variable thickness are numerically investigated. Based on the algorithm for the problem solution, a program was compiled in the Delphi programming environment.

## 2 Materials and Methods

Consider a viscoelastic orthotropic rectangular plate of variable thickness  $h = h(x, y)$  with sides  $a$  and  $b$  under the action of axial dynamic loads. Let the plate undergo dynamic loading along side  $a$  with a periodic load  $P(t) = P_0 + P_1 \cos(\Theta t)$  ( $P_0, P_1 = \text{const}$ ,  $\Theta$  is the frequency of external periodic load). A mathematical model



of the problem is constructed in a geometrically nonlinear statement according to the classical Kirchhoff–Love theory. We assume that the plate has initial deflections  $w_0 = w_0(x, y)$ .

In this case, physical dependence between stresses  $\sigma_x, \sigma_y, \tau_{xy}$  and strains  $\varepsilon_x, \varepsilon_y, \gamma_{xy}$  is taken in the form [2, 20]:

$$\begin{aligned}\sigma_x &= B_{11}(1 - \Gamma_{11}^*)\varepsilon_x + B_{12}(1 - \Gamma_{12}^*)\varepsilon_y, \quad (x \leftrightarrow y, 1 \leftrightarrow 2), \\ \tau_{xy} &= 2B(1 - \Gamma^*)\gamma_{xy},\end{aligned}\quad (1)$$

where  $\Gamma^*, \Gamma_{ij}^*$  are the integral operators with the relaxation kernels  $\Gamma(t)$  and  $\Gamma_{ij}(t)$ , respectively:

$$\Gamma^*\phi = \int_0^t \Gamma(t - \tau)\phi(\tau)d\tau, \quad \Gamma_{ij}^*\phi = \int_0^t \Gamma_{ij}(t - \tau)\phi(\tau)d\tau, \quad i, j = 1, 2,$$

$$B_{11} = \frac{E_1}{1 - \mu_1\mu_2}, \quad B_{22} = \frac{E_2}{1 - \mu_1\mu_2}, \quad B_{12} = B_{21} = \mu_1 B_{22} = \mu_2 B_{11}, \quad B = \frac{G}{2},$$

$E_1, E_2$  are the elastic moduli in the direction of the axes  $x$  and  $y$ ;  $G$  is the shear modulus;  $\mu_1, \mu_2$  are Poisson's ratios; here and hereafter, the symbol  $(x \leftrightarrow y, 1 \leftrightarrow 2)$  indicates that the remaining relations are obtained by circular substitution of indices.

The relationship between strains in the middle surface  $\varepsilon_x, \varepsilon_y, \gamma_{xy}$  and displacements  $u, v, w$  in  $x, y, z$  directions, taking into account initial irregularities, is taken in the form [2]:

$$\begin{aligned}\varepsilon_x &= \frac{\partial u}{\partial x} + \frac{1}{2} \left[ \left( \frac{\partial w}{\partial x} \right)^2 - \left( \frac{\partial w_0}{\partial x} \right)^2 \right], \\ \varepsilon_y &= \frac{\partial v}{\partial y} + \frac{1}{2} \left[ \left( \frac{\partial w}{\partial y} \right)^2 - \left( \frac{\partial w_0}{\partial y} \right)^2 \right], \\ \gamma_{xy} &= \frac{\partial u}{\partial y} + \frac{\partial v}{\partial x} + \frac{\partial w}{\partial x} \frac{\partial w}{\partial y} - \frac{\partial w_0}{\partial x} \frac{\partial w_0}{\partial y}\end{aligned}\quad (2)$$

Bending  $M_x, M_y$  and torques  $H$  with (2) have the form [2, 20]:

$$\begin{aligned}M_x &= -\frac{h^3}{12} \left[ B_{11}(1 - \Gamma_{11}^*) \frac{\partial^2(w - w_0)}{\partial x^2} + B_{12}(1 - \Gamma_{12}^*) \frac{\partial^2(w - w_0)}{\partial y^2} \right], \\ &\quad (x \leftrightarrow y, 1 \leftrightarrow 2), \\ H &= -\frac{Bh^3}{3} (1 - \Gamma^*) \frac{\partial^2(w - w_0)}{\partial x \partial y}.\end{aligned}\quad (3)$$

Substituting (1) and (3) into equation of motion [2]

$$\begin{aligned}
\frac{\partial N_x}{\partial x} + \frac{\partial N_{xy}}{\partial y} + p_x - \rho h \frac{\partial^2 u}{\partial t^2} = 0, \quad \frac{\partial N_{xy}}{\partial x} + \frac{\partial N_y}{\partial y} + p_y - \rho h \frac{\partial^2 v}{\partial t^2} = 0 \\
\frac{\partial M_x}{\partial x^2} + \frac{\partial^2 M_y}{\partial y^2} + 2 \frac{\partial^2 H}{\partial x \partial y} + \frac{\partial}{\partial x} \left( N_x \frac{\partial w}{\partial x} + N_{xy} \frac{\partial w}{\partial y} \right) \\
+ \frac{\partial}{\partial y} \left( N_{xy} \frac{\partial w}{\partial x} + N_y \frac{\partial w}{\partial y} \right) + P_x(t) \frac{\partial^2 w}{\partial x^2} + q - \rho h \frac{\partial^2 w}{\partial t^2} = 0
\end{aligned} \tag{4}$$

we get a system of integro-differential equations in partial derivatives of the form:

$$\begin{aligned}
& h \left[ B_{11}(1 - \Gamma_{11}^*) \frac{\partial \varepsilon_x}{\partial x} + B_{12}(1 - \Gamma_{12}^*) \frac{\partial \varepsilon_y}{\partial x} + 2B(1 - \Gamma^*) \frac{\partial \varepsilon_{xy}}{\partial y} \right] \\
& + \frac{\partial h}{\partial x} [B_{11}(1 - \Gamma_{11}^*) \varepsilon_x + B_{12}(1 - \Gamma_{12}^*) \varepsilon_y] + 2B \frac{\partial h}{\partial y} (1 - \Gamma^*) \varepsilon_{xy} - \rho h \frac{\partial^2 u}{\partial t^2} = 0, \\
& h \left[ B_{22}(1 - \Gamma_{22}^*) \frac{\partial \varepsilon_y}{\partial y} + B_{21}(1 - \Gamma_{21}^*) \frac{\partial \varepsilon_x}{\partial y} + 2B(1 - \Gamma^*) \frac{\partial \varepsilon_{xy}}{\partial x} \right] \\
& + 2B \frac{\partial h}{\partial x} (1 - \Gamma^*) \varepsilon_{xy} + \frac{\partial h}{\partial y} [B_{21}(1 - \Gamma_{21}^*) \varepsilon_x + B_{22}(1 - \Gamma_{22}^*) \varepsilon_y] - \rho h \frac{\partial^2 v}{\partial t^2} = 0, \\
& D \left[ B_{11}(1 - \Gamma_{11}^*) \frac{\partial^4 (w - w_0)}{\partial x^4} + (8B(1 - \Gamma^*) + B_{12}(1 - \Gamma_{12}^*) + B_{21}(1 - \Gamma_{21}^*)) \right. \\
& \quad \left. \frac{\partial^4 (w - w_0)}{\partial x^2 \partial y^2} + B_{22}(1 - \Gamma_{22}^*) \frac{\partial^4 (w - w_0)}{\partial y^4} \right] \\
& + \frac{\partial^2 D}{\partial x^2} \left( B_{11}(1 - \Gamma_{11}^*) \frac{\partial^2 (w - w_0)}{\partial x^2} + B_{12}(1 - \Gamma_{12}^*) \frac{\partial^2 (w - w_0)}{\partial y^2} \right) \\
& + 2 \frac{\partial D}{\partial x} \left[ B_{11}(1 - \Gamma_{11}^*) \frac{\partial^3 (w - w_0)}{\partial x^3} + (B_{12}(1 - \Gamma_{12}^*) + 4B(1 - \Gamma^*)) \frac{\partial^3 (w - w_0)}{\partial x \partial y^2} \right] \\
& + 2 \frac{\partial D}{\partial y} \left[ B_{22}(1 - \Gamma_{22}^*) \frac{\partial^3 (w - w_0)}{\partial y^3} + (B_{21}(1 - \Gamma_{21}^*) + 4B(1 - \Gamma^*)) \frac{\partial^3 (w - w_0)}{\partial x^2 \partial y} \right] \\
& + \frac{\partial^2 D}{\partial y^2} \left( B_{22}(1 - \Gamma_{22}^*) \frac{\partial^2 (w - w_0)}{\partial y^2} + B_{21}(1 - \Gamma_{21}^*) \frac{\partial^2 (w - w_0)}{\partial x^2} \right) \\
& + 8 \frac{\partial^2 D}{\partial x \partial y} B(1 - \Gamma^*) \frac{\partial^2 (w - w_0)}{\partial x \partial y} - \frac{\partial w}{\partial x} \left\{ h \left[ B_{11}(1 - \Gamma_{11}^*) \frac{\partial \varepsilon_x}{\partial x} + B_{12}(1 - \Gamma_{12}^*) \frac{\partial \varepsilon_y}{\partial x} \right. \right. \\
& \left. \left. + 2B(1 - \Gamma^*) \frac{\partial \varepsilon_{xy}}{\partial y} \right] + \frac{\partial h}{\partial x} [B_{11}(1 - \Gamma_{11}^*) \varepsilon_x + B_{12}(1 - \Gamma_{12}^*) \varepsilon_y] + 2B \frac{\partial h}{\partial y} (1 - \Gamma^*) \varepsilon_{xy} \right\} \\
& - h \frac{\partial^2 w}{\partial x^2} [B_{11}(1 - \Gamma_{11}^*) \varepsilon_x + B_{12}(1 - \Gamma_{12}^*) \varepsilon_y] - \frac{\partial w}{\partial y} \left\{ h \left[ B_{22}(1 - \Gamma_{22}^*) \frac{\partial \varepsilon_y}{\partial y} \right. \right. \\
& \left. \left. + B_{21}(1 - \Gamma_{21}^*) \frac{\partial \varepsilon_x}{\partial y} + 2B(1 - \Gamma^*) \frac{\partial \varepsilon_{xy}}{\partial x} \right] + 2B \frac{\partial h}{\partial x} (1 - \Gamma^*) \varepsilon_{xy} \right\} \\
& + \frac{\partial h}{\partial y} [B_{21}(1 - \Gamma_{21}^*) \varepsilon_x + B_{22}(1 - \Gamma_{22}^*) \varepsilon_y] \left\} - h \frac{\partial^2 w}{\partial y^2} [B_{21}(1 - \Gamma_{21}^*) \varepsilon_x
\end{aligned}$$

$$+B_{22}(1 - \Gamma_{22}^*)\varepsilon_y] - 4h \frac{\partial^2 w}{\partial x \partial y} B(1 - \Gamma^*)\varepsilon_{xy} + P_x(t) \frac{\partial^2 w}{\partial x^2} + \rho h \frac{\partial^2 w}{\partial t^2} = q \tag{5}$$

The system of Eq. (5) with the corresponding boundary and initial conditions describes the motion of a viscoelastic orthotropic rectangular plate of variable thickness under the action of a periodic load  $P(t) = P_0 + P_1 \cos(\Theta t)$  taking into account initial imperfections.

In calculations, the singular kernels of the Koltunov–Rzhanitsyn type [21] are used as relaxation kernels  $\Gamma(t), \Gamma_{ij}(t), i, j = 1, 2$ :

$$\Gamma(t) = Ae^{-\beta t} t^{\alpha-1}, (0 < \alpha < 1), \Gamma_{ij}(t) = A_{ij}e^{-\beta_{ij}t} t^{\alpha_{ij}-1}, (0 < \alpha_{ij} < 1) \tag{6}$$

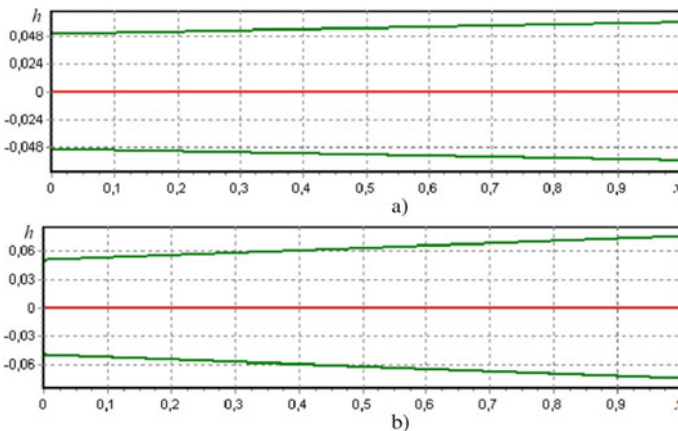
Let the plate thickness change according to the following law  $h(x) = \frac{1}{2}h_0(1 + \alpha * x)$ ; i.e., a linear increase in the plate thickness is observed (Fig. 1). Here,  $\alpha^*$  is a parameter characterizing the variability of the thickness;  $h_0$  is the plate thickness corresponding to  $\alpha^* = 0$ .

A solution to system (5) satisfying the boundary conditions of the problem is sought with respect to the displacements  $u$  and  $v$ , and deflection  $w$  in the form

$$u(x, y, t) = \sum_{n=1}^N \sum_{m=1}^M u_{nm}(t)\phi_{nm}(x, y), v(x, y, t) = \sum_{n=1}^N \sum_{m=1}^M v_{nm}(t)\phi_{nm}(x, y),$$

$$w(x, y, t) = \sum_{n=1}^N \sum_{m=1}^M w_{nm}(t)\psi_{nm}(x, y) \tag{7}$$

Substituting (7) into the system of Eq. (5) and performing the Bubnov–Galerkin procedure, taking into account dimensionless quantities



**Fig. 1** Change in plate thickness depending on parameter  $\alpha^*$ : **a**  $\alpha^* = 0.2$ ; **b**  $\alpha^* = 0.5$

$$\frac{u}{h_0}, \frac{v}{h_0}, \frac{w}{h_0}, \frac{w_0}{h_0}, \frac{x}{a}, \frac{y}{b}, \frac{h}{h_0}, \lambda = \frac{a}{b}, \delta = \frac{b}{h_0}, q^* = \frac{q}{E} \left( \frac{b}{h_0} \right)^4, \frac{\Theta}{\omega}, \omega t$$

and maintaining the previous notations, the following system of basic resolving nonlinear integro-differential equations is obtained

$$\begin{aligned} & \sum_{n=1}^N \sum_{m=1}^M a_{k \ln m} \ddot{u}_{nm} - \eta_1 \left\{ \sum_{n=1}^N \sum_{m=1}^M \left[ [(1 - \Gamma_{11}^*)d_{1k \ln m} + (1 - \Gamma^*)d_{2k \ln m}]u_{nm} \right. \right. \\ & \left. \left. + [(1 - \Gamma_{12}^*)d_{3k \ln m} + (1 - \Gamma^*)d_{4k \ln m}]v_{nm} \right] \right. \\ & \left. + \sum_{n,i=1}^N \sum_{m,j=1}^M \left[ (1 - \Gamma_{11}^*)d_{7k \ln mij} + (1 - \Gamma_{12}^*)d_{8k \ln mij} + (1 - \Gamma^*)d_{9k \ln mij} \right] (w_{nm}w_{ij} - w_{0nm}w_{0ij}) \right\} = 0, \\ & \sum_{n=1}^N \sum_{m=1}^M b_{k \ln m} \ddot{v}_{nm} - \eta_2 \left\{ \sum_{n=1}^N \sum_{m=1}^M \left[ [(1 - \Gamma_{21}^*)e_{1k \ln m} + (1 - \Gamma^*)e_{2k \ln m}]u_{nm} \right. \right. \\ & \left. \left. + [(1 - \Gamma_{22}^*)e_{3k \ln m} + (1 - \Gamma^*)e_{4k \ln m}]v_{nm} \right] \right. \\ & \left. + \sum_{n,i=1}^N \sum_{m,j=1}^M \left[ (1 - \Gamma_{22}^*)e_{7k \ln mij} + (1 - \Gamma_{21}^*)e_{8k \ln mij} + (1 - \Gamma^*)e_{9k \ln mij} \right] (w_{nm}w_{ij} - w_{0nm}w_{0ij}) \right\} = 0 \\ & \sum_{n=1}^N \sum_{m=1}^M c_{k \ln m} \ddot{w}_{nm} + \eta_3 \sum_{n=1}^N \sum_{m=1}^M p_{k \ln m}^2 (1 - 2\mu_{k \ln m} \cos \Theta t) w_{nm} \\ & - \eta_3 \left\{ \sum_{n=1}^N \sum_{m=1}^M \left[ [\Gamma_{11}^* f_{5k \ln m} + \Gamma_{12}^* f_{6k \ln m} + \Gamma_{22}^* f_{7k \ln m} + \Gamma_{21}^* f_{8k \ln m} + \Gamma^* f_{9k \ln m}] w_{0nm} \right] \right. \\ & - \eta_3 \left\{ \sum_{n,i=1}^N \sum_{m,j=1}^M w_{nm} \left[ (1 - \Gamma_{11}^*)\xi_{1k \ln mij} + (1 - \Gamma_{21}^*)\xi_{2k \ln mij} \right. \right. \\ & \left. \left. + (1 - \Gamma^*)\xi_{3k \ln mij} \right] u_{ij} + \left[ (1 - \Gamma_{22}^*)\xi_{4k \ln mij} + (1 - \Gamma_{12}^*)\xi_{5k \ln mij} + (1 - \Gamma^*)\xi_{6k \ln mij} \right] v_{ij} \right\} \\ & + \sum_{n,i,r=1}^N \sum_{m,j,s=1}^M w_{nm} \left[ (1 - \Gamma_{11}^*)g_{5k \ln mijrs} + (1 - \Gamma_{12}^*)g_{6k \ln mijrs} + (1 - \Gamma_{22}^*)g_{7k \ln mijrs} \right. \\ & \left. + (1 - \Gamma_{21}^*)g_{8k \ln mijrs} + (1 - \Gamma^*)g_{9k \ln mijrs} \right] (w_{ij}w_{rs} - w_{0ij}w_{0rs}) \Big\} = 0 \\ & u_{nm}(0) = u_{0nm}, \dot{u}_{nm}(0) = \dot{u}_{0nm}, v_{nm}(0) = v_{0nm}, \dot{v}_{nm}(0) = \dot{v}_{0nm}, \\ & w_{nm}(0) = w_{0nm}, \dot{w}_{nm}(0) = \dot{w}_{0nm} \end{aligned} \tag{8}$$

where the constant coefficients entering this system are related to coordinate functions and their derivatives:

$$\begin{aligned} p_{klnm}^2 &= f_{5klnm} + f_{6klnm} + f_{7klnm} + f_{8klnm} + f_{9klnm} - 4\pi^2 \lambda^2 p_{klnm}^* \delta_0; \\ \mu_{klnm} &= \frac{2\pi^2 \lambda^2 p_{klnm}^*}{p_{klnm}^2} \delta_1. \end{aligned}$$

Based on the developed algorithm, a program in the Delphi algorithmic language was compiled.

### 3 Results and Discussion

Integration of system (8) was carried out using a numerical method based on the use of quadrature formulas [17]. The calculation results for various physical and geometric parameters are shown in graphs, Figs. 2 and 3. The dependence of the change in thickness has the following form:  $h = 1 + \alpha^*x$ ,  $h_0 = h(0) = const$ , where  $\alpha^*$  is the parameter of thickness change.

The effect of inhomogeneous material properties on the plate behavior was studied (Fig. 2).

As seen from the figure, an increase in parameter  $\Delta = \sqrt{E_1/E_2}$  determining the degree of anisotropy (curve 1— $\Delta = 1$ ; curve 2— $\Delta = 1.5$ , and curve 3— $\Delta = 2$ ) leads to a rapid increase in the amplitude of oscillations.

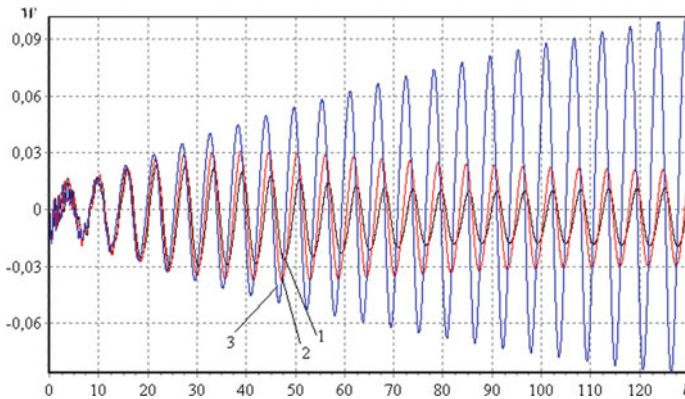


Fig. 2 Dependence of deflections on time at  $\Delta = 1$  (1); 1.5 (2); 2 (3)

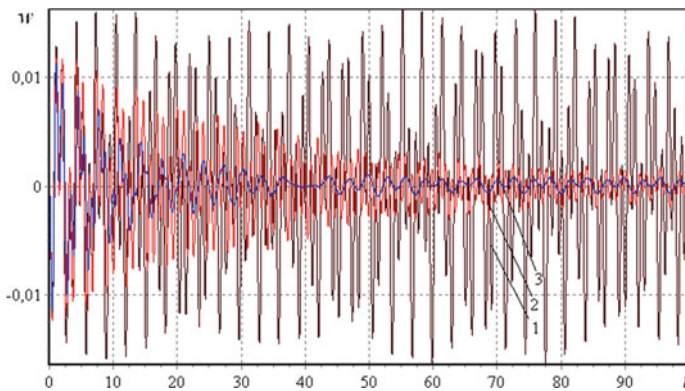


Fig. 3 Dependence of deflections on time

In Fig. 3, various curves correspond to the results obtained by various theories. Curve 1 corresponds to the elastic case, curve 2 to the results obtained taking viscosity into account in shear direction only ( $A = 0.05$ ,  $A_{ij} = 0$ ,  $i, j = 1, 2$ ), and curve 3 to the case when viscosity is taken into account in all directions ( $A = A_{ij} = 0.05$ ,  $i, j = 1, 2$ ).

The results obtained confirm the need to take into account the viscoelastic properties of the material not only in shear direction, but in other directions as well.

## 4 Conclusion

A mathematical model, method, and computer program have been developed for evaluating the parametric vibrations of a viscoelastic orthotropic rectangular plate of variable thickness, taking into account geometric nonlinearity under the action of periodic loads.

The effect of the change in physico-mechanical and geometric parameters of the plate material on the amplitude–time characteristics and stress–strain state is estimated.

The method proposed in this work can be used for various types of thin-walled structures, such as plates, panels, and shells of variable thickness.

## References

1. Bolotin, V.V.: The Dynamic Stability of Elastic Systems. Holden-Day, San Francisco (1964)
2. Volmir, A.S.: The nonlinear Dynamics of Plates and Shells. Foreign Technology Division Wright-Patterson Air Force, USA, Ohio (1974)
3. Amabili, M.: Nonlinear damping in large-amplitude vibrations: modelling and experiments. *Nonlinear Dyn.* 93 (2018). <https://doi.org/10.1007/s11071-017-3889-z>
4. Amabili, M., Alijani, F., Delannoy, J.: Damping for large-amplitude vibrations of plates and curved panels, part 2: Identification and comparisons. *Int. J. Non. Linear. Mech.* 85 (2016). <https://doi.org/10.1016/j.ijnonlinmec.2016.05.004>
5. Amabili, M.: Nonlinear vibrations of viscoelastic rectangular plates. *J. Sound Vib.* 362, 142–156 (2016). <https://doi.org/10.1016/j.jsv.2015.09.035>
6. Karpov, V.V.: Geometrically Nonlinear Problems for Plates and Shells and Methods for Solving Them. Publishing house ASV; SPbSACU, SPb (1999)
7. Karpov, V.V., Semenov, A.A.: Mathematical models and algorithms for studying strength and stability of shell structures. *J. Appl. Ind. Math.* 11, 70–81 (2017). <https://doi.org/10.1134/S190478917010082>
8. Tyukalov, Y.Y.: Finite element models in stresses for plane elasticity problems. *Mag. Civ. Eng.* 77, 23–37 (2018). <https://doi.org/10.18720/MCE.77.3>
9. Tyukalov, Y.Y.: Calculation method of bending plates with assuming shear deformations. *Mag. Civ. Eng.* (2019). <https://doi.org/10.18720/MCE.85.9>
10. Mochalin, A.A.: Modeling free oscillations of an isotropic cylindrical shell with a variable thickness and density. *J. Mach. Manuf. Reliab.* 44, 434–438 (2015). <https://doi.org/10.3103/S1052618815030127>

11. Kochurov, R., Avramov, K.: V: On effect of initial imperfections on parametric vibrations of cylindrical shells with geometrical non-linearity. *Int. J. Solids Struct.* **49**, 537–545 (2012). <https://doi.org/10.1016/j.ijsolstr.2011.10.023>
12. Kurpa, L.V., Mazur, O.S., Tkachenko, V.V.: Parametric vibration of multilayer plates of complex shape. *J. Math. Sci.* **203**, 165–184 (2014). <https://doi.org/10.1007/s10958-014-2098-2>
13. Darabi, M., Ganesan, R.: Nonlinear dynamic instability analysis of laminated composite thin plates subjected to periodic in-plane loads. *Nonlinear Dyn.* **91**, 187–215 (2018). <https://doi.org/10.1007/s11071-017-3863-9>
14. Huynh, H.Q., Nguyen, H., Nguyen, H.L.T.: Non-linear parametric vibration and dynamic instability of laminated composite plates using extended dynamic stiffness method. *J. Eng. Technol.* **6**, 170–185 (2017)
15. Kumar, R., Dutta, S.C., Panda, S.K.: Linear and non-linear dynamic instability of functionally graded plate subjected to non-uniform loading. *Compos. Struct.* **154**, 219–230 (2016). <https://doi.org/10.1016/j.compstruct.2016.07.050>
16. Kumar, R., Mondal, S., Guchhait, S., Jamatia, R.: Analytical approach for dynamic instability analysis of functionally graded skew plate under periodic axial compression. *Int. J. Mech. Sci.* **130**, 41–51 (2017). <https://doi.org/10.1016/j.ijmecsci.2017.05.050>
17. Abdikarimov, R.A., Khodzhaev, D.A.: Computer modeling of tasks in dynamics of viscoelastic thin-walled elements in structures of variable thickness. *Mag. Civ. Eng.* 83–94 (2014). <https://doi.org/10.5862/MCE.49.9>
18. Khodzhaev, D., Abdikarimov, R., Vatin, N.: Nonlinear oscillations of a viscoelastic cylindrical panel with concentrated masses. In: *MATEC Web of Conferences* (2018). <https://doi.org/10.1051/mateconf/201824501001>
19. Abdikarimov, R., Khodzhaev, D., Vatin, N.: To calculation of rectangular plates on periodic oscillations. In: *MATEC Web of Conferences* (2018). <https://doi.org/10.1051/mateconf/201824501003>
20. Ilyushin, A.A.: *Plasticity. Foundations of General Mathematical Theory*. Lenand, Moscow (2016)
21. Mal'tsev, L.E.: The analytical determination of the Rzhantsyn-Koltunov nucleus. *Mech. Compos. Mater.* **15**, 131–133 (1979)

# Methods and Technologies for Protecting Pharmaceutical Products in Polymer Packaging from Counterfeiting



Tamara Chistyakova , Roman Makaruk , Ilya Sadykov ,  
and Christian Kohlert 

**Abstract** This article considers the problem of protecting pharmaceutical products with polymer packaging from counterfeiting. This issue has grown vital in almost the entire world, as the significant harm can come not only to the producer, but the legitimate producer, but the consumers as well. Due to this, the issue of protecting these products against forgery, and creating and improving existing approaches to anti-forgery protection, becomes a crucial one. The authors suggest methods and technologies for protecting pharmaceutical products' polymer packaging based on modern ideas from IT and manufacturing such as image recognition, client-server software architecture, mobile apps, digital signatures, luminophores, and PVC film. Testing the authors' approach showed the effectiveness of the presented methods and technologies. The results should be of interest to companies producing pharmaceuticals.

**Keywords** Pharmaceutical products · Polymer packaging · Counterfeiting · Protection against forgery · Image recognition · Hardware–software solution · Encryption · Identification

## 1 Introduction

At present, the volume of counterfeit production in certain industries is comparable to the volume of legitimate production. This problem is prevalent in practically every field of economic activity, including pharmaceutical production. Counterfeit medicine makes up 10 to 80% of the overall pharmaceutical sales in Russia,

---

T. Chistyakova (✉) · R. Makaruk  
St. Petersburg State Institute of Technology, St. Petersburg, Russia  
e-mail: [chistb@mail.ru](mailto:chistb@mail.ru)

I. Sadykov  
Engineering-Server, Atlassian, Sydney, Australia

C. Kohlert  
Klößner Pentaplast Europe GmbH & Co.KG, Montabaur, Germany



which provides the counterfeiters with annual income of approximately 7 billion dollars. Worldwide, counterfeit medicine sales amount to 600 billion US dollars. Counterfeiting has become a major social issue, not just because such products cause the legitimate producers loss of trust and income, but also because counterfeit pharmaceuticals can lead to severe health issues and even death [1, 2, 3, 4, 5, 6, 7, 8].

Traditional anti-counterfeiting methods such as holograms and radiofrequency identifiers have a number of drawbacks and can only be applied to finished products, which is inadequate for pharmaceuticals, for example, as only the packaging itself could be protected. Furthermore, an analysis of the methods used currently shows that increasing the protection's effectiveness can only be accomplished with non-deterministic algorithms based on randomness, as this increases the probability that the security features will not be fully reproduced. The currently existing protection methods utilizing magnetic bits and metallic nanoparticles resolve this issue, however are extremely expensive. Further, any protection method used for food or medical products must avoid making the packaging toxic. An additional issue is that there is no full software–hardware solution that takes into account the peculiarities of pharmaceutical production, the production volume of which is in the billions [1, 9].

Thus, developing a software package of methods, models, and forms of counterfeit protection for polymer packaging of pharmaceutical products produced in large quantities, as well as a computer system that enables encryption and identifying packaging, is financially justified.

## 2 Problem Definition

An overview of the anti-counterfeiting systems on the market shows us that the field is actively developing. Among the existing systems, there are some that address the flaws of the traditional protection methods; however, they have their own issues, such as needing to label each package with a unique mark. Table 1 shows the compared

**Table 1** Relative characteristics of counterfeit protection systems

System	Encryption: technology and label element	Identification: reading the label	Protection level
Tesa scribos	Special stickers. Several levels of protection. Labeling equipment	Special equipment, a closed system	High
ForgeGuard	Special non-unique labels. Single protection level	Special scanning equipment	Low
RFID (various systems)	A label with an antenna and chip. One invisible protection level. Special equipment	Special readers. A closed system	High

characteristics of several systems currently on the market [6, 7, 8, 10, 11, 12]. None of the available technologies are adequate for full protection of pharmaceutical packages, as the production volumes of pharmaceutical products greatly complicate any attempt to mark every single one (due to production costs). Furthermore, pharmaceutical packaging may become deformed during use (e.g., when a customer pops a tablet out of its packaging), which makes sticker untenable for the task as well.

Having analyzed the existing anti-counterfeiting systems, we can ascertain the necessary characteristics of our anti-counterfeiting protection method software package. A generalized functional scheme of the anti-counterfeiting protection method package is presented in Fig. 1, where the following notation is used:  $Z_i$ —a package’s digital signature;  $F_i$ —identification result output parameter vector;  $V_i$ —the vector of configurable parameters of the deformable package area coordinate assessment;  $R_i$ —three-point circumscribed circle radius, in pixels;  $I$ —encryption picture in either “.jpg” or “.bmp” format;  $X_i$ —encoding input parameter vector.

The proposed process for product protection consists of several steps:

- Labeling the product with a unique code;
- Entering this code into a registry of legitimate codes;
- Testing product legitimacy by scanning this code;
- Searching for the scanned code in the legitimate code database.

The first two steps together make up the “encryption” stage. At that stage, the unique package code is entered into the legitimate code registry. The remaining

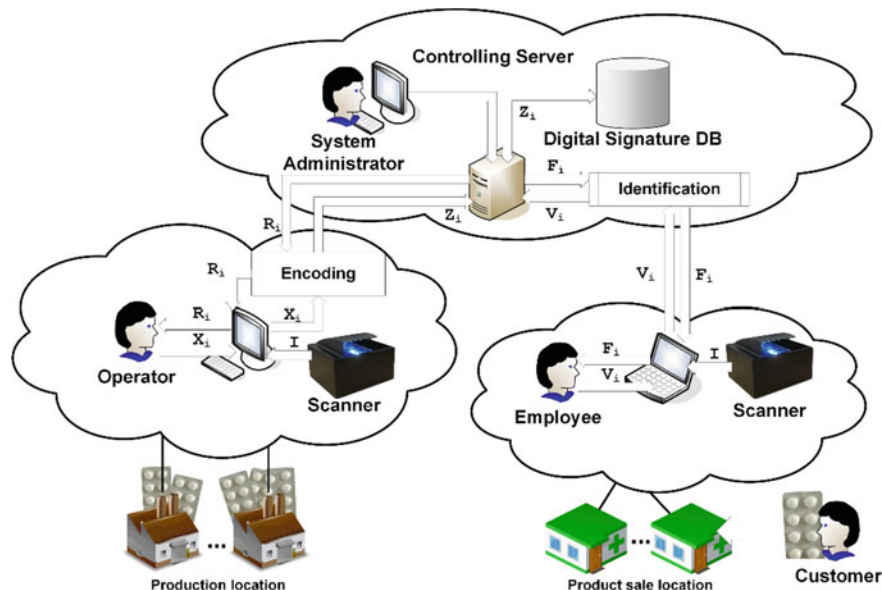


Fig. 1 Software package overall architecture

steps are merged into the “identification” stage. The product package is identified by looking up its code in the registry of legitimate codes.

The created anti-counterfeiting system satisfies the following requirements:

- It uses protection elements that are not labels and cannot be detected with the naked;
- It uses a unique code for each separate package, one based on an element of randomness in order to prevent its reconstruction;
- It can encode a large number of products (up to a billion per product type a year);
- It can identify a product sample within reasonable time (no more than a minute);
- It can partially identify the product (within the input error range) should the product’s packaging become deformed;
- It allows for encryption and identification parameters to be reconfigured to suit the customer’s requirements.

Keeping in mind the overall functional scheme of the anti-counterfeiting method software package as well as the specifics of the subject area, the developed software package consists of the following elements:

- The encryption subsystem, which consists of package scanning (photograph) equipment as well as the software, which takes the digital photograph and uses the user-specified configuration and unique code creation algorithm to calculate that code and send it to the server to be stored in legitimate code registry.
- The identification subsystem, which consists of package scanning (photograph) equipment (it is possible to use a smartphone for this provided it has the appropriate app installed), as well as software that can get take the digital photograph and, using user-specified configuration and code formation selection, produces the unique code for that package and sends it to the server to verify its presence in the legitimate code registry.
- The server component consists of a database containing information regarding products that the software produces codes for, the registry of valid codes, as well as a component that can enter valid codes into the registry or verify code presence therein.

To implement the developed methods and technologies for antiforgery protection the following two-level approach was suggested. Luminophore particles (which cannot be seen by bare eye) are randomly spread in the polymer film, the package is made of. Activating these particles requires ultraviolet or infrared light, or smartphone camera flash. The cost of creating such protection elements is less than 0.01 cents/square meter, and the luminophore content in the resulting polymer film is only 0.001%, which corresponds approximately to 1–5 pigment particles per square centimeter of film [9, 13, 22, 23]. The luminophores found to be most appropriate to the task are photoluminophores. This is primarily due to the composition of the substances, as well as the simpler and more universal excitation method for this type of luminophore.

### 3 Polymer Packaging Encryption

The photoencryption process depends on such parameters as the encryptable object (polymer film or credit card), selected encryption element (As of right now, the software package permits one of three elements—a triangle, a circle, and a rectangle), diameter of the applied particle (in microns), and the permitted encryption variance.

The first step of the encryption is image recognition, i.e., selecting the  $n$  brightest areas from the set and presenting them as points on the packaging material surface. Then, out of those  $n$ ,  $k$  (the size of the hash) points are randomly selected. The size of the hash, just as the encryption method, can be configured by the manufacturer depending on production volumes, product protection level demands, and scanning equipment resolution. In the second encryption step, a set of points  $n$  is analyzed and separated into subsets. It is important to note that during normal product use by the customer (e.g., when taking tablets out of their blister packs), the luminophore microparticles can shift from their initial position, which complicates identification. In order to prevent deformation from blocking packaging identification, it is necessary to discard points in damaged areas  $S_{\text{def}}$  from consideration [9, 14, 15, 16, 17, 18, 19, 22, 23].

The total number of points  $n_i$ , recognized on the  $i$ -th package is formed randomly and depends on the number of luminophore particles distributed on its surface:

$$n_i = k + l + m_i + o_i, \quad (1)$$

where  $3 < l < k$ ;  $m_i < n_i$ ;  $k \leq n_i - m_i$ ;  $n_i$  is the total number of points on the packaging;  $m_i$  is the number of points within deformed areas;  $k$  is the number of bright points;  $l$  is the number of points used during encryption;  $o_i$  is the number of remaining points.

Thus, polymer packaging encryption occurs in three steps:

1. An operator gets a vector of deformed areas  $O$  based on image parameters by overlaying a mask in the form of geometric primitives  $J$  as vectors on the deformed areas  $S_{\text{def}}$  and removes them from the packaging surface image  $S_{\text{pack}}$ .
2. Based on the encoding parameters  $X_i$ , a geometric code of the polymer packing  $Y_i$  is formed, taking into account the deletion of the deformable packing regions  $O$ .
3. A digital signature  $Z_i$  is formed based on the polymer package geometric code  $Y_i$ , and encoded packaging characteristics  $G_i$  are created. It ensures code uniqueness when encrypting the  $i$ -th package, for  $i = 1 \dots N$  ( $N$  is packaging production volume, its volume is calculated in billions).

### 4 Identifying Polymer

In order to check the legitimacy of the packaging (perform identification), image recognition of elements in geometric alignments must occur once more, followed by

a search for the created digital signature in the database so as to assess the product type and check for counterfeiting. During identification, the input is the processed package image, the brightest points  $k$ , the amount of them, and a randomly selected set  $l$  of points from  $k$  which is used to calculate the digital signature. The complexity of the digital signature depends directly on their amount. For example, for the triangle method of encryption, if  $l = 4$ , then the number of triangles is equal 4. At  $l = 5$ , the number of triangles is already 10, and at 6 points it becomes 20. Accordingly, the number of attributes making up the digital signature is twice the amount of triangles [9, 14, 15, 16, 17, 18, 19, 22, 23].

Packaging identification occurs in two stages. The first stage is verifying a full match for the digital signature based on  $l$  points. The second is iteration over all combinations of points from  $k$  with  $l$  and comparison of the geometric element attribute values  $r_j$  that were created based on  $p$ -th combination of points with the values  $r_{c_{p,j}}$  from the legitimate digital signature DB:

$$\forall r_{p,j} : |r_{c_{p,j}} - r_{p,j}| \leq \mu, p = 1 \dots Q_{\max \text{ full}}, j = 1 \dots N_l \quad (2)$$

where  $N_l$  is the number of attributes to be saved that together make up the digital signature;  $Q_{\max \text{ full}}$  is the maximum number of checks permitted to be done at the first identification stage;  $\mu$  is the maximum permitted geometric element (triangle, circle) attribute variance compared to the saved values (in degrees, pixels, and square pixels).

If at least one match of the locally created digital signatures and the information stored on the database occurs, the package is confirmed as legitimate. The maximum number of checks at the first stage equal:

$$Q_{\max \text{ full}} = C_k^l = \frac{k!}{l!(k-l)!}, N_l = f(M_i) \quad (3)$$

where  $M_i$  is the package encryption method.

Should the first step finish without a match, an attempt to find a partial digital signature match using  $l-1$  points is made:

$$\forall r_{p,j} : |r_{c_{p,j}} - r_{p,j}| \leq \mu, p = 1 \dots Q_{\max \text{ part}}, j = 1 \dots N_{l-1} \quad (4)$$

where  $Q_{\max \text{ part}}$  is the maximum number of checks at the second identification step;  $N_{l-1}$  is the number of saved digital signature attributes when building geometric elements using  $l-1$  points.

The maximum number of checks at this step equals:

$$Q_{\max \text{ part}} = C_k^{l-1} = \frac{k!}{(l-1)!(k-l+1)!}, N_{l-1} = f(M_i) \quad (5)$$

If a partial match is found between at least one of the partial digital signatures and one digital signature stored in the database, the packaging is accepted as legitimate,

though the user gets a warning about possible package deformation. The package is confirmed to be counterfeit if neither the first nor the second identification step produce a match [9, 14, 15, 16, 17, 18, 19, 22, 23].

## 5 Polymer Film Encryption and Identification Algorithms

The input parameters for the encryption algorithms are the coordinates of the selected bright spots in the photographs. The algorithm output is parameters of the resulting geometric elements. The main requirement for an encryption algorithm is consistency (a set of points will produce the same output unless the points are changed). The reason for its primacy being that orientation is not controlled during the polymer film photographing and can be different during encryption and identification. However, the same points are recognized in both pictures, and their relative positions remain unchanged, though their coordinates may. As a result, the point processing algorithm was created in such a way that any point set orientation on a plane will produce the same parameters.

In order to keep the encoding system general, a library of encryption methods that uses various geometric models has been developed. Creation of these geometric models uses  $l$  random points from an array  $B_k$  of the  $k$  brightest points. Each method is characterized by  $r$  geometric models based on which the digital signature ( $Z$ ) is created. The checksum of the digital signature  $A$  is a number calculated for each set of geometric models and depends on the encryption method used:

- Using triangle edges, it saves the two minimal angles  $r_j = \{a_{j1}, a_{j2}\}$  of each  $j$ -th triangle for  $j = 1 \dots u$ . The overall number of attributes saved is expressed using the formula:

$$N_l = u \cdot 2 = C_l^3 \cdot 2 = \frac{l!}{3 \cdot (l-3)!} \quad (6)$$

The checksum of the  $j$ -th packaging's digital signature is calculated according to the:

$$A_i = \sum_j^n (a_{j \text{ med}} + a_{j \text{ min}}), n = u = \frac{N_l}{2}, i = 1 \dots N \quad (7)$$

where  $a_{j \text{ med}}$  and  $a_{j \text{ min}}$  are the average and minimal values for the edges of the  $j$ -th triangle.

- Using the radii of the circumscribed circles, we save the  $r_j = R_j$  of each  $j$ -th circle for  $j = 1 \dots N_l$ . The overall number of attributes saved is expressed using the formula:

$$N_l = u = C_l^3 = \frac{l!}{6 \cdot (l-3)!} \quad (8)$$

The checksum of the  $i$ -th package's digital signature is calculated according to the formula:

$$A_i = \sum_{j=1}^n R_j \cdot (n - j + 1) - R_n, n = u = N_l, i = 1 \dots N, \quad (9)$$

where  $R_j$  is the radius of the  $j$ -th circle.

Using the area of the triangles we made with triangulation preserves the areas  $r_j = S_j$  of each  $j$ th triangle for  $j = 1 \dots N_l$ . The overall number of attributes saved is expressed using the formula:

$$N_l = u = l + l_{\text{int}} - 2, j = 1 \dots N_l, \quad (10)$$

where  $l_{\text{int}}$  is the number of points internal to the triangle.

The digital signature checksum is calculated according to the formula:

$$A_i = \sum_{j=1}^n S_j \cdot (n - j + 1) - S_n, n = u = N_l, i = 1 \dots N, \quad (11)$$

where  $S_j$  is the value of the area of the  $j$ -th triangle.

Thus, the digital signature of the  $i$ -th package is a set of the following parameters:

$$Z_i = \{r_{ij}, A_i, l, k, M, f, t, d\}, j = 1 \dots u, i = 1 \dots N, \quad (12)$$

A description of the encryption algorithm that enables us to create the digital signature for the  $i$ -th package is presented in Fig. 2.

The identification algorithm (see Fig. 3) allows establishing the degree of pharmaceutical packaging legitimacy with consideration for the input maximum deviation of the encryption geometric element parameters  $\mu$  from the values of the digital signature. The input parameters for the identification algorithm are encryption parameters, the scanned image, the selected encryption method, as well as the maximum permitted deviation of the identification values  $r_j$  compared to the stored values  $r_{cj}$ . The algorithm's output is the degree of packaging legitimacy  $F \in \{W_{\text{true}}, W_{\text{fake}}, W_{\text{part}}\}$ . The identification algorithm allows us to ensure the legitimacy of whole and partially deformed packaging by verifying the full and partial consistency of the digital signature possibilities and the  $i$ -th identifiable package with the legitimate digital signature in the database.

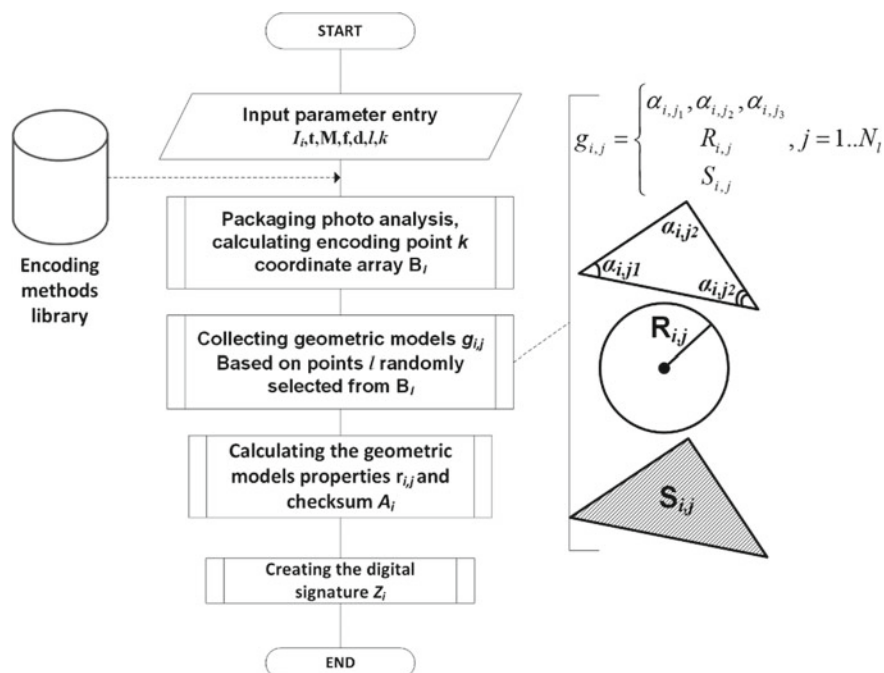


Fig. 2 Generalized digital signature algorithm

## 6 Testing

The proposed pharmaceutical product polymer packaging protection methods and technologies have already been implemented and went through testing at Klöckner Pentaplast Europe GmbH & Co. KG polymer film plants in Europe and Russia. Testing was done using EP-73 polymer film, produced according to GOST 25250-88 at polymer film plant “OOO Klöckner Pentaplast Rus” in Saint Petersburg, and the average identification time per package was no more than 30 s even with over a million fake digital signatures in the database [1, 9, 20, 21, 22, 23].

The specialized version of the software adapted to mobile devices was tested at the joint polymer film center of “Klöckner Pentaplast GmbH” and Saint Petersburg State Institute of Technology. Sample data is presented in Table 2.

Testing results: The “Lum\_04 Form3” sample provides the most system stability and result reproducibility due to:

- An appropriately rounded shape;
- Conformity to the minimal luminophore spot size;
- Luminophore spot glow period.

The proposed methods and technologies for product protection are patented in Germany and Europe [22, 23].



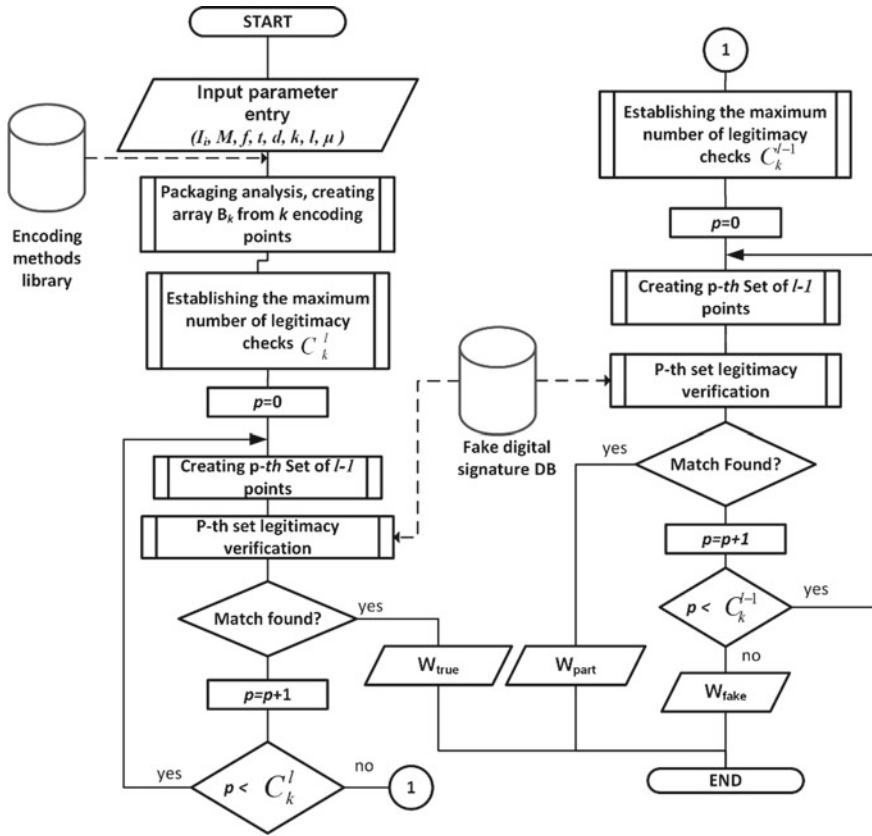


Fig. 3 Pharmaceutical packaging identification algorithm

Table 2 Data for testing the mobile app

Samples	Luminophore	Film	Production
LUM_02 Form 1	LWB520—0.08 ppm, 20 points/cm <sup>2</sup>	Transparent 200 microns	Calender
LUM_04 From 1	HK300—0.18 ppm, 40 points/cm <sup>2</sup>	Transparent 200 microns	Rolling
LUM_04 Form 3	HK300—0.04 ppm, 10 points/cm <sup>2</sup>	Colored with pink pigment, and filled with chalk. 350 microns	Rolling

## 7 Conclusion

The developed encryption methods and technologies offer an opportunity to select one of a few pharmaceutical packaging counterfeiting protection types using various

protective features and objects. The developed identification algorithm also allows partial package identification with a preset maximum geometric element parameter deviation from the encrypted value.

Testing showed that the cost of such a system is minor due to the low concentration of pigments in the product, and cheap, widespread data processing equipment (lights, camera). This technology is offered to all customers and users of “Klößner Pentaplast GmbH” polymer films for protecting their products against counterfeiting. Using the proposed physical and mathematical protection methods as well as the encryption/identification algorithms enables pharmaceutical package identification within a reasonable timespan. This software package is flexible and can be configured for various product types and anti-counterfeit pharmaceutical packaging encryption methods.

**Acknowledgements** The authors are grateful for the experimental data provided by “Klößner Pentaplast Europe GmbH & Co. KG”.

## References

1. Chistyakova, T.B., Sadykov, I.A., Kohlert, C., Ivanov, A.B.: Methods of coding and identification of pharmaceutical production to provide a protection against forgery. In: Information Technologies, pp. 52–57 (2011) (in Russian)
2. Cockburn, R., Newton, P., Agyarko, E., Akunyili, D., White, N.: The global threat of counterfeit drugs: why industry and governments must communicate the dangers. *PLoS Med* **2**(4), e100 (2005). <https://doi.org/10.1371/journal.pmed.0020100>
3. Newton, P., Green, M., Fernández, F., Day, N., White, N.: Counterfeit anti-infective drugs. *Lancet Infect. Dis.* **6**(9), 602–613 (2006). [https://doi.org/10.1016/S1473-3099\(06\)70581-3](https://doi.org/10.1016/S1473-3099(06)70581-3)
4. Kramer, A.: Drug piracy: a wave of counterfeit medicines washes over Russia. In: *The New York Times*, Sept. 5, 2006. URL: <https://www.nytimes.com/2006/09/05/business/worldbusiness/05fake.html>. Last accessed 2019/09/02
5. Newton, P., Green, M., Fernández, F.: Impact of poor-quality medicines in the ‘developing’ world. *Trends Pharmacol. Sci.* **31**(3), 99–101 (2010). <https://doi.org/10.1016/j.tips.2009.11.005>
6. Wilson, J.M., Grammich, C., Chan, F.: Organizing for brand protection and responding to product counterfeit risk: an analysis of global firms. *J. Brand Manag.* (2016). <https://doi.org/10.1057/bm.2016.12>
7. Schuha, G., Haag, C.: How to prevent product piracy using a new TRIZ-based methodology. In: *Procedia Engineering* (2011). <https://doi.org/10.1016/j.proeng.2011.03.128>
8. Potdar, M., Chang, E., Potdar, V.: Applications of RFID in pharmaceutical industry. In: *Proceedings of the IEEE International Conference on Industrial Technology* (2006). <https://doi.org/10.1109/ICIT.2006.372726>
9. Kohlert, C., Kohlert, M., Chistyakova, T., Ivanov, I., Sadykov, I.: Counterfeit-proofing based on the principle of randomness. *Kunststoffe Int.* **7**, 32–35 (2010)
10. Fujifilm ForgeGuard Anti-Counterfeit Label. <https://ideasmodern.com/ideas/fujifilm-forgeguard-anti-counterfeit-label/>. Last accessed 2019/09/02
11. Market-leading security technologies for product and brand protection. [https://www.tesa-scribos.com/eng/security\\_technologies](https://www.tesa-scribos.com/eng/security_technologies). last accessed 2019/09/02
12. Manning, L.: Food safety and brand equity. *Br. Food J.* (2007). <https://doi.org/10.1108/00070700710761491>

13. Bakhmet'ev, V.V., Tamamura, H., Nakanishi, I., Korsakov, V.G., Sychev, M.M.: Synthesizing white-luminescence SrS:Pr powder phosphor. *J. Optical Technol.* **78**(7), 449–451 (2011). <https://doi.org/10.1364/JOT.78.000449>
14. Levachkine, S., Velázquez, A., Alexandrov, V., Kharinov, M.: Semantic analysis and recognition of raster-scanned color cartographic images. *Lect. Notes Comput. Sci.* **2390**, 178–189 (2002)
15. Kharinov, M.V.: Image segmentation method by merging and correction of sets of pixels. In: *Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications)*, vol. 23, Issue 3, pp. 393–401 (2013). <https://doi.org/10.1134/S1054661813030061>
16. Kharinov, M.V.: Image segmentation using optimal and hierarchical piecewise-constant approximations. In: *Pattern Recognition and Image Analysis (Advances in Mathematical Theory and Applications)*, vol. 24, issue 3, pp. 409–417 (2014). <https://doi.org/10.1134/S1054661814030092>
17. Konstantinov, I.S., Lazarev, S.A., Shulyak, B.Yu.: Analysis of image feature points detecting methods. *Information Systems and Technologies* **3**(107), 33–39 (2018). (In Russian)
18. Zhilyakov, E.G., Konstantinov, I.S., Chernomorets, A.A.: Decomposition of images into additive components. *Int. J. Imaging Robot.* **16**, 1–8 (2016)
19. Zhilyakov, E.G., Konstantinov, I.S., Chernomorets, A.A., Bolgova, E.V.: Image compression subband method. *Int. J. Soft Comput.* **10**, 442–447 (2015)
20. Makaruk, R.V.: The software solution for the analysis of the characteristics and evaluation of information security level of a computer network based on fuzzy models. In: R. V. Makaruk, A. A. Bolshakov (eds.) *Software Engineering*, № 6, pp. 268–282 (2016) (in Russian). <https://doi.org/10.17587/prin.7.268-282>
21. Zegzhda, D., Zegzhda, P., Pechenkin, A., Poltavtseva, M.: Modeling of information systems to their security evaluation. In: *SIN '17 Proceedings of the 10th International Conference on Security of Information and Networks*, pp. 295–298. <https://doi.org/10.1145/3136825.3136857>
22. Kohlert, C., Schmidt, B., Egenolf, W., Chistyakova, T.: Patent DE 10 2008 032 781 A1. Verpackungsfolie für Produktauthentifizierung, Authentifizierungsverfahren und –system
23. Kohlert, C., Schmidt, B., Egenolf, W., Chistyakova, T.: Patent WO 2010/003585 A1. Packaging film for product authentication, authentication method and system

# Solving Multicriteria Optimization Problem for an Oil Refinery Plant



Dmitri Kostenko , Vyacheslav Shkodyrev , and Vadim Onufriev 

**Abstract** This article describes the process of multicriteria optimization of a complex industrial control object using Pareto efficiency. The object is being decomposed and viewed as a hierarchy of embedded orgraphs. Performance indicators and controlling factors lists are created based on the orgraphs and technical specifications of an object, thus allowing to systematize sources of influence. Using statistical data archives to train, the neural network approximates key sensors data to identify the model of the controllable object and optimize it.

**Keywords** Decomposition · Pareto efficiency · Multicriteria optimization · Identification · SPEA2 · Neural network

## 1 Introduction

Multicriteria optimization is a process of simultaneous optimization for two or more conflicting functions within one point [1]. The need to simultaneously optimize different functions, often presented as an array of key performance indicators (KPIs), exists in business and industrial production. Solving such tasks is important from both theoretical and practical standpoints.

For instance, in paper [2], authors consider optimizing the rim thickness of the ring of a high-ratio planetary system with straight gears. Using a set of simulations, they were able to introduce a solution, improved by multiple criteria. The work [3] focuses on evaluating significance of different criteria for multicriteria optimization in technical systems. In [4], authors successfully applied Pareto optimality principle to choose a set of hardware for a modular robot. Paper [5] describes the process of optimization of a set of fuzzy parameters to resolve a temperature control problem. In [6], authors are dealing with similar task for a floor heating system, resolving it using simulation software and a first-order nonlinear differential model. Another work on heating-related multicriteria optimization present in [7], where authors are

---

D. Kostenko (✉) · V. Shkodyrev · V. Onufriev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [Zaba-1@bk.ru](mailto:Zaba-1@bk.ru)

using the VIKOR method. Paper [8] analyzes relationships between KPIs for business structures, taking a more general approach to the problem.

Oil refinement facility, as the one viewed in this paper, is an advantageous direction for multicriteria optimization because of its complexity. Optimization process is preceded by the decomposition [9] of the refinement sequence, which is followed by the model identification [10]. Pareto optimality principle [11] is used in conjunction with the aforementioned models to build a front of optimal values.

Part of the refinement process takes place inside a refraction unit (RU). One of these units was taken as a prototype for our model. The model got the following KPIs assigned [12]:

- Quality (matching degree between the output and sample values)
- Performance (output product volume)
- Efficiency (resource usage potency)
- Reliability (equipment failures per unit of time)
- Safety (emergencies per unit of time).

Rectification consists of a wide array of parameters, up to several hundreds of characteristics per one refraction unit. These include the splitting section, column head, and column plates temperature and pressure. Inside and outside of the rectification column both sequential (multi-layered raw oil refinement, raw oil heating, raw oil pumping, etc.) and parallel (vapor condensation) processes are taking place. Rectification technology also includes transition products into the process, thus applying an additional, horizontal level of hierarchy between the operations. It is also essential to account for the time delay, added by the inertia of the system itself and enforced by the continuous operating mode.

Consequently, processes from the highest levels of the hierarchy have unobvious connections with the lower-level processes thus making it impossible to use simple functions like  $y = f(g, u)$  to represent dependencies between them. However, it is essential to be able to influence the top-level processes by changing parameters of the low-level processes and vice versa.

In this work, the aforementioned problem is resolved by decomposing a complex system (such as refraction unit) down to individual units and processes. The resulting structure is represented as a graph. The KPI set takes the top level of the hierarchy. Every KPI is divided into several summands of a lower hierarchy levels. The step is repeated until the summand can be unambiguously interpreted by the  $y = f(x)$  type of dependency. Dependencies are identified using a neural network trained on the RU statistical data archive. Going up by one hierarchy level changes dependency to  $y = g(f(x))$ . Ascending by the hierarchical tree allows to determinate a clean dependency between a KPI and an input parameter from the bottom of the hierarchy.

However, the top-level key performance indicators may directly contradict each other. For instance, raising performance by forcing aggressive operating parameters will inevitably cause the growth in equipment failures and an overall reduction of reliability, which in turn damages efficiency of the refraction unit or the refinery as a whole. The “Good—Fast—Cheap” triangle encourages us to use a multicriteria optimization algorithm to balance out conflicting key performance indicators.

The aim of this work is to perform a multicriteria optimization to find a Pareto optimal solution. This allows us to find a safe combination of controllable parameters able to keep the target indicators inside the given target intervals.

This analysis is based on statistical data archive taken from a working refinery. It was used to build the graphical representations of dependencies between performance and temperature, characterizing the lowest hierarchy level. In order to optimize the top-level KPIs by changing the bottom-level controllable parameters, a strong correlation must be revealed. To grant it, a dependence model identification has been performed [13].

## 2 Data Identification

The refraction unit consists of oil preheater with a heat-exchange unit, a fractionating column, a refrigerator, and a boiler. The preheated oil is injected into the column feeding zone to be divided into vapor and solid phases. During the rectification process isopentane is extracted from the top part of the column as a fractionator overhead. Heavier fractions are taken from the plates in the middle of the rectification column. The heaviest part, the long residuum, is extracted from the bottom part of the column [14]. A simplified scheme, showing distillation inputs and outputs targeted by this work, is present on Fig. 1.

The stripper temperature ( $U_4$  in Table 1) has been chosen for the optimization. Temperature variation was aimed at maximizing fractions 240–300 and 300–350 output volumes ( $G_2$  and  $G_3$  in Table 1).

To identify dependencies (see Fig. 2), a neural network (NN) was utilized. It was trained on statistical data obtained during 24 h of the prototype column work.

The neural network consists of 1 input, 1 hidden, and 1 output layer. Input and output layers both contain 1 neuron, while the hidden layer contains 10. The NN was trained using the backpropagation method. Hidden and output layers contain a sigmoid activation function (1) (Fig. 3):

$$F(x) = \frac{e^x}{(e^x + 1)} \tag{1}$$

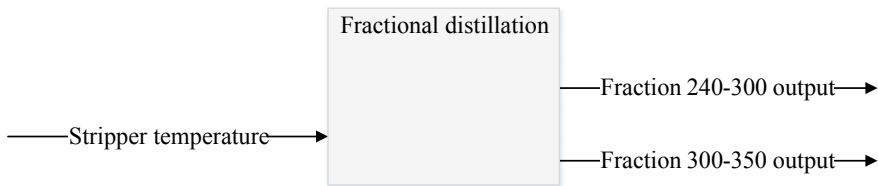
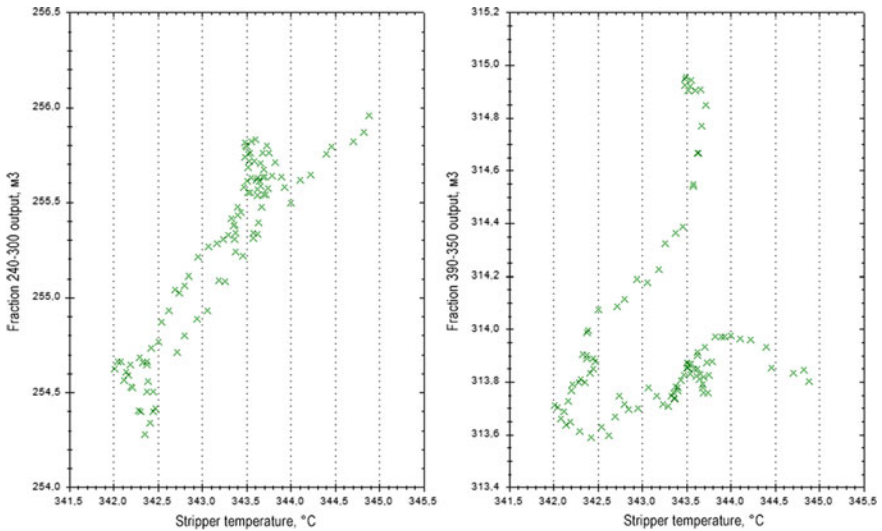


Fig. 1 Simplified scheme of the rectification process

**Table 1** Controllable parameters and target indicators of the controllable object

Input parameters	Controllable parameters (U)		Target indicators (G)	
Raw oil	U <sub>1</sub>	Input oil volume	G <sub>1</sub>	Fraction 140–240 output volume
	U <sub>2</sub>	Input oil temperature	G <sub>2</sub>	Fraction 240–300 output volume
	U <sub>3</sub>	Input oil pressure	G <sub>3</sub>	Fraction 300–350 output volume
	U <sub>4</sub>	Stripper temperature	G <sub>4</sub>	Long residuum output volume
	U <sub>5</sub>	Stripper pressure	G <sub>5</sub>	Fractionator overhead output volume
	U <sub>6</sub>	Plate 23 temperature		
	U <sub>7</sub>	Plate 36 temperature		
	U <sub>8</sub>	Plate 49 temperature		

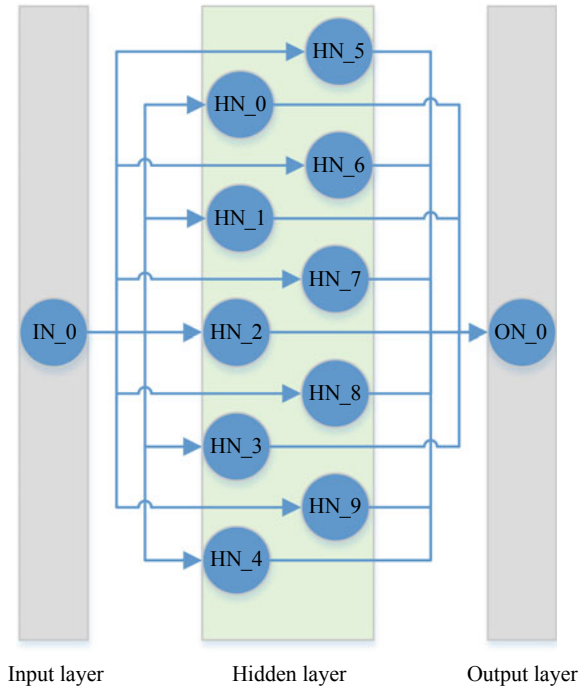


**Fig. 2** Dependencies between fraction outputs and temperature

The general formula for the NN is (2):

$$ON = f \left( \sum_{k=0}^{10} f \left( \sum_{k=0}^{10} IN \times IHN \right) \times OHN \right) \tag{2}$$

**Fig. 3** Neural network scheme



Here  $f$  stands for sigmoid activation function (1), while IN, ON, IHN, and OHN are present in Tables 2 and 3.

Application of the aforementioned neural network to input data allowed to identify dependencies between the stripper temperature and output volumes of 240–300 and 300–350 fractions (see Fig. 4).

To verify the correlation between the models (lines on Fig. 4) and statistical data (crosses on Fig. 4), correlation coefficients  $\rho_{x,y}$  have been calculated. The left graph has  $\rho_{x,y} = 0.76312$  and the right graph has  $\rho_{x,y} = 0.90781$ . Coefficient formula (3) is present below, and  $\sigma_x$  and  $\sigma_y$  are the mean values of the corresponding selections.

$$\rho_{x,y} = \frac{\frac{1}{n} \sum_{t=1}^n X_t Y_t - (\frac{1}{n} \sum_{t=1}^n X_t)(\frac{1}{n} \sum_{t=1}^n Y_t)}{\sigma_x \sigma_y} \tag{3}$$

Correlation numbers could be improved by increasing the size of the training selection for the neural network. Amount of NN's hidden layers and/or hidden neurons is also subjects to change. But in order not to deviate from the main theme of the work, achieved correlations have been accepted as sufficient.



**Table 2** Hidden neural layer for fraction 240–300

Neyron	Input (IN)	Output (ON)	Input weight (IHN)	Output weight (OHN)
0	0.5329	0.4675	0.9646	0.5356
1	−0.4489	0.4218	−0.8125	−1.0897
2	0.6221	0.6083	1.1261	0.6041
3	−0.1149	0.5999	−0.2079	−0.7524
4	0.7458	0.5611	1.3499	0.8198
5	2.5314	0.7924	4.5817	3.1874
6	−1.0151	0.3031	−1.8373	−1.8294
7	1.5302	0.7177	2.7695	1.8181
8	−0.0096	0.4126	−0.0174	−0.3680
9	−0.3281	0.3505	−0.5939	−0.8191

**Table 3** Hidden neural layer for fraction 300–350

Neyron	Input (IN)	Output (ON)	Input weight (IHN)	Output weight (OHN)
0	−0.4983	0.2103	−0.9457	0.7026
1	−3.5414	0.0843	−6.7410	−3.9722
2	−0.7764	0.3018	−1.4716	1.3219
3	0.7931	0.4216	1.5021	−1.4161
4	−0.7696	0.3039	−1.4587	1.3393
5	−0.1638	0.1935	−0.3109	0.2529
6	−0.1228	0.2843	−0.2331	0.1390
7	−0.2815	0.2045	−0.5342	0.4098
8	−0.5891	0.2605	−1.1171	0.9753
9	0.0778	0.2562	0.1475	−0.1713

Having mathematical models of interconnections between the basic parameters of the oil refinement process in place, it is now possible to compare them and define the optimal points according to the multicriteria optimization method.

### 3 Pareto Optimality

Pareto optimality is a state of allocation of resources from which it is impossible to reallocate so as to make any one individual or preference criterion better off without making at least one individual or preference criterion worse off [15].

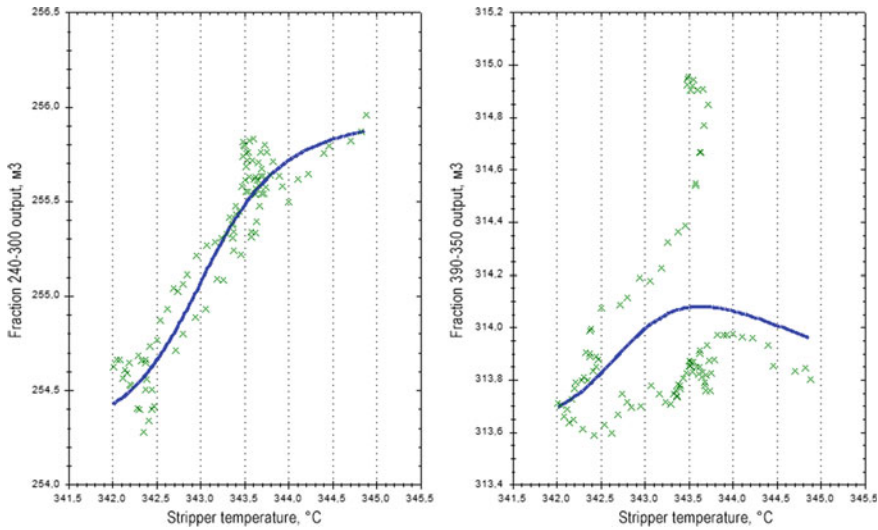


Fig. 4 Graphical representation of the identified models

Pareto front within the range of target functions is a combination of solutions, which do not dominate each other, but dominate every other solutions within the search space at the same time. It means that it is impossible to find a single solution able to excel every other solution at reaching every target. Mathematically such problem can be formulated as follows: one must find a vector  $X^* = [x_1^*, x_2^*, \dots, x_n^*]^T$ , that would optimize a vector of target functions  $F(X) = [f_1(X), f_2(X), \dots, f_k(X)]^T$  while having  $m$  inequality constraints  $g_i(X) \leq 0, i = \overline{1, m}$  and  $p$  equality constraints  $h_j(X) = 0, j = \overline{1, p}$ .

Here  $X^* \in R^n$  is a solution vector;  $F(X) \in R^k$  is a vector of target functions every single one of which must be optimized [16].

Strength Pareto Evolutionary Algorithm 2 (SPEA2) [17] was used for Pareto optimization. Despite its relatively old age, it is a well-tested algorithm, effective for select applications [18], including more representative spread of non-dominated solutions [19], and was chosen over others, including VEGA, FFGA [20], and NPGA [21].

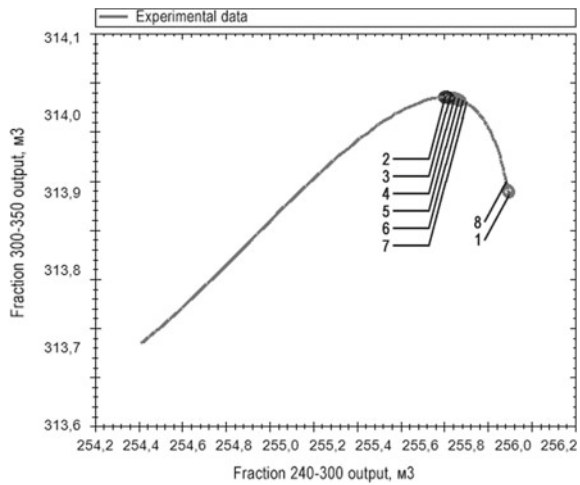
SPEA2 algorithm can be summarized in 6 steps:

- Step 1, Initialization: generate an initial population  $P_0$  and create an empty archive (external set)  $\overline{P}_0 = \emptyset$ . Set  $t = 0$ .
- Step 2, Fitness assignment: calculate fitness values of individuals in  $P_t$  and  $\overline{P}_t$ .
- Step 3, Environmental selection: copy all nondominated individuals in  $P_t$  and  $\overline{P}_t$  to  $P_{t+1}$ . If the size of  $P_{t+1}$  exceeds  $\overline{N}$ , then reduce  $P_{t+1}$  using the truncation operator, otherwise if the size of  $P_{t+1}$  is less than  $\overline{N}$ , fill  $P_{t+1}$  using dominated individuals in  $P_t$  and  $\overline{P}_t$ .

- Step 4, Termination: if  $t \geq T$  or another stopping criterion is satisfied, then set A to the set of decision vectors represented by the non-dominated individuals in  $P_{t+1}$ . Stop.
- Step 5, Mating selection: perform a binary tournament selection with replacement on  $P_{t+1}$  in order to fill the mating pool.
- Step 6, Variation: apply the recombination and mutation operators to the mating pool and set  $P_{t+1}$  to the resulting population. Increment generation counter ( $t = t + 1$ ) and go to Step 2.

SPEA2 algorithm has been utilized to find the front of temperatures, maximizing the output of fractions 240–300 and 300–350. Previously identified models have been used as “experimental data” on the graph seen on Fig. 5.

**Fig. 5** Pareto front



**Table 4** Complete stats of the Pareto front

Point number	Temperature (°C)	Fraction 240–300 output volume ( $M^3$ )	Fraction 300–350 output volume ( $M^3$ )
1	344.823852539063	255.869043673089	313.914703058703
2	343.325988769531	255.623350989602	314.04678075533
3	343.371765136719	255.636549610692	314.046650398148
4	344.712951660156	255.866950335304	313.918779855307
5	343.374877929688	255.661217222795	314.045467958112
6	343.35595703125	255.672724869121	314.044445521759
7	343.364624023438	255.649165327318	314.046207859264
8	343.407836914063	255.683708140522	314.043155299305

As a result, we got a Pareto front, consisting of eight points (see Table 4). The points are sorted in descending order of preference, thus making the first point the most optimal.

## 4 Conclusion

In the course of the work, a refraction unit has been decomposed, and its input, controllable, and target parameters were extracted and listed. Based on statistical data analysis, the neural network was trained and then used to identify dependency models between the RU parameters. The models allowed us to define eight points of Pareto front.

We are planning on extend the sphere of practical application of this method. First of all, we all have to decompose complete sets of dependencies from the basic controllable parameters up to top level KPIs. This would allow us to see key performance indicators of separate units and the whole refinery changing in real time.

Further software development enables us to revert the process and guess controllable parameters, able to sustain a predefined set of high-level KPIs. Such instrument will not only allow to optimize complex processes, but will also ensure much more effective control over them.

## References

1. Steuer, R.: Multiple Criteria Optimization: Theory, Computations, and Application. John Wiley & Sons, New York (1986)
2. Mihailidis, A.K., Pupaza, C.: Multicriteria optimization of planetary systems. Annals of DAAAM for 2011 & Proceedings of the 22nd International DAAAM Symposium, Volume 22, No. 1. In: Editor B. Katalinic, DAAAM International, Vienna, Austria, EU (2011)
3. Malakov, I., Zaharinov, V.: Computer aided determination of criteria priority for structural optimization of technical systems. Proc. Eng. **69**, 735–744 (2014). <https://doi.org/10.1016/j.proeng.2014.03.049>
4. Andreev, V., Kim, V., Pletenev, P.: Using pareto optimum to choose module's computing platforms of mobile robot with modular architecture. In: Katalinic, B. (ed.) Proceedings of the 29th DAAAM International Symposium, pp.0559–0565, Published by DAAAM International, Vienna, Austria. <https://doi.org/10.2507/29th.daaam.proceedings.081>
5. Nahlovsky, T.: Optimization of fuzzy controller parameters for the temperature control of superheated steam. In: Proceedings of the 25th DAAAM International Symposium on Intelligent Manufacturing and Automation, vol. 100, pp. 1547–1555, Published by Elsevier Ltd. (2015)
6. Maric, P., Bevanda, I.: Parameter Estimation And Optimization of the Hydronic Floor Heating System as a Basis for Predictive Control of the Zone Level Energy. In: Katalinic, B. (ed.) Proceedings of the 29th DAAAM International Symposium, pp.1038–1045, Published by DAAAM International, Vienna, Austria. <https://doi.org/10.2507/29th.daaam.proceedings.148>
7. Civic, A., Vucijak, B.: Multi-criteria optimization of insulation options for warmth of buildings to increase energy efficiency. In: Proceedings of the 24th DAAAM International Symposium on Intelligent Manufacturing and Automation, 2013. Procedia Engineering, vol. 69, pp. 911–920. Published by Elsevier Ltd (2014)

8. Durkacova, M., Lavin, J., Karjust, K.: KPI optimization for product development process. In: Katalinic, B. (ed.) *Annals of DAAAM for 2012 & Proceedings of the 23rd International DAAAM Symposium*, vol. 23, no. 1, CDROM version, Published by DAAAM International, Vienna, Austria (2012)
9. Yaochu, J.: Pareto-optimality is everywhere: from engineering design, machine learning, to biological systems. In: *2008 3rd International Workshop on Genetic and Evolving Systems*, pp. 1–1. IEEE (2008)
10. Kostenko, D., Kudryashov, N., Maystrishin, M., Onufriev, V., Potekhin, V., Vasiliev, A.: Digital twin applications: diagnostics, optimisation and prediction. In: *Proceedings of the 29th DAAAM International Symposium*, pp. 574–581. DAAAM International, Vienna (2018)
11. Evans, G.W., Stuckman, B., Mollaghasemi, M.: Multicriteria optimization of simulation models. In: *1991 Winter Simulation Conference Proceedings*, pp. 894–900. Phoenix, Arizona (1991)
12. Key performance indicators in the oil & gas industry. <https://www.performancemagazine.org/key-performance-indicators-oil-bp/>. Last accessed 2019/01/23
13. Schleicha, B., Anwer, N., Mathieu, L., Wartzack, S.: Shaping the digital twin for design and production engineering. *CIRP Ann. Manuf. Technol.* **66**(1), 141–144 (2017)
14. Alfke, G., Irion, W.W., Neuwirth, O.S.: Oil refining. In: *Ullmann's Encyclopedia of Industrial Chemistry* (2007)
15. Keeney, R., Raiffa, H.: *Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge University Press, Cambridge (1993)
16. Coello Coello, C.A., Christiansen, A.D.: Multi-objective optimization of trusses using genetic algorithms. *Comput. Struct.* **75**(6), 647–660 (2000)
17. Zitzler, E., Laumanns, M., Thiele, L.: SPEA2: Improving the Strength Pareto Evolutionary Algorithm. In: *TIK Report №103*. Switzerland, Zurich (2001)
18. Tang, Y., Reed, P., Wagener T.: How effective and efficient are multiobjective evolutionary algorithms at hydrologic model calibration? *Hydrology and Earth System Sciences Discussions*. European Geosciences Union, vol. 10, issue 2, pp. 289–307 (2006)
19. Chołodowicz, E., Orłowski, P.: Comparison of SPEA2 and NSGA-II applied to automatic inventory control system using hypervolume indicator. *Stud. Informatics Control* **26**(1), 67–74 (2017)
20. Fonseca, C.M., Fleming, P.J.: Genetic algorithm for multiobjective optimization, formulation, discussion and generalization. In: *Genetic Algorithms: Proceeding of the Fifth International Conference*, pp. 416–423. California (1993)
21. Horn, J.N., Nafpliotis, A.L., Goldberg, D.E.: A niched Pareto genetic algorithm for multiobjective optimization. In: *Proceedings of the First IEEE Conference on Evolutionary Computation*, IEEE World Congress on Computational Intelligence, pp. 82–87. IEEE Service Center, Piscataway (1994)

# Methods and Techniques for Increasing the Accuracy of Continuous Non-invasive Blood Pressure Measurement Under Dynamic Loads



Gleb Zaitsev , Alexei Vassiliev , and Quang-Kien Trinh 

**Abstract** In this paper, various methods for measuring blood pressure (BP) are considered and discussed in detail. Among those, we focus on the indirect BP measurement using the volume clamp method (VC). We considered a hardware solution using the SAKR-2 (Ltd Intox) device for non-invasive BP measurement. Based on our experimental results, this solution is well suited to BP dynamics measure for the patients both at rest and during movement. Specifically, by analyzing the characteristics of multiple measured results in different scenarios, we have proposed several post-processing techniques to remove the systematic errors during measurements (e.g., due to the measurement condition) and enhance accuracy. Compared to the conventional direct BP measurement methods in the radial artery (RA) using commercial device (S5 monitor), both systolic blood pressure (SBP) and diastolic blood pressure (DBP) indicators have strong relationship. The measured BP shifts with respect to hydrostatic changes in blood pressure are fairly match to the theoretical predetermined value during repetitive measurements.

**Keywords** Medical applications · Blood pressure sensors · Continuous non-invasive blood pressure measurement · Embedded measurement systems

## 1 Introduction

Measurement of blood pressure (BP) essentially is one of the most commonly important measurements used in outpatient, inpatient facilities, as well as in everyday life because it provides important indicators of the physiological processes of the patient's cardiovascular system.

---

G. Zaitsev (✉) · A. Vassiliev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [zaitsevgleb@gmail.com](mailto:zaitsevgleb@gmail.com)

Q.-K. Trinh  
Le Quy Don Technical University, Hanoi, Vietnam

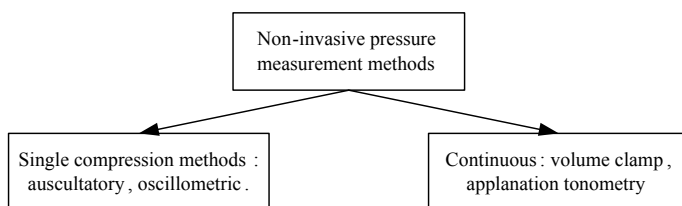
For more than a hundred years, the scientific community has made great efforts to develop algorithms and techniques for the implementation of blood pressure measurement devices using indirect measurement approaches. BP direct measurement undoubtedly is one of the most widely adopted BP measurement methods in cardiology practice during surgery, in intensive care units and also for research purposes [1]. This method typically is carried out for a relatively limited range of indicators and does not require either a complicated interpretation of physical processes or a highly accurate measuring system [2]. Nonetheless, this method remains fundamentally an invasive approach, where the complexity of the measurement procedure significantly simplifies the instrument design. The main advantages of non-invasive (indirect) methods are the exclusion of trauma, protecting the integrity of the patient's skin and eliminating the risk of infection [3, 4].

From other point of view, in any measuring system, it is a necessity to take into account the dynamic properties of the target signal. Specifically, in BP measurement, BP reflects hemodynamics and continuously on each heartbeat. Instantaneous BP values are represented by a pressure pulse. The characteristic parameters of the pulse are distinguished by anacrotic rise, dicrotic wave, and incisura [5, 6]. Capturing the pattern of the BP signal is the most difficult task in non-invasive measurements. Most instruments allow to evaluate BP using extremals of systolic blood pressure (SBP) pulse, diastolic blood pressure (DBP) or based on empirical relationships. They also are able to quantify the BP pulse (PBP) and average pressure as an integral indicator calculated from a series of heartbeats [7].

Non-invasive BP measurement methods can be conditionally divided into the following groups (see Fig. 1):

- methods for measuring BP parameters by compressing the vascular system in a single compression–decompression cycle of air in the cuff;
- continuous BP measurement methods.

In medical practice, the Korotkov method [8] has been long recognized as the “golden standard” for BP measuring. This method calculates the estimated pressure value indirectly extracted from a series of brachial artery pulses. However, this method permits evaluating only one SBP and DBP value per single measurement, and the resulting values vary with respect to different heartbeat periods. The error of the method hence is statistically determined by the magnitude of the variations in SBP, DBP and the rate of pressure decrease in the cuff. This leads to inevitable



**Fig. 1** Classification of BP measurement methods

ambiguity in determining the SBP as the presence of “auscultatory failure” (short-term disappearance of Korotkov tones) and does not allow determining the blood pressure in patients with severe rhythm disturbance. This is due to the fact that BP pattern is significantly different at every heartbeat [9].

Korotkov’s method is suitable for characterizing and extracting diurnal BP dynamics from periodic measurements, meanwhile, in various types of medical diagnosis and research, the instantaneous pressure caused by shorter heartbeat may reflect important information. Continuous measurements permit recording in detail the reaction of the cardiovascular system to a given external load (orthostatic test, stress test, etc.) [10–12]. Despite the fact that the measured value is associated with a short period, it carries significant information about the dynamics of the system, and the transient process can be further used to characterize the system during a short measurement.

Among non-invasive continuous BP measurement methods, the volume clamp (VC) (known as the Peñáz method) is one of the most popular methods. As the fundamental principle, BP measurement is carried out using a pneumatic cuff on a finger [13]. The reliability of measured BP by the Peñáz method has been solidly proven both for patients at rest [14–16] and for patients in BP changing conditions (such as an orthostatic test) [17, 18]. Nonetheless, these works focused only on comparing the average blood pressure obtained by various methods while the accuracy in recording pressure dynamics and variations for every single heartbeat were not considered in detail.

This paper systematically studies the accuracy aspect of the BP measurement using the volume clamp approach. We then proposed techniques for enhancing the reliability of the measured BP for the patient both at rest and during movement condition.

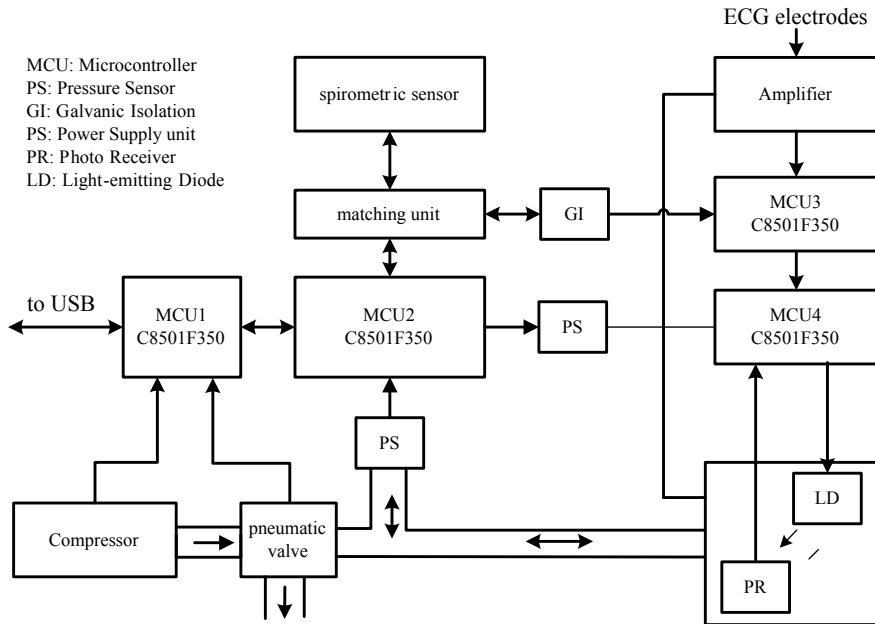
The remaining of the paper is organized as follows. Section 2 describes the research methodology and the experimental setup. Section 3 presents and compares the main experimental results. Further development and enhancement techniques are discussed in Sect. 4 before the conclusions are drawn in Sect. 5.

## 2 Methodology and Experimental Setup

The VC method in this work has been implemented using an embedded hardware platform. In particular, we adopted the Spiroarteriocardiorhythmograph (SAKR-2) system from Ltd. “Intox” described in the study [19, 20], whose structure is shown in Fig. 2.

The SAKR-2 device based on the VC method is calibrated and configured before every BP measurement. During the setup, a tacho-oscillogram is used for estimating SBP, DBP, and average pressure values. The obtained parameters then are used to set the initial settings for continuous BP measurement using the VC method. The measurement device does not yet support correction for peripheral pressure





**Fig. 2** Block diagram of the electronic-pneumatic device «SAKR»

components associated with the shoulder. The latter basically is a hydrostatic pressure component caused by the relative height of the finger position with respect to the heart. The measurement devices either have no post-processing for the values measured in the shoulder. This leads to interesting questions: how accurately the device registers small-scale physiological pressure fluctuations, and how accurately the device reproduces the pressure curve during significant and abrupt measurement condition shift from the original setting. In fact, the accuracy of recording changes in pressure is especially noticeable when monitoring pressure during stress tests.

Evaluation of the accuracy in measuring the BP dynamics using the VC method was carried out in two series of experiments:

1. comparison of pressures obtained by the method of RA and direct method;
2. checking the coincidence of theoretical and measured pressure in the finger with “hydrostatic” changes in blood pressure by a predetermined value.

The measurements were carried out at the Scientific Institution of Experimental Medicine. The study involved seven patients aged from 39 to 76. In a motionless patient, the BP was simultaneously recorded by the direct method with RA on the left hand and the VC method on the right-hand finger.

Direct measurement of blood pressure was performed using a bedside S5 monitor from Datex-Ohmeda [21]. Accordingly, an invasive sensor recorded blood pressure in the radial artery. The measurement results were displayed every 5 s.

From the quantified SBP and DBP values, the constant component, calculated by averaging recorded BP signals, was removed in order to extract the BP variable components. The constant component then was analyzed separately. For the variable component, we adopted a statistical method to quantify the accuracy of recorded BP dynamics using the VC method. Specifically, the standard deviations and correlation coefficients between the BP values measured by the invasive method and the SAKR-2 device were statistically evaluated and compared.

In the second series of tests, the dynamics of blood pressure measured on the finger for every single heartbeat was compared with the pressure measured on the wrist by the oscillometric method with significant changes in BP in the limb. The change in pressure was set by the vertical movement of the limb relative to the heart level. With a vertical movement of an unstressed arm, finger pressure  $P(h)$  (in mmHg) is determined by the following formula:

$$(h) = P_0 + \frac{\rho_K}{\rho_{Hg}}(h_0 - h) + \Delta P \quad (1)$$

where  $P_0$  is the systolic or diastolic pressure in the finger at a height of  $h_0$ ;  $\rho_K$  is the blood density;  $\rho_{Hg}$  is the density of mercury;  $h$  is the current height of the finger;  $\Delta P$  is the pressure variation due to respiration, the influence of the baroreflex, and other physiological factors.

A change in the height of the hand relative to the level of the heart leads to a noticeable pressure drop. Accordingly, pressure variations can introduce not only random but also systematic errors, for example, associated with the influence of a baroreflex during hydrostatic changes in peripheral pressure. An Omron R2 tonometer, which measures pressure on the wrist, was used as a monitoring device. The tonometer indicator values were used to exclude the systematic error caused by the physiological characteristics of the subject.

The methodology of these studies was as follows. The subject's hand in an extended relaxed position was located on a convenient stand, and this stand can be rotated to two fixed positions relative to the subject's shoulder. The fixed positions were chosen so that the height of the cuff changed exactly by 520 mm, which corresponds to a change in pressure of 40 mmHg. In each of the fixed positions, a pressure measurement was performed with a tonometer after 10 s displacement. Averaging was carried out according to the results of five measurements in each position.

Measurements by the SAKR-2 device were performed in a similar manner, i.e., the subject's arm in an extended relaxed position was located on a convenient stand, which can be rotated to two fixed positions. The device setup was carried out in the upper arm position, then the device performed continuous BP measurements. For every 15 s, the stand with the hand rotates to the opposite fixed position. The positions were carefully chosen so that the height of the finger with the cuff changed exactly by 520 mm, which corresponds to a 40 mmHg change in BP.

The average BP values measured by the SAKR-2 device in each of the stationary positions were compared with the average BP values measured by the Omron R2

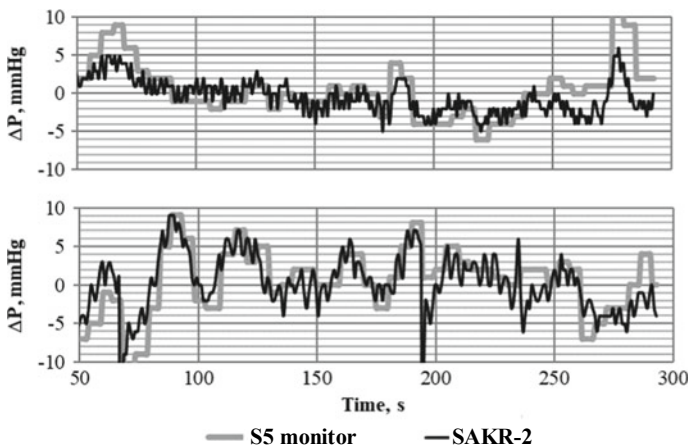
tonometer. Changes in systolic (SBP), diastolic blood pressure (DBP), and pulse blood pressure (PBP) relative to the highest pressure were analyzed. The detail is presented and discussed in the subsequent section.

### 3 Experimental Results

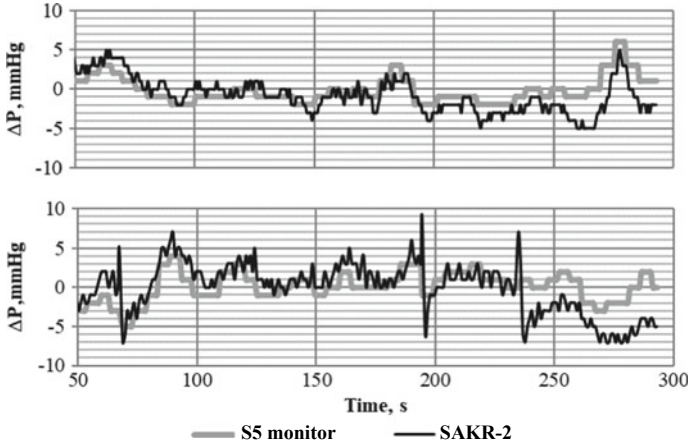
#### 3.1 Comparison of Non-invasive BP Measurement in the Finger and Invasive BP Measurement in the Radial Artery at Rest

The BP dynamics are most clearly manifested when observing the BP variable components retrieved from the direct method and recorded by the SAKR device for every single heartbeat. For this purpose, the average level during the synchronous measurement was previously removed from the recorded data. Figures 3 and 4 show examples of SBP dynamics in direct synchronous measurements using the S5 monitor and SBP measured by SAKR-2 using the VC method.

From the graphs presented in Figs. 3 and 4, the overall recorded signals patterns using S5 monitor and SAKR-2 are very much in line. However, there is a large discrepancy in the absolute value due to the averaging effect in the invasive devices. On the last graph, extrasystole moments are visible that did not appear on the monitor due to averaging of invasive results. The standard deviation for SBP and DBP is summarized in Table 1. The correlation coefficient calculated for zero time lag between signals.



**Fig. 3** Comparison of resting SBP variations during synchronous measurement by a S5 monitor in the radial artery and by the SAKR-2 in the finger



**Fig. 4** Comparison of the DBP variations at rest during synchronous measurement by the S5 monitor in the radial artery and by the SAKR-2 in the finger

**Table 1** Standard deviation and correlation coefficients calculated from synchronous BP measurements by the SAKR-2 device and direct BP measurements in the radial artery after removal of the stationary components

#	Std Dev* SBP, mmHg	Std Dev* DBP, mmHg	r** SBP (p-value)	r** DBP (p-value)
1	3.3	2.9	0.5 ( $p < 0.05$ )	0.3 ( $p < 0.05$ )
2	3.0	2.0	0.5 ( $p < 0.05$ )	0.6 ( $p < 0.05$ )
3	2.0	1.6	0.9 ( $p < 0.05$ )	0.7 ( $p < 0.05$ )
4	1.6	1.0	0.8 ( $p < 0.05$ )	0.9 ( $p < 0.05$ )
5	3.5	3.1	0.7 ( $p < 0.05$ )	0.4 ( $p < 0.05$ )

*Std Dev* standard deviation, *r* correlation coefficient, and *p* probability value.

From Figs. 3 and 4, there is a notable gradual pressure drift when measured by the SAKR device caused by the change in the blood supply to the finger. After performing a linear approximation on the data, we found that, on average, systolic pressure (diastolic pressure) decreases at a rate of  $0.02 \pm 0.001$  mmHg/s ( $0.025 \pm 0.007$  mmHg/s). The described signal attenuation in a five-minute recording possibly causes a pressure decrease of 6 mmHg. The drift in the first minutes of measurement by the VC method is associated with the displacement of blood from the capillary vessels, which affects the constant level of the photosignal in the finger. This drift takes 3–5 min. After this time, it is necessary to repeat the adjustment of the control system.

**Table 2** Measured BP shift with lowering of the arm by 520 mm (predicted change is 40 mmHg)

№	Oscillometric method (Omron R2)			Volume clamp method (SAKR-2)		
	$\Delta$ SBP $\pm$ Std Dev, mmHg	$\Delta$ DBP $\pm$ Std Dev, mmHg	$\Delta$ PBP $\pm$ Std Dev, mmHg	$\Delta$ SBP $\pm$ Std Dev, mmHg	$\Delta$ DBP $\pm$ Std Dev, mmHg	$\Delta$ PBP $\pm$ Std Dev, mmHg
1	51 $\pm$ 10	44 $\pm$ 4	7 $\pm$ 9	49 $\pm$ 3	37 $\pm$ 1	12 $\pm$ 2
2	46 $\pm$ 10	43 $\pm$ 3	3 $\pm$ 9	40 $\pm$ 5	38 $\pm$ 2	2 $\pm$ 3
3	41 $\pm$ 6	37 $\pm$ 4	4 $\pm$ 8	36 $\pm$ 6	31 $\pm$ 3	5 $\pm$ 5
Average	48 $\pm$ 9	41 $\pm$ 5	8 $\pm$ 8	42 $\pm$ 6	36 $\pm$ 3	6 $\pm$ 6

$\Delta$ —average pressure shift at different heights of the measurement point relative to the heart.

### 3.2 *Quantifying the Accuracy of the BP Measurement Under a Significant Measurement Condition Change*

To determine the level of BP changes using the hydrostatic method, we used an Omron R2 carpal tonometer as the reference for the BP shift. Table 2 shows the difference between results of BP measurement in different arm level. Results are presented for oscillometric method (Omron R2) and VC method (SAKR-2).

The reliability of BP measurement by the volume clamp method is strongly affected by the outflow of blood from the finger. In many cases, finger BP obtained both by the oscillometric method and by the VC method significantly exceeds the BP indicators in the shoulder. In such cases, it is observed a significant redness of the finger portion that is far from the occlusion site. This is due to the fact that the outflow of venous blood is completely blocked; thus, the average pressure of the finger area increases correspondingly. The described case is also in line with the deterioration in the reproduction of the BP pulse when measured by the VC method. After adjusting the device to lower the pressure in the limb, raising it relative to the level of the heart, the diastolic pressure practically does not change, while the systolic pressure significantly decreases.

## 4 Discussion on Further Improvements

The blood pressure profile measured by the SAKR-2 showed many similarities to the BP profile obtained by the direct measurement method in the radial artery (standard deviation of results less than 3.5 mmHg). In all cases, signal dependences are observed ( $p$ -value  $<$  0.05) and the correlation coefficients of SBP are greater than 0.5.

The linear trend, observed during the first minutes of measurement by the RA method, is associated with the displacement of blood from capillary vessels, which affects the constant level of the photosignal in the finger. This blood displacement occurs within 3–5 min. After this period has passed, it is necessary to recalibrate

the control system. Alternatively, a preliminary procedure for stabilizing the photo-signal (displacing blood from under the finger cuff) with excessive pressure can be conducted. During this procedure, the transient response of regulation would be determined by the dynamic characteristics of the photosignal associated with a particular patient.

In absolute measurements of blood pressure by the direct method in the radial artery and by SAKR-2 in the finger, very often, significant differences between two approaches, especially regarding the diastolic blood pressure, were observed. The discrepancy may be due to a combination of the following factors:

- measurements were taken in various places of the arterial bed. As the pulse wave advances, its shape and amplitude change due to changes in the resistance and stiffness of the vessels.
- the heights of the forearm and finger relative to the heart were not taken into account.

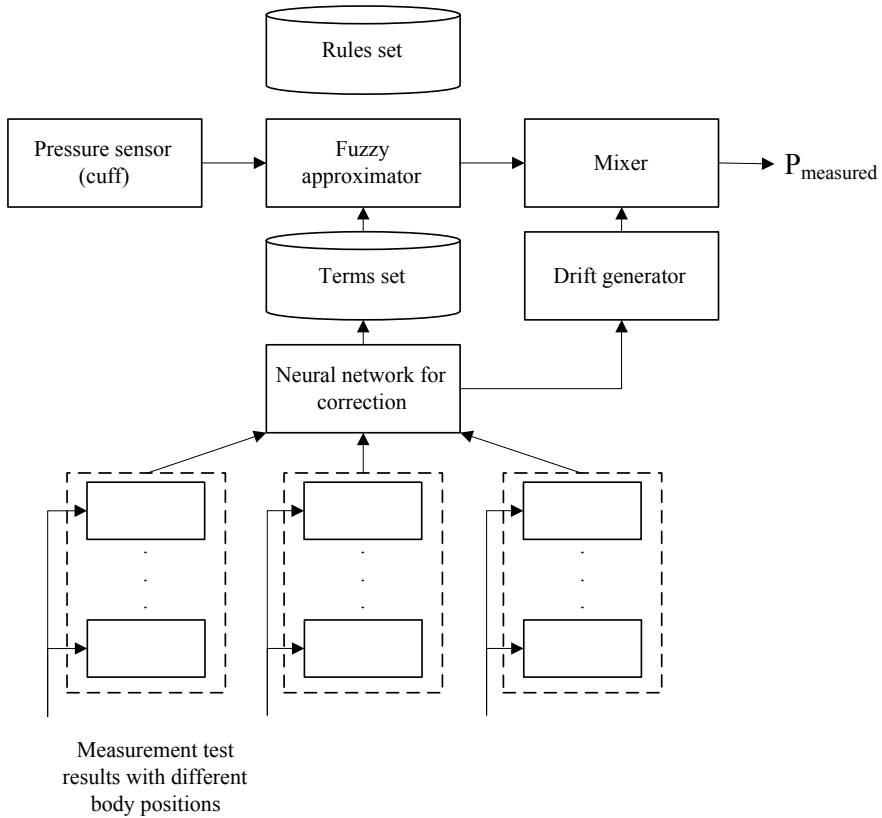
Obtaining the usual estimations of blood pressure, corresponding to measurements in the shoulder at the level of the heart, requires constructive and algorithmic techniques for improving the measuring system.

It is required to develop a system for adjusting the values of peripheral BP at the shoulder for each measurement. Due to significant changes in the signal at different heartbeats, signals describing hemodynamics in the shoulder should be simultaneously recorded while measuring the BP pressure in the finger. The correction value should be based not only on final integral indicators, such as SBP and DBP but on the corresponding continuous signals recorded at different parts of the system (BP signal from the shoulder cuff, finger cuff, microphone, ECG, photosignal, etc.).

One possible solution for BP correction is to use an additional finger sensor. This reference value helps automatically adjusting the pressure change during the movement of the patient's arm. Nonetheless, with changing the patient's body position during measurement (e.g., during an orthostatic test), this technique may not guarantee good level of accuracy.

Analyzing the total existing features of the BP measurement system, it is suggested that potential techniques for correcting measurement results should be based on intelligent data processing algorithms, i.e., machine learning approaches. The structure of such a correction system is shown in Fig. 5. In the proposed system, all modules are implementable on embedded hardware. This will be our main focus in future work, and detailed discussion on that structure is beyond the content of this paper.

Data from the pressure sensor needs to be intelligently corrected in depends on a large number of parameters. In addition to the body positions, following parameters may affect the result: the pulse wave propagation velocity, the error in the regulation of the photo signal, the heart rate, etc. These parameters are sent to the neural network. The results of the neural network through a set of terms are provided to a fuzzy approximator, which corrects the original signal.



**Fig. 5** Structure of the correction system for BP measurement

## 5 Conclusions

The method that measures BP for a single heartbeat in the finger is well suited for use under physical activity, in which the change in BP relative to the resting state is used to evaluate the final results. Despite the fact that this method is non-invasive, it still is capable to record continuously the relative BP dynamics with a fairly high level of accuracy. With potential further improvements proposed above, this BP measurement method could potentially replace and eventually overcome distinct disadvantages of the conventional invasive BP measurement method in majority cases of medical practice.

## References

1. Agabiti-Rosei, E., et al.: Central blood pressure measurements and antihypertensive therapy: a consensus document. *Hypertension* **50**, 154–160 (2007)
2. Gupta, B.: Invasive blood pressure monitoring. *Update Anaesthesia* **28**, 37–42 (2012)
3. Durie, M., Beckmann, U., Gillies, D.M.: Incidents relating to arterial cannulation as identified in 7.525 reports submitted to the Australian incident monitoring study (AIMS-ICU). *Anaesth Intensive Care* **30**(1), 60–65 (2002)
4. Scheer, B., Perel, A., Pfeiffer, U.J.: Clinical review: complications and risk factors of peripheral arterial catheters used for haemodynamic monitoring in anaesthesia and intensive care medicine. *Crit Care* **6** (2002)
5. Esper, S.A., Pinsky, M.R.: Arterial waveform analysis. *Best Pract. Res. Clinical Anaesthesiol.* **28**, 363–380 (2014)
6. Thiele, R.H., Durieux, M.E.: Arterial waveform analysis for the anesthesiologist: past, present, and future concepts. *Anesth Analgesia* **113**(4), 766–776 (2011)
7. Avolio, A.P., van Bortel, L.M., Boutouyrie, P., et al.: Role of pulse pressure amplification in arterial hypertension: experts opinion and review of the data. *Hypertension* **54**(2), 375–383 (2009)
8. Williams, B., et al.: 2018 ESC/ESH guidelines for the management of arterial hypertension. *Eur. Heart J.* **39**(31), 3021–3104 (2018)
9. Pickering, T.G.: Blood pressure variability and ambulatory monitoring. *Curr. Opin. Nephrol. Hypertens.* **2**(3), 380–385 (1993)
10. Marchenko, V., Noskin, L., Pivovarov, V., et al.: Prognostic and diagnostic value of moderate intensity stress test with blood pressure control. In: *The Scientific Notes of the I. P. Pavlov St. Petersburg State Medical University*, vol. 21, issue 4, pp. 18–21 (2014)
11. Miyai, N., Arita, M., Miyashita, K., et al.: Blood pressure response to heart rate during exercise test and risk of future hypertension. *Hypertension* **39**(8), 761–766 (2002)
12. Schultz, M.G., Otahal, P., Cleland, V.J., et al.: Exercise-induced hypertension, cardiovascular events, and mortality in patients undergoing exercise stress testing: a systematic review and meta-analysis. *Am. J. Hypertens.* **26**(3), 357–366 (2013)
13. Penaz, J.: Photoelectric measurement of blood pressure, volume and flow in the finger. In: *Proceedings of the Conference Committee of the 10th International Conference on Medicine and Biological Engineering*, pp. 104, Dresden (1973)
14. Guelen, I., Westerhof, B.E., Van Der Sar, G.L., Van Montfrans, G.A., Kiemeneij, F., Wesseling, K.H., Bos, W.J.: Finometer, finger pressure measurements with the possibility to reconstruct brachial pressure. *Blood Press Monit.* **8**(1), 27–30 (2003)
15. Schutte, A.E., Huisman, H.W., Van Rooyen, J.M., Malan, N.T., Schutte, R.: Validation of the Finometer device for measurement of blood pressure in black women. *J. Hum. Hypertens.* **18**(2), 79–84 (2004)
16. Silke, B., McAuley, D.: Accuracy and precision of blood pressure determination with the Finapres: an overview using re-sampling statistics. *J. Hum. Hypertens.* **12**(6), 403–409 (1998)
17. Imholz, B.P., Settels, J.J., Van der Meiracker, A.H., Wesseling, K.H., Wieling, W.: Non-invasive continuous finger blood pressure measurement during orthostatic stress compared to intra-arterial pressure. *Cardiovasc. Res.* **24**(3), 214–221 (1990)
18. Parati, G., Casadei, R., Groppelli, A., Di Rienzo, M., Mancia, G.: Comparison of finger and intra-arterial blood pressure monitoring at rest and during laboratory testing. *Hypertension* **13**(6 Pt 1), 647–655 (1989)
19. Pivovarov, V.V.: A Spiroarteriocardiorhythmograph. *Biomed. Eng.* **40**(1), 45–47 (2016)
20. Pivovarov, V.V., Zaytsev, G.K., Sizov, V.V.: Adaptive exercise stress system SAKR-VELO for load tests. *Biomed. Eng.* **49**(2), 74–78 (2015)
21. Datex-Ohmeda S/5 Anesthesia Monitor. [https://www.ardusmedical.com/wp-content/productManuals/Anesthesia/Datex\\_Ohmeda\\_5%20Tech.pdf](https://www.ardusmedical.com/wp-content/productManuals/Anesthesia/Datex_Ohmeda_5%20Tech.pdf). Last access 2019/10/31



# Computer Modeling of Robust Control of Vibrationless Movement of Multi-mode Flexible Structures



Vladimir A. Prourzin , Kiseon Kim , and Georgy Shevlyakov 

**Abstract** The article is devoted to rejection of unwanted dynamic reactions of flexible structures. The problem of vibrationless movement of an elastic object when natural oscillations are absent and the reaction of the system does not exceed the static reaction is considered. To analyze the oscillations of the system, the maximum response spectrum and the residual spectrum are used. The vibrationless movement property is defined as restrictions on these spectra specified by the control signal. The problem of vibrationless movement is solved by input shaping methods. The shaping filter built on fixed values of the eigenfrequencies of the system is unstable when the frequencies deviate from the set values. To solve this problem, robust modifications of the method are proposed. The high complexity of the considered problem requires computer modeling. In particular, using computer modeling, the problem of the choice of a shaping filter with the property of maximum robustness is solved.

**Keywords** Modeling · Robustness · Oscillation · Vibration damping · Vibrationless · Input shaping

---

V. A. Prourzin (✉)

Institute of Problems of Mechanical Engineering, Russian Academy of Sciences, St. Petersburg, Russia

e-mail: [proursin@gmail.com](mailto:proursin@gmail.com)

K. Kim

Gwangju Institute of Science and Technology, Gwangju, South Korea

G. Shevlyakov

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

# 1 Introduction

## 1.1 Preliminaries

Flexible structures are characterized by dynamic responses to disturbing and controlling actions. These undesirable reactions cause significant problems with positioning accuracy, throughput, fatigue, and safety for many types of systems [1].

Here, we consider the problem of designing control of the movement of an elastic object in such a way that natural elastic vibrations of the object do not arise at all, in other words, the vibrationless motion of an elastic object. This problem setting requires some clarification. When the movement of an elastic object under an external force, elastic deformations always arise. These deformations are of the two types: the static deformations, in which the reaction of elastic bonds compensates the external load and the dynamic vibrations with eigenfrequencies. If control is applied only to one of the masses of a multi-mass system, then static deformations cannot be excluded. Thus, considering the vibrationless movement, we are talking about the total excluding of their natural vibrations.

Under vibrationless motion, the elastic system monotonously transitions from one state of static equilibrium to another with changing of constant control force regimes. In this case, natural vibrations are absent, and the reaction of the system does not exceed the corresponding static reaction.

## 1.2 State of the Art

The problem of displacement without natural vibrations is considered in [2], where the goal is achieved by designing a smooth and slowly changing control action. In this case, a significant increase of time to achieve the goal occurs. It is practically interesting to look for vibrationless solutions, in which the time of achieving the goal increases only slightly.

An example of vibrationless control of a two-mass elastic system is given in [3], where a continuous trapezoidal control law is used. The duration of the sides of the trapezoid (linearly increasing and decreasing sections) is equal to the period of natural oscillations of the system. In these sections, the deformation of the system changes monotonously, and during the constant control regime, this deformation is also constant and equal to the static reaction. The control duration increases by an amount comparable to the period of the system natural oscillations.

The general method of damping of residual vibrations of a linear system with one degree of freedom by summing the initial signal with the same signal shifted in time by a half-period of natural oscillations is proposed in [4]. As a result, the exposure duration increases by a half-period of natural oscillations. Oscillation suppression in the active control section is not considered. There is no generalization of this approach to systems of higher dimension.

A known method of reducing elastic vibrations is called as the method of input shaping filter. A review of studies of this method over the past half century is given in [5]. The method consists of the formation of a control signal by the operation of convolution of a driving action with a pulse transition characteristic representing the sum of shifted impulses (Dirac delta functions). As a result of the convolution operation, the constructed control signal is the sum of a certain number of instances of the driving action multiplied by weight coefficients and shifted in time.

Note that in well-known works, the study of the effectiveness of input shaping methods is based on the analysis of Fourier amplitude spectrum of controls. This approach allows us to study only the residual vibrations of the system and does not guarantee a decrease in the amplitudes of vibrations in the active control interval of time.

In [6], a rigorous definition of vibrationless control is given. To exclude natural vibrations both during the displacement process and after reaching the final state, the problem of controlling the motion of a multi-mass elastic object is posed and solved. The constructed control can be implemented using a shaping filter, the main elements of which are time delay schemes. The duration of control is equal to the duration of the initial control, increased by the sum of half-periods of the natural oscillations of the object.

There arises an instability problem when using input shaping control [7]. When the eigenfrequencies of the system deviate from the set values, natural oscillations occur. The task of control design insensitive to finite deviations of natural frequencies is important—such control is called robust. Methods for constructing robust controls for damping residual oscillations are discussed in [7–10]. Also, the problem of maximally robust control design is considered. The statement and general approaches to solving such problems are described in [11–14]. Here, we consider the problem of constructing the maximum robust vibrationless controls and computer modeling methods for solving it.

The remainder of the paper is as follows. In Sect. 2, the problem setting of designing of robust vibrationless control is given. In Sect. 3, the methods and algorithms of the solution of the posed problem are considered. In Sect. 4, the results of computer modeling are presented. In Sect. 5, some conclusions are drawn.

## 2 Problem Setting

Consider an elastic object the mechanical model of which is a multi-mass system with linear elastic bonds. A scalar control force acting along a given direction is applied to one of the masses. Next, highlight the mass  $m_0$  is called the carrier mass. A control force  $F(t)$  is applied to the carrier mass. The control  $F(t)$  is bounded and finite:  $|F(t)| \leq f$ ,  $F(t) = 0$  for  $t < 0$  and  $t > T$ . The remaining  $n$  masses  $m_k$ ,  $k = 1, 2, \dots, n$ , are called attached masses. The absolute displacement of the  $k$ th mass is denoted by  $y_k$ . Newton's equations form a system of  $n + 1$  differential equations of the second order:

$$M\ddot{y} + Cy = bF(t). \quad (1)$$

Here,  $y = (y_0, y_1, \dots, y_n)^T$  is mass coordinate vector,  $M$  is diagonal mass matrix,  $C$  is symmetric stiffness matrix, and  $b = (1, 0, \dots, 0)^T$  is vector of dimension  $n + 1$ .

System (1) is reduced by a special linear change of variables  $y = Rx$  to a system of  $n + 1$  independent equations in normal coordinates [15]:

$$\ddot{x}_0 = g_0 F(t); \quad (2)$$

$$\ddot{x}_k + \omega_k^2 x_k = g_k F(t), k = 1, \dots, n. \quad (3)$$

Here,  $M_0 = \sum_{i=0}^n m_i$  is the total mass of the system,  $x_0 = \frac{1}{M_0} \sum_{i=0}^n m_i y_i$  is coordinate of the center of mass of the system,  $x = (x_0, x_1, \dots, x_n)^T$  is vector of normal coordinates, the coefficient  $g_k$  is the  $k$ th component of the vector;  $g = R^{-1}b$  has the meaning coefficient of involving in the oscillations of the  $k$ th eigenform,  $g_0 = M_0^{-1}$ . All initial values of system variables (2), (3) are zeros.

The elastic object is not connected to the base and has the form of oscillations with zero eigenfrequency  $\omega_0 = 0$ . This mode corresponds to the movement of an object as a solid. We assume that the remaining eigenfrequencies are positive and pairwise distinct:  $0 = \omega_0 < \omega_1 < \dots < \omega_n$ . The following control problem is posed.

**Problem 1.** To construct a bounded control  $F(t)$ , which moves the center of mass (2) from the zero initial state to a given state in a finite time  $T$ , providing that the system of oscillators (3) is transferred from the initial state of rest to the final state of rest without natural oscillations (the vibrationless or oscillation-free motion of the elastic system).

The purpose of the control may consist, for example, in giving the center of mass the required acceleration (the problem of a non-oscillating start), in accelerating the center of mass to a given speed or in moving the center of mass to a given distance.

Eigenfrequencies of system (1) are usually estimated with errors:  $\omega_k = \omega_k^*(1 \pm \delta_k)$ , where  $\omega_k^*$  is the exact frequency value,  $\omega_k$  is the frequency estimate, and  $\delta_k$  is the relative estimation error. Next, the following problem arises.

**Problem 2.** To construct a robust vibrationless control  $F(t)$  that solves Problem 1 for all frequencies lying in a given range of values:  $\omega_k \in [\omega_k^*(1 - \delta), \omega_k^*(1 + \delta)]$ ,  $k = 1, \dots, n$ .

### 3 Approaches and Methods

Now we give a strict definition of vibrationless. First consider a linear oscillator with a eigenfrequency  $\omega$ :

$$\ddot{x} + \omega^2 x = F(t) \quad (4)$$

The initial conditions are  $x(0) = \dot{x}(0) = 0$ . The notations are introduced:  $r(t; F) = \omega^2 x(t; F)$  is elastic reaction to the control action  $F(t)$ ,  $F_{\max} = \max_t |F(t)|$ . In [6], the vibrationless displacement of a linear oscillator is defined as a solution that consists of successive states of static equilibrium. In the transition from one state of static equilibrium to another, the solution changes monotonically without natural oscillations. The vibrationless movement of system (1) is defined as the vibrationless movement of each oscillator in system (3).

Let us give a different, more general and convenient for verification definition of vibrationless. The first fundamental property of the vibrationless motion is that the module of oscillator response does not exceed the maximum values of the control module:  $|r(t; F)| \leq \max_{t \geq 0} |F(t)|$ . The second fundamental property is the absence of residual oscillations after the end of control. We put these properties in the definition of vibrationless.

It is known [16] that the function of frequency  $\omega A(\omega; F)$ , where  $A(\omega; F)$  is the module of Fourier transform of the time function  $F(t)$ , is equal to the amplitude of the residual oscillations of oscillator (4). This function is called the residual spectrum. We introduce the function of the relative residual spectrum  $R(\omega; F)$  as the residual spectrum related to the maximum of control:

$$R(\omega; F) = \frac{\omega A(\omega; F)}{\max_t |F(t)|} = \frac{\omega}{F_{\max}} \left| \int_0^\infty F(t) e^{-i\omega t} dt \right|.$$

Another function, the dynamic coefficient  $K(\omega; F)$ , shows how many times the maximum response of an elastic oscillator with frequency  $\omega$  is greater than the static reaction caused by the maximum control value  $F(t)$ :

$$K(\omega; F) = \frac{\max_t |r(t; F)|}{\max_t |F(t)|} = \frac{\omega}{F_{\max}} \max_t \left| \int_0^t F(s) \sin \omega(t - s) ds \right|.$$

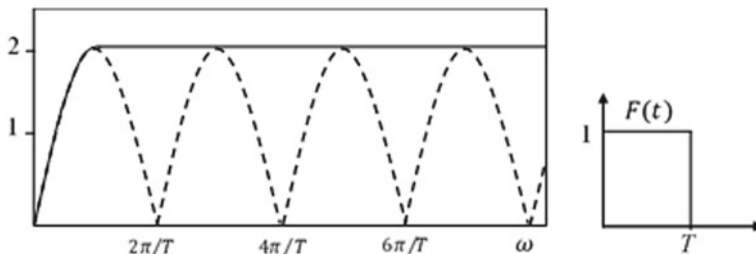
The numerator of the dynamic coefficient is an indicator known as the spectrum of maximum reactions. The following definitions are introduced.

**Definition 1.** The control  $F(t)$  and the solution of Eq. (4) with the eigenfrequency  $\omega = \omega_0$  are called *vibrationless* if the following conditions are satisfied:  $K(\omega_0; F) \leq 1$  and  $R(\omega_0; F) = 0$ .

**Definition 2.** The control  $F(t)$  and the solution of Eq. (1) with eigenfrequencies  $\omega_1, \omega_2, \dots, \omega_n$  are called *vibrationless* if the following conditions are satisfied:  $K(\omega_k; F) \leq 1, R(\omega_k; F) = 0, k = 1, \dots, n$ .

Controls that are vibrationless in the sense of [6] satisfy Def.1 and Def. 2. We indicate the following properties of the dynamic coefficient [17].

1.  $K(0; F) = 0$ .
2. There is a minimum frequency  $\omega_l$  such that  $K(\omega; F) \geq 1$  for  $\omega \geq \omega_l$ .
3.  $K(\omega; F) \geq R(\omega; F)$ .



**Fig. 1** Dynamic coefficient  $K(\omega; F)$  (solid line) and the residual spectrum  $R(\omega; F)$  (dashed line) of a rectangular pulse

4. Scale and shift transformations:  $K(\omega; \tilde{F}) = |p|K(\frac{\omega}{q}; F)$ , where  $\tilde{F}(t) = pF(q(t - \tau))$ , for any  $p, q > 0, \tau \geq 0$ .
5. If there is a piecewise continuous derivative of  $F(t)$ , then  $\lim_{\omega \rightarrow \infty} K(\omega; F) = F_{\max}$ .

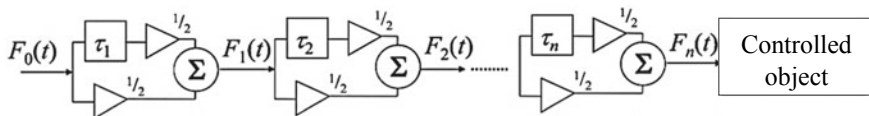
Consider, for example, the control  $F(t)$  in the form of a rectangular pulse of unit amplitude. We construct a dynamic coefficient and a residual spectrum for it (Fig. 1). We see that the dynamic coefficient has values not exceeding unity in the low-frequency region, and the residual spectrum vanishes at  $\omega = 2\pi k/T, k = 0, 1, 2, \dots$ . These sets of frequencies intersect only at  $\omega = 0$ . Thus, the conditions of Definition 2 for a rectangular pulse are satisfied only at  $\omega = 0$ .

The solution to Problem 1 is given in [6], and it is given by the following recursive procedure. Let system (1) have  $n$  nonzero eigenfrequencies  $\omega_k, k = 1, \dots, n$ . The control  $F_0(t)$ , optimal by time [18], is constructed for Eq. (2), which solves the problem of moving the center of mass. Let  $T_0$  be the minimum time to reach the goal. The control  $F_0(t)$  is the initial control. Using the half-periods of natural oscillations  $\tau_k = \pi\omega_k^{-1}$  as time shifts, a recurrent sequence of functions is constructed:

$$F_k(t) = \frac{1}{2}(F_{k-1}(t) + F_{k-1}(t - \tau_k)), k = 1, \dots, n \tag{5}$$

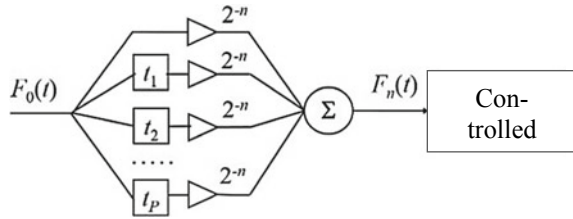
**Proposition** (solution of Problem 1) [6]. The control  $F_n(t)$  constructed according to algorithm (5) with the initial control  $F_0(t)$  solves Problem 1 during the time  $T_n = T_0 + \sum_{i=1}^n \tau_i$  and satisfies Def. 2.

Procedure (5) can be implemented using a cascade digital filter with  $n$  delay lines (Fig. 2). The control  $F_n(t)$  can also be represented using the input shaping filter



**Fig. 2** Cascade digital filter circuit that implements vibrationless control

**Fig. 3** Scheme of the input shaping filter that implements oscillatory control



(Fig. 3):

$$F_n(t) = \frac{1}{2^n} \sum_{i=0}^P F_0(t - t_i).$$

The filter includes  $P = 2^n - 1$  delay lines. The delay values  $t_i$  are all possible partial sums of half-periods  $\tau_1, \dots, \tau_n$ :  $t_0 = 0, t_1 = \tau_1, \dots, t_n = \tau_n, t_{n+1} = \tau_1 + \tau_2, \dots, t_p = \sum_{i=1}^n \tau_i$ .

The control  $F_n(t)$  has a significant drawback: the obtained vibrationless property is unstable. Even small deviations of the eigenfrequencies from the given values lead to the appearance of natural oscillations. To cope with this, we turn to the solution of Problem 2.

The Fourier transform of finite functions cannot vanish on a finite frequency interval. Therefore, it is impossible to build a robust control that ensures vibrationless on any finite frequency interval, if you require strict fulfillment of the conditions of Def. 2. Next, let weaken the requirements for vibrationless: set the permissible value  $\varepsilon > 0$  and consider the solutions, for which the following restrictions are satisfied:

$$K(\omega_k; F) \leq 1 + \varepsilon, R(\omega_k; F) \leq \varepsilon, k = 1, \dots, n. \tag{6}$$

We assume that the natural frequency  $\omega_k$  is known up to the segment  $\Omega_k(\delta)$ :  $\omega_k \in \Omega_k = [\omega_k^*(1 - \delta), \omega_k^*(1 + \delta)]$ ,  $k = 1, \dots, n$ . It is required to construct a control  $F(t)$  that solves Problem 1 and satisfies conditions (6) for all frequencies belonging to given segments. This control is called robust.

The parameter  $\Delta$  is an allowable error when estimating the natural frequencies of an object, at which the vibrationless motion in the sense of (6) is preserved. This parameter is not known in advance and we want to make it maximum. For a control  $F(t)$ , form the set of frequencies  $\Omega(F) = \{\omega : K(\omega; F) \leq 1 + \varepsilon, R(\omega; F) \leq \varepsilon, k = 1, \dots, n\}$  for which conditions (6) hold. The indicator function is introduced:  $I(\delta; F) = 1$ , if  $\bigcup_k \Omega_k(\delta) \subset \Omega(F)$  and  $I(\delta; F) = 0$ , if  $\bigcup_k \Omega_k(\delta) \not\subset \Omega(F)$ . The functional is set as

$$\delta_\varepsilon(F) = \max\{\delta : I(\delta; F) = 1\}.$$

### Problem 3 (the maximum robust control).

Let a constant  $\varepsilon$  and a set  $\mathcal{F}$  of admissible controls be given. It is required to construct a control  $F_\varepsilon^*(t)$  delivering a maximum to the functional  $\delta_\varepsilon(F)$  at  $F \in \mathcal{F}$ .

Previously, this problem was considered for the case of suppression of the residual spectrum [7–10, 19–22]. If residual spectrum is set equal to zero at the modeling frequencies, then this approach is called the zero vibration (ZV) shaper. The control  $F_n(t)$  we built relates to this case. The 5% insensitivity of the ZV shaper is 0.06 ( $\delta_{0.05}(F) = 0.03$ ).

Robust method, called specified-insensitivity (SI) shaping, suppresses a specified range of frequencies [19]. The most straightforward method for generating a shaper with specified insensitivity to frequency errors is the technique of frequency sampling. In the simplest case, two frequencies are selected in which the residual spectrum is equal zero. This approach is called the extra-insensitive (EI) approach [20]. If three frequencies are selected, the approach is called two-hump EI.

Here, the solution to Problem 3 is based on the application EI and two-hump EI approach. To solve this problem, computer simulation methods and numerical optimization methods are used.

## 4 Results

The problem of the vibrationless set of speed  $v_0$  is simulated. The initial control  $F_0(t)$  is a rectangular pulse with amplitude  $f$  and duration  $T = v_0/f$ . Due to the properties of the dynamic coefficient  $K(\omega_k; F)$  and the residual spectrum  $R(\omega_k; F)$  and the input shaping method, Problem 3 is sufficient to solve for a certain frequency  $\omega_0$ . Then, this solution can be extended on the remaining frequencies.

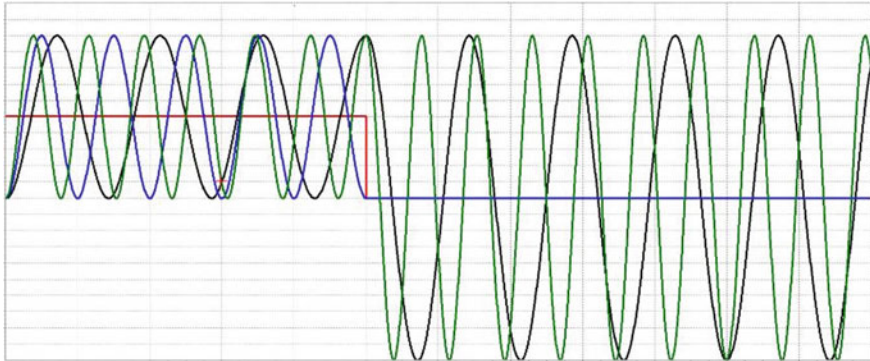
We consider Eq. (4) and set the frequency  $\omega_0^*$ , the frequency range  $\Omega_0 = [\omega_0^*(1 - \delta), \omega_0^*(1 + \delta)]$ , the set  $\Omega(F)$ , indicator function:  $I(\delta; F) = 1$ , if  $\Omega_0(\delta) \subset \Omega(F)$  and  $I(\delta; F) = 0$  otherwise. Following the EI approach, we assign two frequencies  $\tilde{\omega}_1$  и  $\tilde{\omega}_2$  ( $\tilde{\omega}_1 < \omega_0^* < \tilde{\omega}_2$ ) and construct the control  $F_2(t)$  according to algorithm (5), where the half-periods are  $\tilde{\tau}_1 = \pi\tilde{\omega}_1^{-1}$ ,  $\tilde{\tau}_2 = \pi\tilde{\omega}_2^{-1}$ . The set of controls  $\mathcal{F}$  is a two-parameter family of functions  $F_2(t; \tilde{\tau}_1, \tilde{\tau}_2)$ . The frequencies  $\tilde{\omega}_1^*$  and  $\tilde{\omega}_2^*$  should be found to provide a maximum to the functional  $\delta_{0.05}(F)$ . When constructing the maximum robust vibrationless control of multi-frequency system (1), instead of the initial set of  $n$  natural frequencies, we take a set of  $2n$  frequencies of the form  $\frac{\tilde{\omega}_1^*}{\omega_0^*}\omega_k, \frac{\tilde{\omega}_2^*}{\omega_0^*}\omega_k, k = 1, \dots, n$ , and construct the control  $F_{2n}(t)$  according to algorithm (5). The three-point solution using the two-hump EI approach is made similarly the two-point IE approach and consists in finding the three frequencies  $\tilde{\omega}_1, \tilde{\omega}_2$ , and  $\tilde{\omega}_3$ . The results of solving the problem at  $\varepsilon = 0.05$  are summarized in Table 1.

The initial control  $F_0(t)$  and reactions at different frequencies are given in Fig. 4. The maximum robust control  $F_3^*(t)$  and reactions at different frequencies are given in Fig. 5. The behavior of the dynamic coefficient and the residual spectrum in the neighborhood of the frequency  $\omega_0^*$  is presented in Fig. 6.



**Table 1** The result of solving the maximum robust control problem

Method	$\max \delta_{0,05}(F)$	Frequencies
ZV	0.03	$\omega_0^*$
Two-point method (EI)	0.18	$\omega_1 = 0.85\omega_0^*, \omega_2 = 1.15\omega_0^*$
Three-point method (two-hump EI)	0.38	$\omega_1 = 0.65\omega_0^*, \omega_2 = 0.96\omega_0^*, \omega_3 = 1.35\omega_0^*$



**Fig. 4** Initial control  $F_0(t)$  (red line) and reactions at the frequencies  $0.7 \omega_0^*$  (black line),  $\omega_0^*$  (blue line), and  $1.3 \omega_0^*$  (green line)

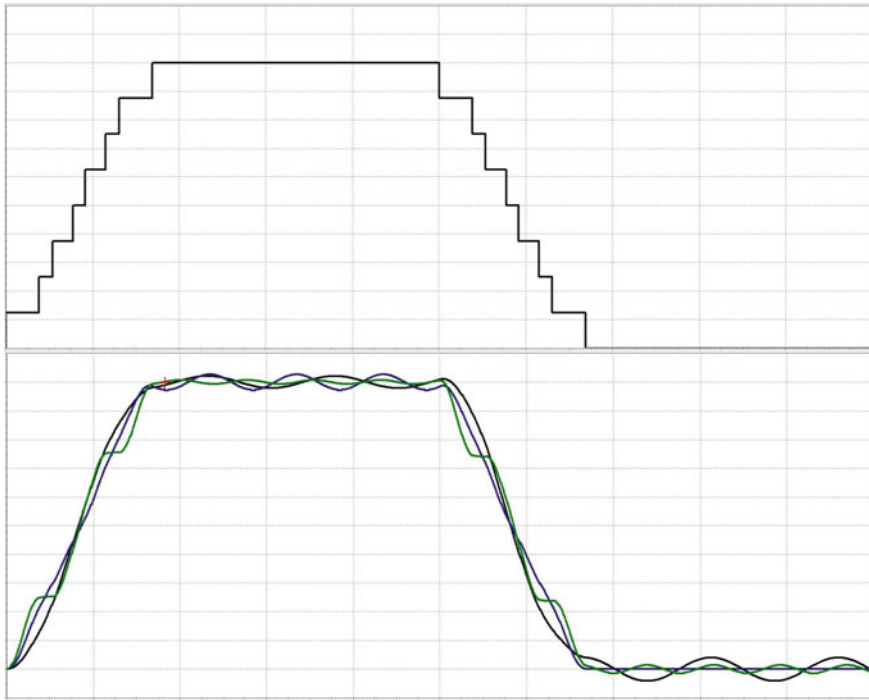
## 5 Conclusion

The obtained results proved to be useful at controlling the motion of the physical elastic structures, where oscillations are undesirable.

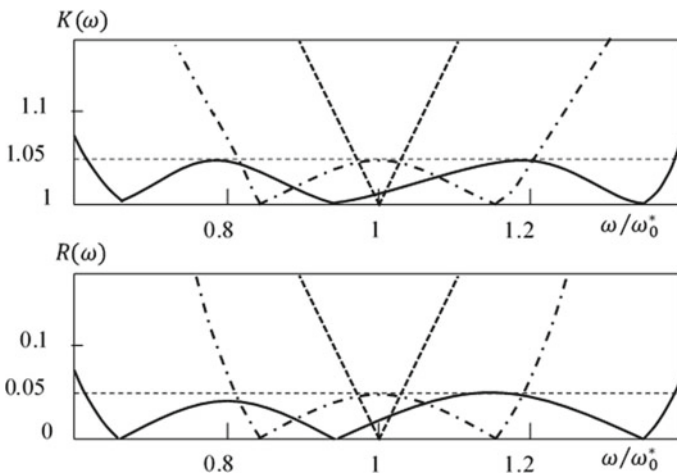
The problem of oscillation damping for the elastic structures is solved in the sense that the dynamic elastic responses are totally excluded. In this case, elastic structures undergo only static deformations that cannot be avoided at motion. This mode of motion and the corresponding control realizing it are named the vibrationless.

The problem of robust control of vibrationless movement of multi-mode flexible structures is solved. The designed robust control provided the minimum sensitivity to perturbations of eigenfrequencies.

The high complexity of the considered problems requires intensive computer modeling: in particular, for solving the problem of the choice of control algorithms with the property of maximum robustness (minimum sensitivity).



**Fig. 5** Maximal robust control  $F_3^*(t)$  and reactions at the frequencies  $0.7 \omega_0^*$  (black line),  $\omega_0^*$  (blue line), and  $1.3 \omega_0^*$  (green line)



**Fig. 6** Dynamic coefficient and the residual spectrum for maximal robust controls built by two-point method (EI) and three-point method (two-hump EI)

## References

1. Genta, G.: *Vibration Dynamics and Control*. Springer, New York (2009)
2. Akulenko, L.D.: Boundary kinematic control of a distributed oscillatory system. *J. Appl. Math. Mech.* **71**(6), 862–868 (2007)
3. Gannel, L.V., Formal'skii, A.M.: Control for minimizing vibrations in systems with compliant elements. *Journal of Computer and Systems Sciences International* **52**(1), 117–128 (2013).
4. Smith, O.J.M.: Posicast control of damped oscillatory systems. *Proc. IRE* **45**(9), 1249–1255 (1957)
5. Singhose, W.: Command shaping for flexible systems: a review of the first 50 years. *Int. J. Precis. Eng. Man.* **10**(4), 153–168 (2009)
6. Prourzin, V.A.: Control of elastic plant movement without excitation on eigen-oscillation. *Autom. Remote Control* **78**(12), 2141–2153 (2017)
7. Singh, T., Vadali, S.R.: Robust time-optimal control: a frequency domain approach. *J. Guidance Control Dyn.* **17**, 346–353 (1994)
8. Vaughan, J., Yano, A., Singhose, W.: Comparison of robust input shapers. *J. Sound Vib.* **315**, 797–815 (2008)
9. Pao, L., Lau, M.: Robust input shaper control design for parameter variations in flexible structures. *J. Dyn. Syst. Meas. Contr.* **122**, 63–70 (2000)
10. Sung, Y.G., Singhose, W.: Robustness analysis of input shaping commands for two-mode flexible systems. *IET Control Theor. Appl.* **3**, 722–730 (2009)
11. Prourzin, V.A.: Equivalent gaming formulations of the problem of designing the maximum robust controls. *Autom. Remote Control* **66**(8), 1305–1315 (2005)
12. Shevlyakov, G.L., Vil'chevski, N.O.: *Robustness in Data Analysis: Criteria and Methods*. VSP, Utrecht-Boston-Tokio (2002)
13. Shevlyakov, G.L., Vil'chevski, N.O.: *Robustness in Data Analysis*. De Gruyter, Boston (2011)
14. Shevlyakov, G.L., Oja, H.: *Robust Correlation: Theory and Applications*. Wiley, New York (2016)
15. Weaver, Jr., W., Timoshenko, S.P., Young, D.H.: *Vibration Problems in Engineering*. Wiley & Sons, New York (1990)
16. Prourzin, V.A.: A constrained scalar control for the motion of a system of oscillators with damping residual oscillations. *J. Comput. Syst. Sci. Int.* **46**(4), 521–531 (2007)
17. Prourzin, V.A.: A problem of optimal shock isolation of elastic objects. *Mech. Solids* **39**(2), 22–29 (2004)
18. Boltyanskii, V.G.: *Matematicheskie metody optimal'nogo upravleniya (Mathematical Methods of Optimal Control)*. Nauka, Moscow (1969)
19. Singhose, W., Seering, W., Singer, N.: Input shaping for vibration reduction with specified insensitivity to modeling errors. In: *Proc. Japan-USA Sym. on Flexible Automation*, pp. 307–313 (1996)
20. Park, U.H., Lee, J.W., Lim, B.D., Sung, Y.G.: Design and sensitivity analysis of an input shaping filter in the ZPlane. *J. Sound Vib.* **243**, 157–171 (2001)
21. Bhat, S.P., Miu, D.K.: Precise point-to-point positioning control of flexible structures. *ASME J. Dyn. Syst. Meas. Control* **112**(4), 667–674 (1990)
22. Singh, T., Heppler, G.R.: Shaped input control of a system with multiple modes. *ASME J. Dyn. Syst. Meas. Control* **115**, 341–347 (1993)

# Feature-Based Plant Seedlings Classification



Dmitri Jakovlev , Iuliia Kamaletdinova , and Georgy Shevlyakov 

**Abstract** The application of image features in the plant classification task is studied. The used dataset was created at Aarhus University Flakkebjerg Research. This work aims at different approaches in plant species categorization. In this study, the feature-based approach is used. It allows to perform classifying using less computational resources. The features usage is motivated by the following purpose: to distinguish weeds from other plants by selecting the defining features of all classes of plants. The proposed method combines image thresholding, feature selection, and feature extraction for the further multiclass classification by such well-known machine learning algorithms as the support vector machines, K-nearest neighbors, decision tree, and Naive Bayes. To a greater extent, we use computer vision algorithms for the image processing step. The main classification method we choose for the task is the support vector machines: It shows the best performance among other tested algorithms; the K-nearest neighbors algorithm is slightly worse.

**Keywords** Image processing · Feature extraction · Computer vision

## 1 Introduction

The demand for agricultural products is increasing day by day, as the population of the Earth is growing. Even though people work on plant classification algorithms, approaches are still not as efficient and robust as desired. A significant part of this work has still been done by people. The question arises of the efficiency with which human resources are used. We will use exhaustible natural resources wisely and increase harvests if we automatize quality assurance, which objectives are to detect and distinguish weeds among the variety of crop seedlings.

---

D. Jakovlev · I. Kamaletdinova · G. Shevlyakov (✉)  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [georgy.shevlyakov@phmf.spbstu.ru](mailto:georgy.shevlyakov@phmf.spbstu.ru)

All this naturally leads to the idea of automation of the classification process with the help of machine learning algorithms. From the recent experience, neural networks are well suited for image processing, but we have to pay for it by computational costs. On the other hand, we could use less costly algorithms, but they require finer tuning to achieve comparable results.

A feature-based method is a popular approach in image classification problems, there are many advantages and results with a good performance in many related pieces of research, as in [1, 2]. The main idea of this method is to process the only relevant information of an image, and to reduce the feature space dimension as much as possible. Likewise, dimensionality reduction is the way of solving the curse of dimensionality problem formulated by Richard Bellman in 1957 [3]. The suggested method is described in [4].

The goal of our work is to implement segmentation and classification of a specific type of datasets for lower processing time and computational complexity. In this paper, we study binary classifier capabilities on the dataset [5] consisting of images of 12 species and containing the most common weed species in Danish agriculture.

The remainder of the paper is as follows. In Sect. 2, the dataset used is described, and the methods of classification are considered. In Sect. 3, the classification results are given. In Sect. 4, some conclusions are drawn.

## 2 Materials and Methods

### 2.1 Data

The considered dataset is a part of the database that has been recorded at the Aarhus University Flakkebjerg Research station in the collaboration between the University of Southern Denmark and Aarhus University. Images are available to researchers at <https://vision.eng.au.dk/plant-seedlings-dataset/>. The specifics of this dataset is that recorded plants are in different growth stages since detecting a weed in its early stage is the thing which makes the task problematic.

The dataset contains 960 unique plant images of 12 species. The sizes of plant classes are not balanced among themselves—they range from 221 to 654 labeled samples for each class. Original images are cropped by plant boundaries, but their resolutions vary from  $50 \times 50$  px to  $2000 \times 2000$  px. Also, images have different backgrounds—some of them are on the ground, other are on the marked paper.

## 2.2 Data Preprocessing

**Resolution reducing.** We reduce the resolution of all images to the same resolution  $200 \times 200$  px using the bilinear interpolation. The main idea of the bilinear interpolation is that a new image pixel is defined as the weighted sum of neighboring pixels of the original image. It helps to decrease computational complexity and build normalized features [6].

**Segmentation.** The objects of our study are plants, and they are painted green. Therefore, we can create a mask that filters the range of green channel and ignores the other pixels. For these purposes, the hue saturation value (HSV) color model is a suitable representation [7]. In the blue green red (BGR) format, the value of each component depends on the amount of light hitting the object. HSV allows us to distinguish between the image color and brightness. We set the lower and upper bounds of the green color using HSV representation. Then we merely mark the pixels in the green range and get a color mask (Fig. 2b). Now we apply the operation of logical multiplication to the original image, assign the value of the background pixels to a black color value, and get a segmented plant (Fig. 1).

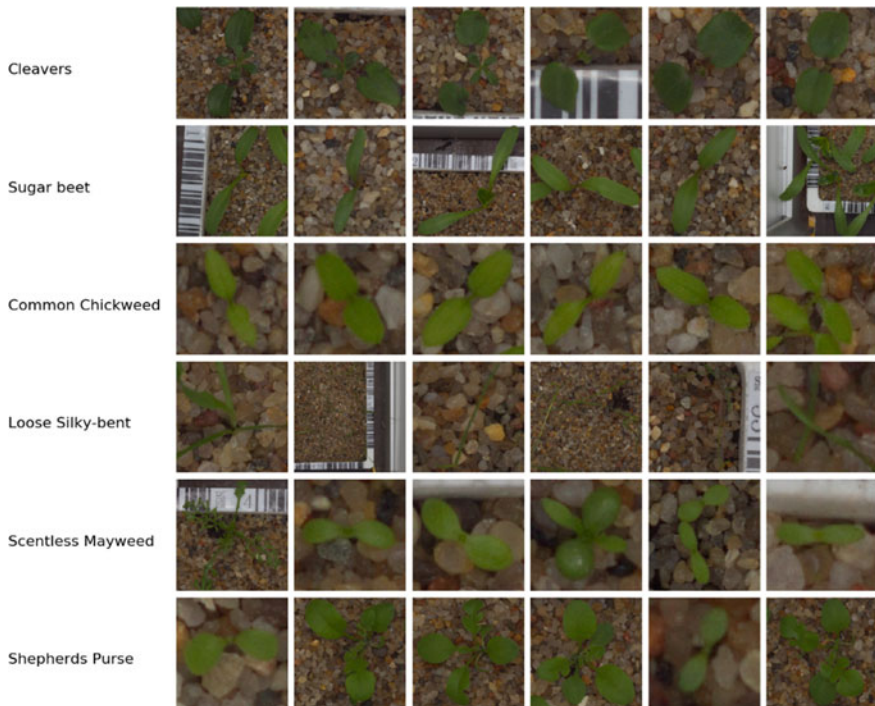
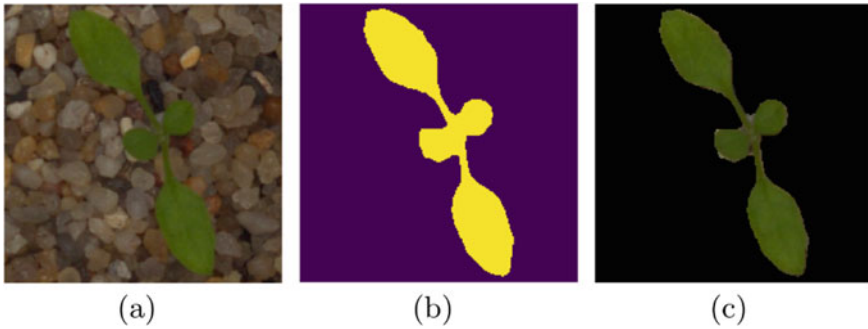
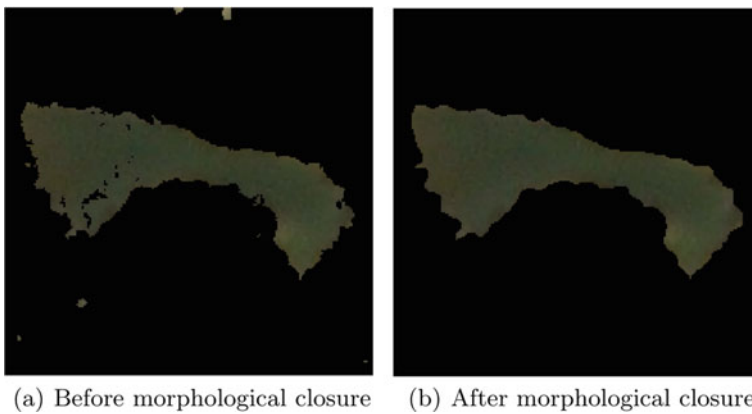


Fig. 1 Data overview



**Fig. 2** **a** Source image, **b** mask, **c** segmented image



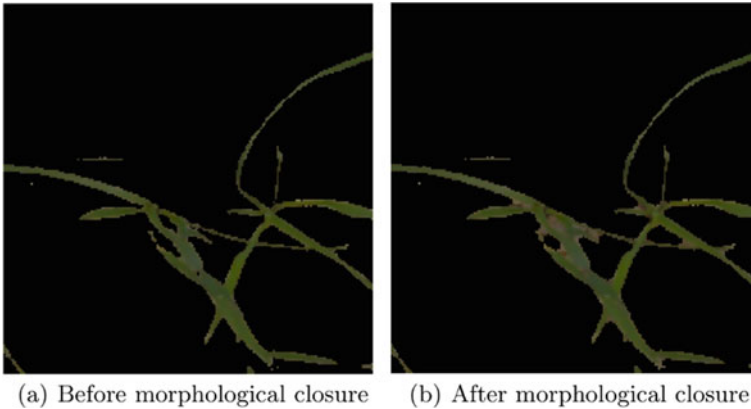
**Fig. 3** Segmentation improvement example

**Denoising.** Segmentation does not always work well (see Fig. 2c). Small areas of the background may fall into the range of green values which distorts the binary mask as in Fig. 3a.

Such drawbacks can be eliminated by the morphological operations of the nonlinear transformation associated with the shape and structure of an image. The morphology is used to study the interaction of an image with a specific structural element, the kernel. The kernel iterates over the entire image and compares the neighborhoods of pixels after which we apply morphological operations [8].

To improve segmentation, we use the operation of the morphological closure—a combination of the dilatation and erosion operations [9].

Then we apply the morphological closure operation to the image (Fig. 4a) by selecting an elliptical core of  $6 \times 6$  px size and delete the remaining objects within an area of less than 160 px. The plant on the image (Fig. 4b) has no cavities, and the background is cleared of non-plant elements. But the morphological closure does not always improve the segmentation result. Estimate the result of processing an image



**Fig. 4** Segmentation degradation example

of the class loose silky-bent in Fig. 4: The cavities corresponding to the background were restored that does not correspond to the desired result. We restrict ourselves to the removal of objects whose contours limit a small area.

### **2.3 Feature Selection**

The features of images define their content. Consideration of a great number of features helps us to recognize the information better. Image features let the classifier propose the output decision. Another advantage of the approach is that it reduces feature space for a machine learning algorithm. We often need only a part of the information on the image, hence we do not need to process and evaluate all the pixels, which can cause additional computational expenses.

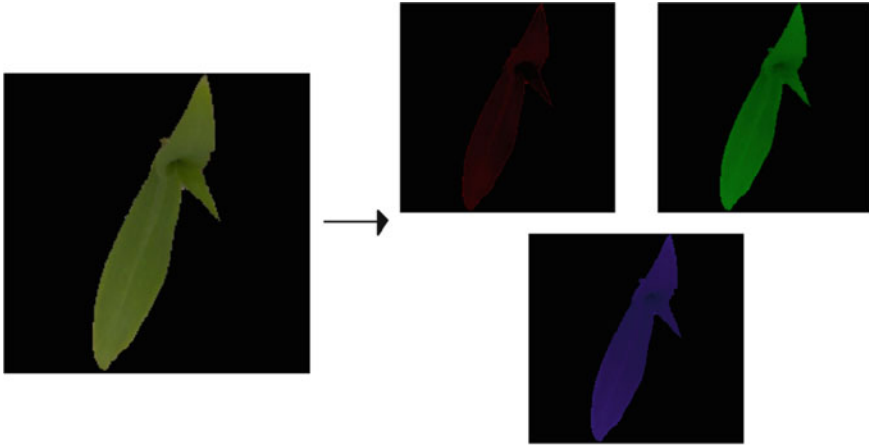
Selecting features is a complicated and convoluted research area itself, the statement is supported by the variety of feature types, and the need of presenting essential properties on an equal basis with the previous assertion.

The goal is to define the set of features describing the dataset in the best way. Supposed features must satisfy the following criteria:

1. The feature space should be low-dimensional
2. The features should not correlate or correlate as little as possible
3. Selected features should represent the content of an image as fully as possible

Now we are going to define the selected features.





**Fig. 5** RGB transformation

## 2.4 Color Features

Overviewing the dataset, we notice that all the plant species are mostly green. Additionally, their images are recorded under specific conditions. We use the RGB color model, which stands for red, green, and blue colors, and calculate features described below (Fig. 5).

Let  $\{x_i^{(k)}\}_{i=1}^N$ , where  $k = 1, 2, 3$  is an index of a channel in the RGB color space, respectively;  $N$  is a total number of the image pixels;  $x_i^{(k)}$  is an  $i$ -th pixel of the  $k$ -th channel. Next, compute the sample mean and standard deviation for each channel:

$$\begin{aligned}\overline{x^{(k)}} &= \frac{1}{k} \sum_{i=1}^N x_i^{(k)}, \\ S^{(k)} &= \sqrt{\frac{1}{N} \sum_{i=1}^N \left(x_i^{(k)} - \overline{x^{(k)}}\right)^2}.\end{aligned}$$

## 2.5 Shape Features

A widely used approach to retrieve shape features is to detect and analyze bounding contours. Here we use the boundary tracing algorithm for the boundary extraction. The designated algorithm [10] is implemented in the OpenCV [11] library for the Python programming language. The studies do not take into account the contours bounding areas below a certain threshold, which is empirically chosen.

Let  $K$  be a number of detected bounding contours above the threshold in the further contour-related characteristics.

**Total perimeter.** For this feature, we count the sum of perimeters of all the areas bounded by contours:

$$P = \sum_{i=1}^K p_i,$$

where  $p_i$  is an  $i$ -th perimeter.

**Entire area.** It includes all the areas bounded by contours:

$$S = \sum_{i=1}^K s_i,$$

where  $s_i$  is an  $i$ -th area.

**Maximal contour area.** Here we analyze the contours bounding maximal areas:

$$S_m = \max s_i, i = 1, \dots, K.$$

**Rectangularity.** One of the methods to estimate rectangularity is to plot minimum bounding rectangle. Rectangularity is the ratio of the entire object area to the minimum bounding rectangle area. This feature represents how rectangular an object is:

$$f_{\text{rect}} = \frac{S}{S_{\text{MBR}}},$$

where  $S$  is the entire area,  $S_{\text{MBR}}$  is the minimum bounding rectangle area.

**Circularity.** Another title of this shape factor is the isoperimetric quotient, and it shows how much area per perimeter is bounded:

$$f_{\text{circ}} = \frac{4\pi A}{P^2},$$

where  $P$  is an entire perimeter;  $A$  is an entire area of all detected elements of a plant.

The correlation matrix of the described features has the form:

Based on the data in Table 1, we conclude that the most linearly dependent features are the entire area and the largest area. This is not true for all classes due to the predominance of plants bounded by the only one contour. Therefore, the largest area feature is not rejected.

**Table 1** Feature correlation matrix

	area	largest_area	perimeter	aspect_ratio	circularity	mean_r	mean_g	mean_b	stddev_r	stddev_g	stddev_b	skew_r	skew_g	skew_b
area	1	0.89	0.25	-0.24	-0.36	-0.0072	0.28	-0.31	-0.071	0.096	0.063	-0.11	-0.24	-0.055
largest_area	0.89	1	0.42	-0.19	-0.23	0.084	0.28	-0.27	-0.00094	0.19	0.19	-0.17	-0.22	-0.17
perimeter	0.25	0.42	1	-0.12	0.59	-0.046	0.06	-0.038	0.29	0.36	0.3	-0.046	0.053	-0.14
aspect_ratio	-0.24	-0.19	-0.12	1	0.026	0.057	-0.037	0.11	0.045	-0.016	0.0031	-0.0074	0.048	-0.0021
circularity	-0.36	-0.23	0.59	0.026	1	-0.072	-0.17	0.17	0.31	0.21	0.15	0.023	0.17	-0.056
mean_r	-0.0072	0.084	-0.046	0.057	-0.072	1	0.83	-0.18	-0.009	0.24	0.41	-0.3	-0.32	0.098
mean_g	0.28	0.28	0.06	-0.037	-0.17	0.83	1	-0.28	-0.063	0.31	0.39	-0.22	-0.45	0.064
mean_b	-0.31	-0.27	-0.038	0.11	0.17	-0.18	-0.28	1	0.19	-0.013	-0.25	0.15	0.38	-0.38
stddev_r	-0.071	-0.00094	0.29	0.045	0.31	-0.009	-0.063	0.19	1	0.71	0.55	-0.005	0.39	0.095
stddev_g	0.096	0.19	0.36	-0.016	0.21	0.24	0.31	-0.013	0.71	1	0.58	-0.41	-0.077	-0.049
stddev_b	0.063	0.19	0.3	0.0031	0.15	0.41	0.39	-0.25	0.55	0.58	1	-0.12	-0.022	0.035
skew_r	-0.11	-0.17	-0.046	-0.0074	0.023	-0.3	-0.22	0.15	-0.005	-0.41	-0.12	1	0.55	0.34
skew_g	-0.24	-0.22	0.053	0.048	0.17	-0.32	-0.45	0.38	0.39	-0.077	-0.022	0.55	1	0.14
skew_b	-0.055	-0.17	-0.14	-0.0021	-0.056	0.098	0.064	-0.38	0.095	-0.049	0.035	0.34	0.14	1

## 2.6 Classification

The main method for solving this task is the support vector machine (SVM) [12], a binary classification algorithm based on building a separating hyperplane. The other methods we apply are the K-nearest neighbors [13], Naive Bayes [14], and decision tree [15] classifiers. These algorithms are implemented in the scikit-learn [16] library for the Python programming language.

We use the radial basis function (RBF) as the kernel function for SVM. This choice is made because the RBF allows to build a hyperplane when the data is not linearly separable.

**Data normalization.** The SVM algorithm is sensitive to non-normalized data, especially when using the RBF kernel, which is just the Euclidian distance. In the case when the feature values are at different intervals, a slight difference in one of them can lead to going out of range in second feature values. The solution is to map all the values into one segment. In this task, we choose the segment [0,1].

## 3 Results

### 3.1 Metric

Results of classification are evaluated by the micro-averaged F-score. Given the positive and negative rates for each class, the resulting score is computed as follows:

$$\text{Precision}_{\text{micro}} = \frac{\sum_{k \in c} TP_k}{\sum_{k \in c} TP_k + FP_k}, \text{Recall}_{\text{micro}} = \frac{\sum_{k \in c} TP_k}{\sum_{k \in c} TP_k + FN_k},$$

**Table 2** Detailed metrics for the SVM classification

Type	Precision	Recall	F-score
Sugar beet	0.90	0.93	0.91
Fat hen	0.87	0.90	0.89
Scentless mayweed	0.85	0.90	0.87
Charlock	0.95	0.93	0.94
Small-flowered cranesbill	0.96	0.99	0.97
Maize	0.95	0.89	0.92
Shepherds purse	0.83	0.71	0.77
Common wheat	0.80	0.89	0.84
Common chickweed	0.96	0.96	0.96
Cleavers	0.89	0.85	0.87
Loose silky-bent	0.83	0.89	0.86
Black-grass	0.76	0.53	0.62
Micro-averaged F-score	0.88		

$$F_{\text{micro}} = \frac{2\text{Precision}_{\text{micro}}\text{Recall}_{\text{micro}}}{\text{Precision}_{\text{micro}} + \text{Recall}_{\text{micro}}}$$

where  $C$  is a set of the plant classes.

The choice of such a metric is supported by the fact that classes are imbalanced. In this case, the influence of classes decreases due to averaging by classification characteristics, not by F-scores.

The classification output is shown in Table 2. The worst classification result is received for the black-grass class. This type of plant is difficult to segment on an equal basis with the loose silky-bent class (Fig. 4), for which the result is significantly better. The black-grass plants have purple roots, but the segmentation algorithm we use does not work properly in the purple color values range. Another cause of this result is in the predominance of the loose silky-bent training data size. One of the highest scores belongs to the common chickweed class (Fig. 2). The segmentation algorithm demonstrates good results on these plants, because there is a definite green color values range, which defines the entire plant. Additionally, plants of this class cover a sizable area. It decreases the chance of treating the parts of the target object as the parts of the background.

### 3.2 Models Comparison

The choice of the SVM is justified by better results in comparison with other classical methods of machine learning. Table 3 shows micro-averaged F-scores for the Naive Bayes, K-nearest neighbors and decision tree classifiers. The Naive Bayes shows the

**Table 3** Metrics for classification

Method	Micro-averaged F-score
Naive Bayes	0.72
kNN	0.84
Decision tree	0.73
SVM	0.88

worst results because it is sensitive to the correlation between features. The decision tree performs vastly worse than SVM, because we use the RBF kernel in SVM. This effect is called “kernel trick” [17], it allows us to work in a transformed space, where the data is linearly separable. Since the k-nearest neighbors algorithm is insensitive to nonlinear data, the result is as good as SVM. These methods are implemented in scikit-learn library. All the experimental results are obtained using cross-validation technique [18].

## 4 Conclusion

In this paper, we apply the feature-based method to the image classification task. The constructed algorithm is implemented and evaluated on the real plant dataset containing images of 12 different types of seedlings. We select and extract features using computer vision algorithms. As it is shown in Table 3, the best performance is reached with the support vector machines algorithm. The detailed result for the best method is shown in Table 2. Some of the classes are not recognized well because of the minor differences between some types of proposed plants. The future work aims at improving segmentation output and at the usage of other types of image features.

## References

1. Shi, L., Wan, Y., Gao, X., Wang, M.: Feature selection for object-based classification of high-resolution remote sensing images based on the combination of a genetic algorithm and tabu search. *Comput. Intell. Neurosci.* **2018**, 1–13 (2018). <https://doi.org/10.1155/2018/6595792>
2. Mingqiang, Y., Kidiyo, K., Joseph, R.: A survey of shape feature extraction techniques. In: *Pattern Recognition*, chap. 3. IntechOpen, Rijeka (2008). <https://doi.org/10.5772/6237>
3. Bellman, R.: *Dynamic Programming*, 1st edn. Princeton University Press, Princeton, NJ, USA (1957)
4. Pechenizkiy, M., Puuronen, S., Tsymbal, A.: Feature extraction for classification in knowledge discovery systems. In: *Knowledge-Based Intelligent Information and Engineering Systems*, pp. 526–532. Springer, Heidelberg (2003)
5. Mosgaard Giselsson, T., Nyholm Jørgensen, R., Kryger Jensen, P., Dyrmann, M., Skov Midtiby, H.: A Public Image Database for Benchmark of Plant Seedling Classification Algorithms. *arXiv e-prints arXiv:1711.05458* (2017)

6. Davis, P.J.: Interpolation and Approximation, 1st edn. Dover Publications, New York, USA (1963)
7. Smith, A.R.: Color gamut transform pairs. In: Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques, pp. 12–19. SIGGRAPH'78. ACM, New York, USA (1978). <https://doi.org/10.1145/800248.807361>
8. Friel, J.J.: Practical Guide to Image Analysis. ASM International (2000)
9. Efford, N.: Digital Image Processing: A Practical Introduction Using Java (with CD-ROM), 1st edn. Addison-Wesley Longman Publishing Co. Inc., Boston, MA, USA (2000)
10. Suzuki, S., Abe, K.: Topological structural analysis of digitized binary images by border following. *Comput. Vis. Graph. Image Process.* **30**, 32–46 (1985)
11. Bradski, G.: The opencv library. *Dr. Dobb's J. Softw. Tools* **25**, 120–125 (2000)
12. Hsu, C.W., Chang, C.W., Lin, C.J.: A practical guide to support vector classification. Tech. rep., Department of Computer Science and Information Engineering, University of National Taiwan, Taipei (2003). <https://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf>
13. Cunningham, P., Delany, S.: k-nearest neighbour classifiers. *Multiple Classifier Syst.* **34**, 1–17 (2007)
14. Rish, I.: An empirical study of the naive Bayes classifier. In: *IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence*, vol. 3, pp. 41–46. IBM, New York (2001)
15. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA (1984)
16. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011)
17. Hofmann, T., Schölkopf, B., Smola, A.: Kernel methods in machine learning. *Ann. Stat.* **36**(3), 1171–1220 (2008)
18. Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2nd edn (2009)

# Medical Training Simulation in Virtual Reality



Vladimir Ivanov , Sergey Strelkov , Alexander Klygach ,  
and Dmitry Arseniev 

**Abstract** Digital technologies have a significant impact on medical training. Growing popularity of virtual and augmented reality changes the trend into virtual simulators. We gave an overview of key market products in the field of medical simulators in order to define main aspects for development of our own system. These aspects are: open surgery, realistic visualization, and haptic feedback. We described in details each of them and how it was implemented in our system. For open surgery was used appendectomy as most common procedure of this type of the surgery. In order to achieve realistic visualization, we implemented three different approaches for creating realistic and accurate 3D models. For haptic feedback, we took Novint Falcon and enhanced it with our custom grip which provides additional degrees of freedom.

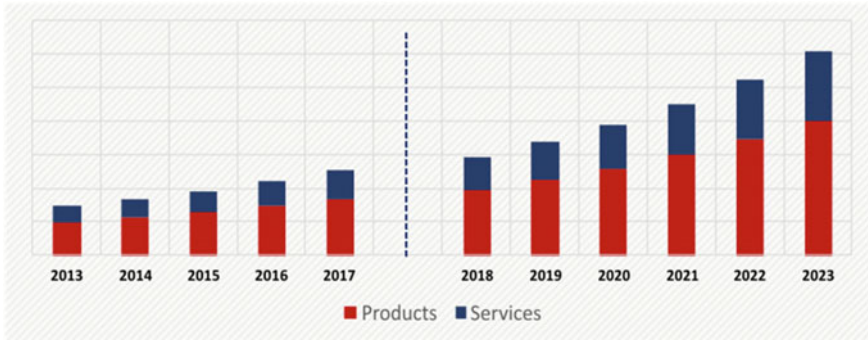
**Keywords** Surgical simulator · Virtual reality · Real-time rendering · 3D visualization · 3D reconstruction · Haptic feedback · Open surgery · Laparoscopy

## 1 Introduction

Medical simulation is changing rapidly since the beginning of twentieth century. This is happening due to a number of reasons. First of all, modern approaches for less invasive surgery redefined the way how surgical procedure is happening, for example endoscopy and robotics surgery [1]. Secondly, dramatic rise of computing power gave an opportunity to implement complex simulations in real time. Finally, new accurate algorithms for a rigid and soft body simulation, realistic 3D visualization, haptic controllers, and virtual reality turn medical simulation from physical models to an area of digital games [2].

---

V. Ivanov (✉) · S. Strelkov · A. Klygach · D. Arseniev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [voliva@rambler.ru](mailto:voliva@rambler.ru)



**Fig. 1** Worldwide surgical simulation market by offering (2013–2023)

## 2 Modern Medical Simulation Market

### 2.1 Market Overview

According to Prescient & Strategic Intelligence data [3], the global surgical simulation market was valued at \$254.7 million in 2017 with clear growing trend and forecasted to increase its value twice in 2023 (Fig. 1).

Another notable trend mentioned in analysis is using augmented reality (AR) and virtual reality (VR) to enhance quality and efficiency of medical training. Thus, from this data we can see that market will continue to grow, where digital technologies will have a major impact.

### 2.2 Medical Simulators

**VirtaMed.** Primarily this company focused on simulator development for orthopedics, gynecology, and urology. Such surgical simulators designed on a single flexible platform with ability to expand and add additional procedures in the future [4]. All simulators were combined with an anatomical model to provide the best tactile feedback and manipulate like in a real life. In addition, for better efficiency each virtual procedure build with guided training, which has specific colored hints and ghost tools to show trainees how to perform different tasks (Fig. 2).

**NeuroVR.** This is a platform for neurological training and enables neurosurgeons to practice skills with a help of virtual reality (Fig. 3). Such system does not depend on real-life models and uses haptic controllers to manipulate in VR. All range of exercises derived from actual patient images, which provides more realistic and accurate image of surgical procedure. The system also captures objectives metrics and measure proficiency in procedures to track a progress in education [5].



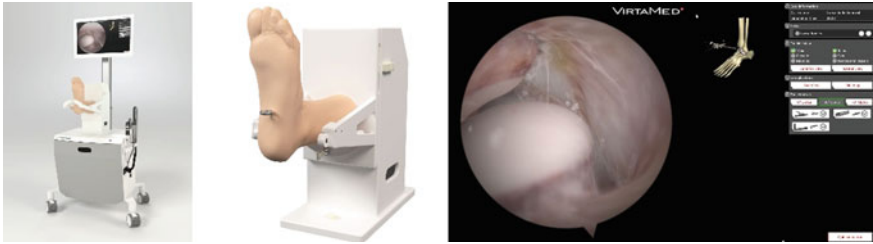


Fig. 2 ArthroS Ankle by VirtaMed AG

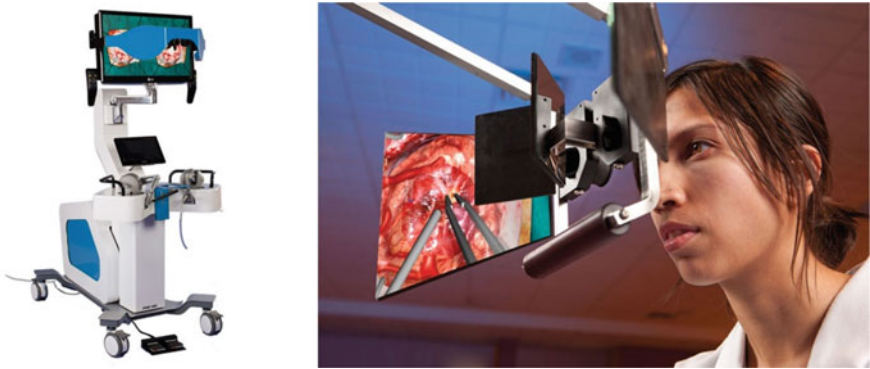


Fig. 3 NeuroVR system with stereoscopic microscopic view

**Surgical Science.** This company focused in development of various simulation products, mostly for laparoscopy and endoscopy [6]. LapSim is the key product designed to improve psychomotor skills using virtual reality with haptic feedback (Fig. 4). It features different modules for laparoscopic exercise that arrange from

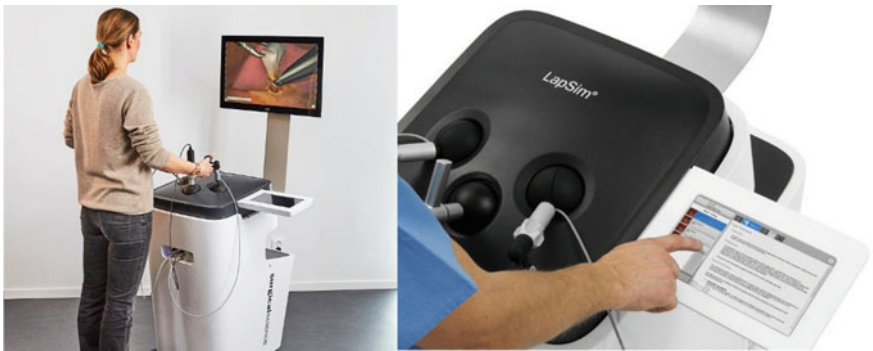
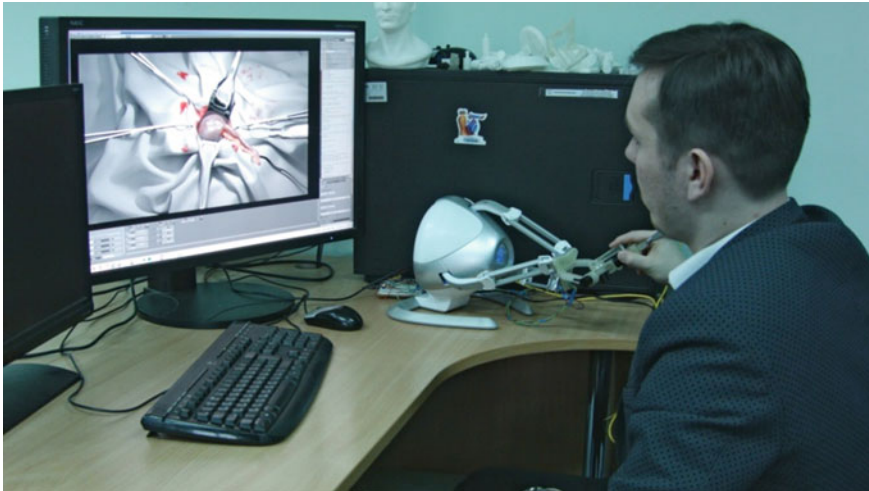


Fig. 4 LapSim with in-house developed haptic system



**Fig. 5** Surgery simulator based on open appendectomy with haptic feedback

navigation to suturing. This system also has a portable version “LapSim essence” [7].

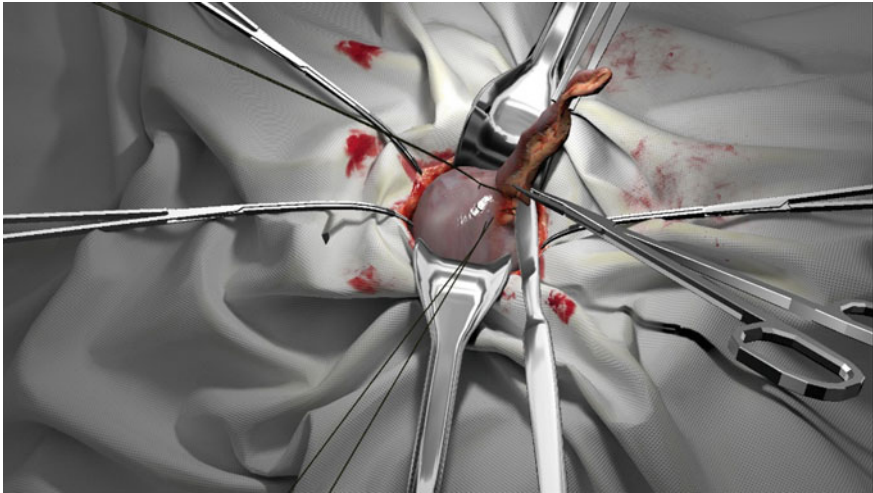
### **3 Our Approach in the Development of Medical Simulation**

#### **3.1 Key Aspects**

Many different surgical simulators that available now on the market have equal disadvantages like simplified 3D visualization and most of all designed for specific surgical approaches, like endoscopy. On the other hand, only few related to open surgery. For this reason, we focused on open surgery simulator for appendectomy [8] with realistic visualization using haptic feedback (Fig. 5).

#### **3.2 Realistic Visualization**

Despite advanced real-time rendering solutions that are available now, it is still difficult to produce realistic image in surgical simulator. This is happening by various reasons due to software limitation. One of them is using built-in graphics solution. For this purpose, we build our system based on modern game engine which allows using physically based shader models [9] and enhances it with our custom physics engine to work with soft tissue (Fig. 6).



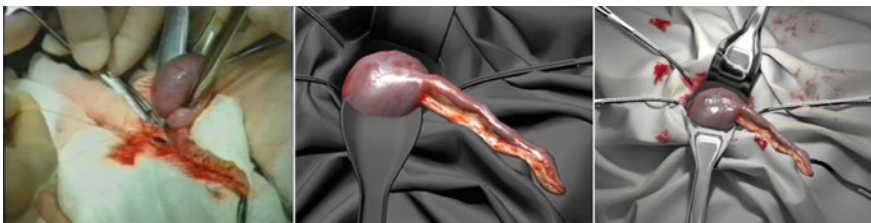
**Fig. 6** Stage, where the mesoappendix is dissected

### 3.3 *Creating Models Based on Patient’s Data and Anatomical Atlases*

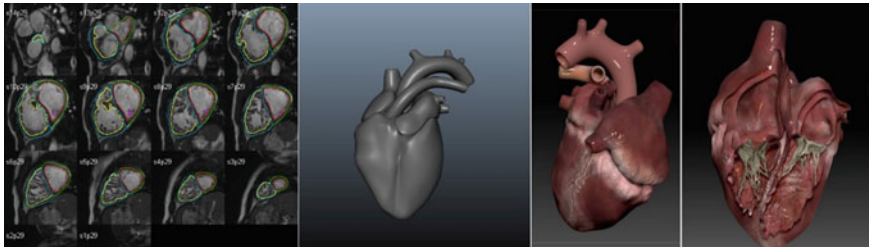
There are few different methods which can be used to produce more accurate result. One of them is taking photographs from real surgical procedure and extracting textures [10] from these images (Fig. 7). It is also possible to build models based on sequence of photographs using photogrammetry solution [11].

Another way is building models based on MRI or CT data [12]. For instance, we can recreate heart model with specific pathology with a help of MRI and contouring data [13]. This achieved through multiple stages (Fig. 8). At first, we built model of ventricles from counterung data and then projected heart master model onto them and finally added textures based on real-life heart images.

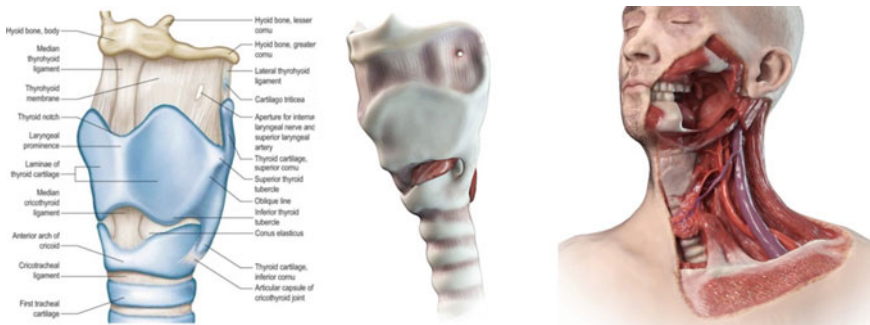
In complex cases, where photography or tomography is not enough [14] to build full model, we are using anatomical atlases. In case of anatomical section to reveal



**Fig. 7** Photograph from surgery procedure (left), appendix model with textures (center), overall view in surgical simulator (right)



**Fig. 8** MRI with contouring data (left), reconstructed heart 3D model (center), heart 3D model with textures and internal structures (right)



**Fig. 9** Larynx image from Gray's anatomy atlas (left), 3D model of larynx (center), neck section with larynx (right)

larynx position, we combined knowledge and data from multiple atlases like Gray's anatomy [15] and 3D atlas of human body. This helped accurately showcase larynx in comparison with other anatomical structures like skull and muscles (Fig. 9).

### 3.4 Haptic Feedback

In order to achieve fully realistic visualization, we should supplement our system with haptic feedback [16]. Not only does it help to sense virtual 3D objects, but also it allows developing right psychomotor skills for surgeons [17].

To implement this technology, we took Novint Falcon haptic device [18] as a basis. Despite this tool was designed primarily for games, it provides accurate haptic feedback with three degrees of freedom (DOF). This also makes it cheapest haptic device on the market, which reduces overall cost of simulator significantly.

In order to use this device as a tool for surgical simulator, some tweaks and encasement are important. One of them was custom-designed grip with extra three DOF [19] for tilting surgical tool (Fig. 10). This tool is based on absolute hall encoders



**Fig. 10** Custom-designed grip to provide extra three-degrees of freedom

[20] and transmits data as separate stream through digital-to-analog converter [21]. In addition, it has a slot for swapping different surgical instruments.

## 4 Conclusion

There are a lot of laparoscopy surgical simulators which are oversaturating market. Open surgery simulators in this case can take a niche on the market and became popular. With help of modern technologies, it became possible to implement such complex simulators. Described solutions for creating 3D models and haptic feedback will help to achieve more realistic result in a cost-efficient way.

## References

1. Kevin, K.: The role of medical simulation. *The Int. J. Med. Robot. Comput. Assist. Surg.* **2**, 203–210 (2006)
2. Frederick, K., Craig, G., Ananda, S.: The role of medical simulation. *BMC Med. Educ.* **10**(50), 1–10 (2010)
3. Surgical Simulation Market: <https://www.psmarketresearch.com/market-analysis/surgical-simulation-market>. Last accessed 2019/10/05
4. Reasons why your center needs a VirtaMed simulator, <https://www.virtamed.com/en/medical-training-simulators/10-reasons-why/>. Last accessed 2019/10/05.
5. NeuroVR Brochure: <https://caehealthcare.com/media/files/TechSheets/NeuroVR-TechSheet.pdf>. Last accessed 2019/10/05.
6. Surgiscience about us page, <https://surgiscience.com/about-us/our-story/>. Last accessed 2019/10/05.
7. Surgiscience LapSim essence page, <https://surgiscience.com/systems/lapsim/lapsim-essence/>. Last accessed 2019/10/05.
8. Sauerland, S., Jaschinski, T., Neugebauer, E.: Laparoscopic versus open surgery for suspected appendicitis. *Cochrane Database Syst. Rev.* (10), 2010.
9. The PBR guide by Allegorithmic—Part 1, <https://academy.substance3d.com/courses/the-pbr-guide-part-1>. Last accessed 2019/10/05.

10. Chen, R., Lu, D., Pan, Y.: Generating Textures of irregular objects from models and photo sequences. *J. Image Graph.* **8** (2003).
11. Petros, P.: *Virtual Prototyping & Bio Manufacturing in Medical Applications*, pp. 46–48. Springer Science (1999).
12. Ylä-Anttila, P., Vihinen, H., Jokitalo, E., Eskelinen, E.L.: 3D tomography reveals connections between the phagophore and endoplasmic reticulum. *Autophagy* **5**(8), 1180–1185 (2009)
13. Strelkov, S., Klygach, A., Ivanov, V.: Creating Hi-detailed heart 3d model based on MRI and contour data and it's representation in augmented reality. *Int. J. Recent Technol. Eng.* **7**(6S5), 254–257 (2019).
14. Keevil, S.F., Gedroyc, W., Gowland, P., Hill, D.L.G., Leach, M.O., Ludman, C.N., McLeish, K., McRobbie, D.W., Razavi, R.S., Young, I.R.: Electromagnetic field exposure limitation and the future of MRI. *The Br. J. Radiol.* **78**(935), 973–973 (2005)
15. Standring, S. (ed.): *Gray's Anatomy E-book: The Anatomical Basis of Clinical Practice*, pp. 586–604. Elsevier Health Sciences (2015).
16. Okamura, A.: Methods for haptic feedback in teleoperated robot-assisted surgery. *Ind. Robot: An Int. J.* **31**(6), 499–508 (2004)
17. Mahmoud, M., Mahmoud, M., Mahmoud, A.: Students' evaluation of a 3DVR haptic device Simodont. Does early exposure to haptic feedback during preclinical dental education enhance the development of psychomotor skills? *Int. J. Dent. Clin.* **6**(2), (2004).
18. Karbasizadeh, N., Aflakiyan, A., Zarei, M.: Dynamic identification of the Novint Falcon Haptic device. In: 4th International Conference on Robotics and Mechatronics, pp. 634–637. IEEE (2016).
19. Knowles, G.J., Mulvihill, M., Uchino, K., Shea, B.: Solid state gimbal system. U.S. Patent 7,459,834, issued December 2, 2008.
20. Shuanghui, H., Yong, L., Minghui, H.: Study on a novel absolute magnetic encoder. In: IEEE International Conference on Robotics and Biomimetics. IEEE, pp. 1773–1776 (2008).
21. Shanmugam, K.S.: Digital and analog communication systems. NASA STI/Recon Technical Report A. 1979;80.

# Application of the Hybrid Model to Numerical Modeling of the Urban Transport Network Topology



Vadim Glazunov , Mikhail Chuvatov , Leonid Kurochkin ,  
Mikhail Kurochkin , Alexander Chernyshev , and Leonid Hanin 

**Abstract** Improving road safety, monitoring the vehicles, and controlling their routes require rapid development of intelligent transportation systems (ITS). The main tool for ITS design is computer modeling. Because such modeling is based on either micro-level or macro-level only, the existing models of transportation systems often represent a trade-off between modeling accuracy and computation time. Hybrid models of transportation networks, allowing one to simulate different parts of the network using micro- and macro-models combined with the ability to synchronize them, are currently not available. In the present work, we justify the need for hybrid model development and suggest a way to have a transition between micro- and macro-levels of transportation network model. The paper contains the results of modeling a segment of urban road network with mixed topology and reports computation time required to simulate the traffic behavior in hybrid model. The proposed solution was implemented using the SUMO microscopic simulator of transportation networks and the original continual model of traffic flows. The suggested approach allows one to greatly reduce the computation time required for modeling the traffic flows on large areas without affecting the accuracy.

**Keywords** Hybrid transport network model · Continual flow model · Transport networks · Macro-model · Micro-model · SUMO · TraCI · Discrete simulator · Mathematical model · Navier–Stokes equations · Modeling transport flow · Intelligent transport systems

---

V. Glazunov (✉) · M. Chuvatov · L. Kurochkin · M. Kurochkin  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [glazunov\\_vv@spbstu.ru](mailto:glazunov_vv@spbstu.ru)

A. Chernyshev  
Ioffe Institute, St. Petersburg, Russia

L. Hanin  
Idaho State University, Pocatello, USA

## 1 Introduction

In recent years, researchers have been particularly interested in the creation and use of intelligent transport networks (ITS), the key purpose of which is to improve road safety, provide drivers and passengers with access to network information services [1], and to monitor and control urban and regional traffic.

The main means of communication used in ITS are global and local wireless data networks. Each ITS participant, a vehicle or an element of ITS infrastructure, can act as a receiver, transmitter, or data repeater. The process of routing data between mobile subscribers of ITS telecommunication networks in local (mesh) networks does not allow organizing continuous reliable access to ITS services.

The main tool used in the design of ITS is computer simulation. Investigation of the dependencies between the properties of road and network traffic is of additional interest. These studies are necessary to build adaptive predictive models that allow one to quickly evaluate the load on the transport and data transmission networks and to plan reliable data transmission to ITS taking into account its dynamics.

## 2 The Main Approaches to the Construction of Transport Network Models

There are two main approaches to the modeling of road traffic, which are based on the use of microscopic and macroscopic models. Microscopic models [2] describe the individual behavior of each road user through a set of mechanistic parameters including the response time of the driver to changes in traffic conditions, the dependence of accelerations, etc. [3]. Discrete microscopic models are the most detailed but at the same time the most demanding on computing resources [4]. An alternative way is to use continual models of road traffic, i.e., macroscopic models.

In constructing such models, an analogy is used with the motion of a liquid or a gas in the framework of a continual-medium approximation where properties of individual molecules are translated into local parameters of the medium. For such a translation, the procedure of averaging over a control volume is used, and the volume is chosen large enough for a large number of molecules to enter it, however much smaller than the linear size of the area [5]. In the case of road networks, it can be argued that the continual approach is applicable on long sections of roads with high traffic density.

One of the first continual mathematical models of road traffic used only the mass balance equation [6]. The main feature of such models was the appearance of non-physical jumps in traffic compression. To eliminate this drawback, the initial equations are reformulated as the momentum conservation equation with additional source terms [7], taking into account the effect of the pressure gradient and relaxation of the velocity field to a certain equilibrium value [8]. The expression for pressure describes the mutual influence of neighboring cars and reflects the driver's behavior



in dense and sparse automobile flows so that with increasing density the pressure also increases.

The road traffic hybrid model developed in [9] combines the properties of microscopic and macroscopic models thus avoiding the constant calculation of the characteristics of all ITS participants based on a description of the road conditions using the average properties of vehicle flows. The use of hybrid models of road traffic will reduce the requirements for the necessary computational resources and thereby enable an implementation of models of urban transport networks and agglomerations required to study the dependencies of road and network traffic.

### 2.1 SUMO—Discrete Micro-level Simulator

One of the popular microscopic transport network simulators [10] is Simulation of Urban MObility (SUMO), a simulator developed at the Institute for Transport Research (Institute of Transportation Systems, Germany). It implements a microscopic discrete road traffic model and provides an open-source framework for modeling traffic flows through road networks. SUMO implements modeling at the microscopic level, providing an individual description of the properties of each road user. The SUMO model allows one to take into account the mutual influence of road users thereby increasing the likelihood of accident-free movement.

Traffic Control Interface (TraCI), a simulator module that provides an application program interface for traffic modeling, allows one to retrieve various characteristics of simulated objects and manipulate their behavior in real time. TraCI uses client-server network architecture to interact with the SUMO. The model configuration diagram is shown in Fig. 1.

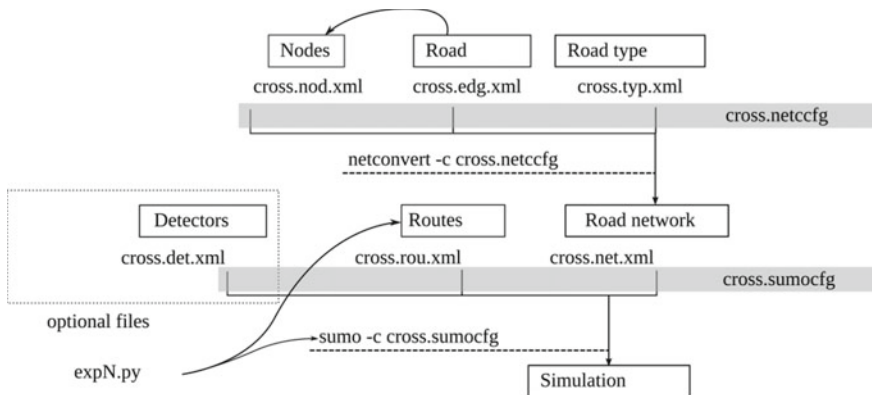


Fig. 1 Model configuration diagram with TraCI

## 2.2 The Macro-level Mathematical Model

The proposed mathematical model is based on investigations done in the field of multiphase flow phenomena [11] and the previously developed MFlow code. The system of equations of road traffic in the continual approximation is equivalent to the system of Navier–Stokes equations for a compressible medium. The model is comprised of the continuity equation and the modified equation of conservation of momentum for a continual medium [12]:

$$\frac{\partial \rho}{\partial t} + \frac{\partial(\rho V)}{\partial r} = 0,$$

$$\frac{\partial(\rho V)}{\partial t} + V \frac{\partial(\rho V)}{\partial r} + \frac{\partial P}{\partial r} = \frac{\rho}{\tau}(V_e - V).$$

The first term on the right-hand side of the momentum conservation equation reflecting the influence of the pressure gradient in hydrodynamics comes from an analogy with the leader–follower micro-model. This term reflects the fact that each driver takes into account the surrounding traffic situation, i.e., accelerates and decelerates in accordance with the behavior of neighboring vehicles. The local value of pressure increases with a decrease in the speed of automobile flow, which leads to an increase of pressure gradient and the tendency of vehicles to leave the region with a high density of vehicles.

The second term on the right-hand side expresses the tendency of the flow to achieve a certain velocity,  $V_e$ , which is optimal under certain conditions.  $V_e$  is the equilibrium flow velocity which can be reached under a uniform and homogeneous regimen of vehicle movement along the road. Parameter  $\tau$  is a characteristic relaxation time of the traffic flow on each section of the road, and in general depends on the flow parameters and the type of a vehicle. The value of  $V_e$  is equal to the maximum velocity allowed in the selected region in a simple case.

## 2.3 Numerical Method for the Macro-model Solution

The numerical method is based on the finite volume method and unstructured grids [13]. The equations of the model are discretized with a second-order accuracy over spatial coordinates. Second-order upwind schemes that satisfy the TVD criterion are used to discretize the convective term. The discretization of time derivatives is implemented using an implicit first-order Euler scheme. The SIMPLE algorithm is used to calculate the pressure field and velocity [14].

The software implementation includes an autonomous builder of the road network and the corresponding computational mesh, as well as a preprocessor and solver. All components were developed using the C++ programming language with the possibility of parallelization using the OpenMP library [15].

### 3 Prerequisites for the Transition Between Micro- and Macro-levels

In a significant part of the tasks associated with the development and research of ITS, there is no need to build a microscopic model [16] provided that each fragment of the road network is operating in a normal mode. Researchers are interested in abnormal situations on road networks such as, for example, the failure of the elements of the ITS communication component in predetermined sections of the road network. To conduct an investigation of emergency situations on a fragment of a road network, it is sufficient to describe the changes in the properties of traffic flows arriving and leaving the area in question. To develop the model that meets these requirements, we proposed to use the continual model described above. The road network of a city or region contains a significant amount of similar elements: crossroads; junction of roads/ramps [17]; straight/curved road sections, etc. In the present work, it is proposed to describe individual elements of the same type of road network using the continual model and, if necessary, proceed to the microscopic description implemented in SUMO, ensuring the transfer of parameters characterizing the average properties of traffic flows calculated by the continual model. The proposed approach will significantly reduce the calculation time of the transport network model [18].

### 4 Interaction Between the micro- and Macro-model Levels

The integration of the continuum and discrete models is realized by transmitting data describing the properties of road traffic on typical fragments [19] of road network (straight sections, crossroads, ramps). Data is transmitted using sockets. This approach allows one to simulate selected sections in parallel. The format of SUMO model output is *time;start/end;vID;speed*, where *time* is model time; *start/end* is the sign of completion or beginning of vehicle movement along the road; *vID* is the vehicle ID; and *speed* is vehicle speed. The format of MFlow model output is an array of structures containing *time2;vID2;speed2*. In a discrete model SUMO, machines are started based on input from MFlow.

The initial speed of the vehicle when it appears in the model is the maximum allowed, that is, 16.7 m/s. The movement of the vehicle as a whole reflects the actual behavior of the drivers. General significant parameters of the models are given in Table 1.

In experiments with straight road sections with a length of 500 and 5000 m, vehicles are launched alternately on each of the two lanes in the only accessible direction. At the same time, as the observation of this experiment showed, vehicles moving in the left lane tend to pick up a slightly higher speed than moving in the right lane, which is consistent with the behavior of drivers in real traffic conditions. In experiments with a T-crossroad, vehicles start in a given direction in a single accessible lane. If the speed of the vehicles and traffic congestion on a busy section

**Table 1** SUMO and MFlow model parameters

Vehicle parameter	Value
Type	Car
Length	4.5 m
Width	1.8 m
Maximum acceleration and deceleration	3 m/s <sup>2</sup>
Distance	2.5 m
Maximum allowed speed on road	16.7 m/s (60.12 km/h)

of the road prohibits one from placing a new vehicle, its launch is canceled until the next attempt. Attempts to start vehicles are carried out every 2, 4, 10, 20, 40, 100 simulation steps. Launching vehicles is connected to simulation steps for experiment repeatability.

To verify the results of modeling with discrete and continual models as well as to evaluate the gain in performance, a vehicle motion scenario was prepared and implemented in SUMO and MFlow solver. The common closed fragment of the road consists of the following blocks:

- T-crossroads, where vehicles are divided in a 1:1 ratio and turn left or right and continue driving;

- Straight section of 5000 m where cars can change lanes;

- T-crossroads, where vehicles are divided in the 1:3 ratio and turn right or continue straight ahead;

- Straight section of 500 m where cars can change lanes.

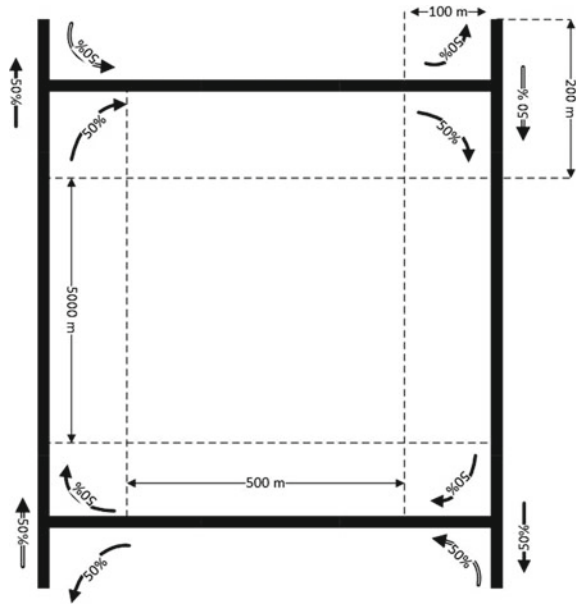
On a scheme shown in Fig. 2, vehicles move clockwise from the lower left corner. After each fragment in the discrete model, detectors are installed that take readings at the entrance and exit of the road section. The geometric dimensions of the intersections are also shown in the figure.

## 5 Experiment

The road network scheme shown in Fig. 2 was used in the numerical experiment. In our experiment, cars moved clockwise, simulating the movement on one-way streets. The arrows indicate the direction of movement of cars as well as the fraction of the car flow (of the entire flow on a straight segment or entering at an intersection) moving in this direction. The number of cars leaving the road network and entering it through all the intersections is assumed the same. This ensures consistent traffic flow properties throughout the entire simulated section of the road network.

This road network scheme was implemented in the SUMO simulator and with the help of a continual model. The simulation of the following components of the scheme was implemented in the continual model: two straight sections of 500 m length and two straight sections of 5000 m length, forming a rectangle; four T-shaped

**Fig. 2** Road network scheme



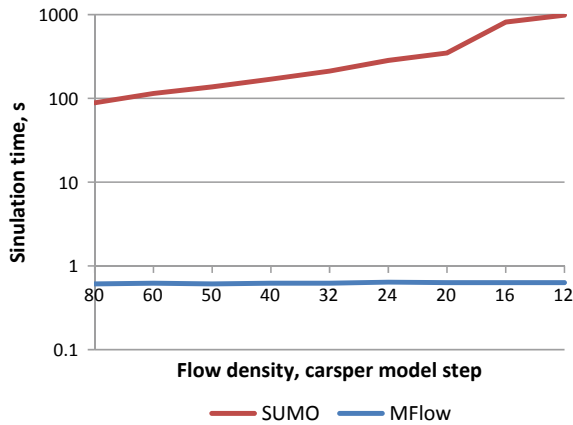
intersections (the length of each exit/entry—100 m), completing the rectangle. It should be noted that the calculation of the individual components of the road network within the framework of the continual model enables parallel calculations, which leads to a further significant reduction in the simulation time [20].

During the experiment the time required to perform the continual and discrete models was estimated. The simulation was carried out on the same computer: Intel (R) Core (TM) i5-4210U CPU @ 1.70 GHz (4 threads, 2 cores) with 4 GB, Debian GNU/Linux Buster 10 memory. Each model was run sequentially on the same single core. Other processes on the core were not being run. To isolate the processor core from the operating system scheduler, the Linux core parameter *isolcpus* = 2, 3 was added, which allowed to disable threads 3 and 4 related to the 2nd processor core. When running the simulation, a free second processor core was used by the command `taskset -c 2 python expN.py`.

During each experiment, the density of the flow of cars was changed and the (astronomical) time required was estimated. At the beginning of the experiments, the road scheme was filled with cars according to the flow properties for the particular experiment (for the continual and discrete models). This guarantees that the mutual influence of vehicles is taken into account from the very beginning of the simulation.

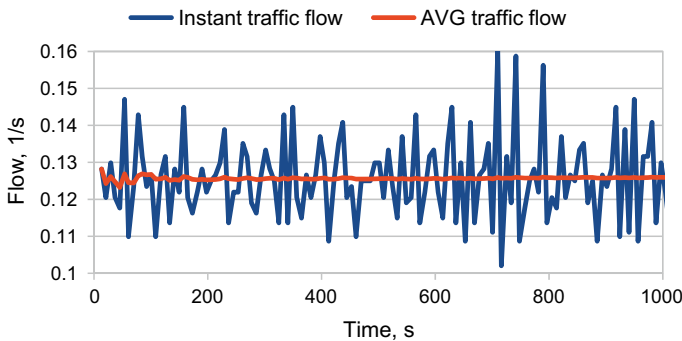
Figure 3 shows how the simulation execution time depends on the flow density of cars. The horizontal axis indicates the interval in model steps (the interval between steps being 0.1 s of model time) of a new car on the road (1 car per 100 or 40, etc. model steps).

It should be noted that for the transition between the micro- and macro-levels of the hybrid model of the transport network, it is necessary to average the discrete



**Fig. 3** Dependence of the time of the simulation on flow density

SUMO data and feed the averages into the continual model. On Fig. 4, the graph of the instantaneous values of the car flow on the scheme section is presented in dark color. A T-shaped intersection was used in this particular example. Despite the fact that the input flow of cars was set uniform and single-speed, the variation of the resulting values in time is significant. This is due to the peculiarity of the model implementation where the behavior of each road user is adjusted to the characteristics of the flow by accelerating and slowing down, as well as changing lane or overtaking other participants. This leads to deviations in speed bringing about a nonstationary behavior of the car flows at the scheme exit. It is obvious that with a given flow of cars at the entrance and the absence of exits (uncontrolled sources and drains for a car flow) the output flow must be in a stationary mode and equal to the flow at the entrance. Time averages were used to determine whether the algorithm has reached a stationary regimen. The difference between the values of total flow at the exit and



**Fig. 4** SUMO simulation results and averaged values for T-shaped intersections with 40-step vehicle start interval

entrance normalized to the input flow was used as a criterion of achieving stationarity. The value of the criterion was set to not exceed some predetermined value (0.001 in our calculations), sufficient to minimize the impact of inaccuracy in the calculation of the flow.

The time-averaged values on Fig. 4 are represented by light curve lines. For each moment of time, the average value was calculated. With the increase in the total simulation time, the average values converged to the value at the input. The fixing of the average value began from the moment when the first running car left the simulated fragment of the road. The dependence presented on Fig. 4 shows that the SUMO model quickly enters the stationary mode, which allows for a transition between the continual and the discrete models.

Based on the results obtained, it can be concluded that the hybrid model of the road traffic is applicable and effective. Within the framework of the hybrid model, a transition criterion and a procedure for averaging data during the transition from the SUMO model to the continual model are proposed, and a complex section of the road network is modeled. It is shown that, within the framework of the proposed approach, numerical modeling can be accelerated not only due to the transition between the discrete and continual models, but also due to the parallel calculation of individual road segments in both the continual and discrete approaches.

## 6 Conclusion

The paper presents a comparison of the execution time of the SUMO simulator, which implements a microscopic model of road traffic, and the MFlow code, which operates within a continual model. The continual model leads to a more than tenfold reduction in the time required to model individual fragments of the road network, which is confirmed by the presented experimental results. The key feature of the original continuum model is almost constant execution time. This is explained by the facts that (1) the proposed solution uses a fixed set of mesh cells to calculate the averaged flow characteristics, and (2) description of each individual traffic participant is eliminated altogether.

Based on the data obtained, it can be concluded that the development of a hybrid model of road traffic is feasible and promising. The reduction in simulation time due to the use of the hybrid model is ensured by transferring from the micro-model the averaged values of the parameters of traffic flow at the input of typical fragments of road networks (straight highways, crossroads) to the macro-model, calculating the averaged values of traffic flow at the output of typical fragments using the macro-model, and transferring the obtained values to the micro-model or for further calculation of typical fragments in the continuum model.

**Acknowledgements** The reported study was supported by RFBR, research project No. 18-07-00430.

## References

1. Chuvatov, M., Glazunov, V., Kurochkin, L., Popov, S.: The technology of management of data about wireless networks for vehicle's telematics map. In: Proceedings of the International Conference on Vehicle Technology and Intelligent Transport Systems, 2016, vol. 1, pp. 138–143
2. Treiber, M., Kesting, A., Helbing, D.: Delays, inaccuracies and anticipation in microscopic traffic models. *Physica A* **360**, 71–88 (2006)
3. Toledo, T.: Integrated driving behavior modeling. Ph.D. thesis, Department of Civil and Environmental Engineering, MIT, Cambridge, MA, 2003.
4. Lim, K.G., Lee, C.H., Chin, R.K.Y., Beng Yeo, K., Teo, K.T.K.: SUMO enhancement for vehicular ad hoc network (VANET) simulation. In: 2017 IEEE 2nd International Conference on Automatic Control and Intelligent Systems (I2CACIS), Kota Kinabalu, 2017, pp. 86–91.
5. Treiber, M., Kesting, A.: Modeling Lane-Changing Decisions with MOBIL. In: Appert-Rolland, C., Chevoir, F., Gondret, P., Lassarre, S., Lebacque, J.P., Schreckenberg, M. (eds.) *Traffic and Granular Flow '07*, pp. 211–220. Springer, Berlin, Heidelberg (2007)
6. Treiber, M., Hennecke, A., Helbing, D.: Congested traffic states in empirical observations and microscopic simulations. *Phys. Rev. E* **62**, 1805–1824 (2000)
7. Helbing, D.: Gas-kinetic derivation of Navier-Stokes-like traffic equations. *Phys. Rev. E* **53**, 2366–2381 (1996)
8. Lighthill, M.J., Whitham, G.B.: A Theory of traffic flow on long crowded roads. *Proc. R. Soc. Lond. A*, **229**, 317–345 (1955).
9. Helbing, D., Hennecke, A., Shvetsov, V., Treiber, M.: Micro- and macro-simulation of freeway traffic. *Math. Comput. Model.* **35**, 517–547 (2002)
10. Haddouch, S., Hachimi H., Hmina, N.: Modeling the flow of road traffic with the SUMO simulator. In: 2018 4th International Conference on Optimization and Applications (ICOA), Mohammedia, 2018, pp. 1–5.
11. Chernyshev, A.S., Schmidt, A.A., Ispolzovaniye eylerovo-eylerovskogo podkhoda dlya modelirovaniya turbulentykh techeniy puzyr'kovykh sred. *Pisma v ZHTF* 39(12), 17–24 (2013) (in Russian)
12. Helbing, D.: Gas-kinetic derivation of Navier-Stokes-like traffic equations. *Phys. Rev. E* **53**, 2366–2381 (1996).
13. Versteeg, H.K., Malalasekera, W.: An Introduction to Computational Fluid Dynamics, 2nd edn. Pearson Education Limited, UK (2007)
14. Patankar, S.: Numerical Heat Transfer and Fluid Flow. Hemisphere Publishing Corporation, NY, USA (1980).
15. Chernyshev, A., Schmidt, A., Kurochkin, L.: Numerical modeling of polydisperse bubbly flows by the OpenMP parallel algorithm. *Proc. Comp. Sci.* **108C**, 1990–1997. ICCS-2017, 12–14 June 2017, Zurich Switzerland, 2017
16. Krauß, S., Wagner, P., Gawron, C.: Metastable states in a microscopic model of traffic flow. *Phys. Rev. E* **55**(304), 55–97 (1997)
17. Hidas, P.: Modeling lane changing and merging in microscopic traffic simulation. *Transp. Res. Part C* **10**(5), 351–371 (2002).
18. Kurochkin, L.M., Chuvatov, M.V., Glazunov, V.V., Chernyshev, A.S.: Comparison of traffic flow simulation results by continuous and discrete-event methods. *Soft Comput. Meas. (SCM'2018)* **1**, 500–503 (2018)
19. Newell, G.F.: A simplified car-following theory: a lower order model. *Transp. Res. B* **36**, 195–205 (2002)
20. Kurochkin, L.M., Chernyshev, A.S., Kurochkin, M.A., Domrachev, D.A., Prokhorov, M.V.: Application of supercomputer technologies for studying the dynamics of polydisperse mediums. *St. Petersburg State Polytech. Univ. J. Comput. Sci. Telecommun. Control Syst.* **10**(1), 63–76 (2017).



# Synchronization Scheme for UWB Wireless Sensor Network System



Iuliia Tropkina , Sergey Zavjalov , and Dong Ge 

**Abstract** Ultra-wide band (UWB) is a modern radio technology that is popularly used in wireless sensor networks (WSNs) due to high resistance to multipath fading in conditions of numerous reflective surfaces. However, a cost of an ultimate device can be very high because of problems of ultra-short impulse processing. Hence, we are focused on the synchronization task to employ the UWB technology for a WSN, which has a large number of nodes. We utilize the non-coherent novel approach to detect a signal at the receiving side and for this scheme, it is necessary to choose preamble that has the best properties. In this paper, we compare the auto-correlation and cross-correlation properties of M-sequences and optical orthogonal codes (OOCs) to reveal the best one for transmission by one user and several users asynchronously. Also, we depict the best preamble threshold value using the Bayesian approach.

**Keywords** UWB · Wireless sensor network · Synchronization · OOC · Orthogonal codes · M-sequences

## 1 Introduction

Over the last decade, WSNs have been successfully applied in many engineering fields such as aircraft, shipbuilding area, and many over fields [1–4]. Nodes of these systems provide the following functions: data collection from different sensors, preliminary data processing, as well as energy harvesting by thermal gradient. The last one is implemented using special thermoelectric generators that ensure decline of power consumption of modules [5–7].

---

I. Tropkina (✉) · S. Zavjalov  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [tropkina2.yua@edu.spbstu.ru](mailto:tropkina2.yua@edu.spbstu.ru)

D. Ge  
Tsinghua University, Haidian District, Beijing, People's Republic of China

It is possible to provide high multipath fading resistance for WSNs. Many studies are focused on reduction of out-of-band emission to avoid effect of intersymbol interference on characteristics of a device [8, 9]. These studies consider intersymbol interference in term of signal specter, but UWB technology allows to solve this problem in time domain. A very short signal duration prevents two signals from overlapping; therefore, the performance of the system becomes higher.

For these UWB systems, it is possible to choose correct multiple access scheme at a receiving side to provide required performance. In addition, pulse duration in UWB systems is of the order of ns; thus, a high-speed ADC is demanded for accurate signal recovery. Nevertheless, such ADCs require a lot of energy and cost. Therefore, the second main problem we are focused on is the problem of synchronization. Some works consider different types of receivers. Basically, there are two main groups of receivers: coherent and non-coherent. A coherent RAKE-type receiver requires large complexity of a scheme and, thus, more cost. However that leads to improvement of final performance of a device [10]. Also, in these receivers, accurate timing acquisition and channel estimation are required. The second broad class of receivers is divided on energy detection receivers and auto-correlation receivers (AcRs). An energy detector (ED) is most simple receiver structure suggested for UWB. It squares the received signal and accumulates its energy over a defined interval using analog hardware [11–13]. The AcR is more robust against inter-pulse and intersymbol interference than an ED [14, 15].

In our paper, we use a novel approach to detect a signal. Instead of a set of elements for integration, we use only one comparator and a D-trigger that simplifies the final circuit. For this type of detection, it is reasonable to use preamble signal search. This approach solves the problem of multiple access, as well. Because of it, the goal of this work is an investigation of an effect of preamble choice on performance characteristics of the scheme. We compare two cases: the simultaneous transmission by several users and the transmission utilizing Aloha access control protocol. In addition, we reveal the best threshold value and explore which a sequence is more stable to length changing.

The rest of the paper is organized as follows. Section 2 introduces the receiver and the transmitter modeling for the UWB WSN, as well as the modeling of multisignal scenario with adding additive white Gaussian noise (AWGN) to the channel. The definition and properties of different sequences, which can be employed as preamble are given in Sect. 3, while the results are discussed in Sect. 4. Conclusions are drawn in the last section.

## 2 System Modeling

In this section, the UWB WSN modeling is represented. The implementation of the real device and performance characteristics of the real scheme are described in [16, 17].

## 2.1 Transmitter

The modeling of the UWB WSN system is shown in Fig. 1. Primarily, the massive with sequences which have the best correlation and auto-correlation properties is created for each user. Data bits are formed and fed into the spreading block, where one of extending sequences is used according to preamble number. In our investigation, we implement OOC as spreading sequences, because it has stability to different cases of signal overlapping. Then, the extended data and corresponding preamble are combined together and modulated using the on-off keying (OOK) modulation scheme. According to this scheme, each chip is replaced by absence or presence of a pulse in each pulse repetition interval. An UWB transmitter sends a stream of ultra-short pulses, occupying several GHz bandwidth. The monocycle shape is the first derivative of a Gaussian impulse that can be described as

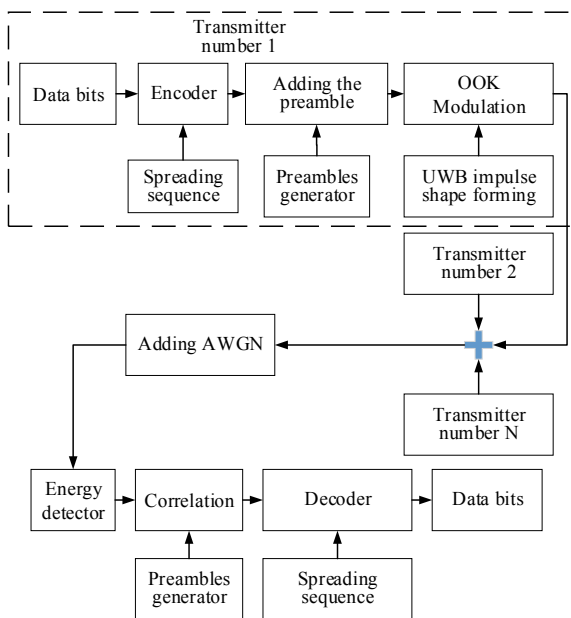
$$w(t) = A_0 \left( \frac{t}{T_w} \right) \exp \left( - \left( \frac{t}{T_w} \right)^2 \right), \quad (1)$$

where  $T_w$  is the monocycle duration;  $A_0$  is the monocycle amplitude.

The final transmitted signal of the  $i$ th bit  $\beta_i$  is described by the following equation:

$$r(t) = \sum_{j=0}^{F-1} \beta_i w(t - iT_b - jT_p), \quad (2)$$

**Fig. 1** Modeling of the UWB receiver



where  $F$  is the number of pulses per data bit;  $T_p$  is the pulse repetition period; and  $T_b = FT_p$  is the bit duration.

At the end, the same packets are added several times to estimate BER performance more accurately.

## 2.2 Signal-to-Noise Ratio Control

The signals from  $N$  receivers are summed over an AWGN channel. To control signal-to-noise ratio (SNR), we use the following approach. Each signal from one of UWB transmitters is multiplied to coefficient  $W$  to reach necessary SNR value. To be precise, in this work, the energy per pulse to noise power spectral density ratio is used to avoid influence whole signal energy due to different number of ones in the extending sequences.

$$W = \frac{1}{\sqrt{\frac{E_p}{N_0(10^{\text{SNR}_{\text{control}}/10)}}}}, \quad (3)$$

where  $E_p$  is the energy of one pulse;  $\text{SNR}_{\text{control}}$  is the controlled value of energy per pulse to noise power spectral density ratio; and  $N_0$ —is the spectral density of randomly formed noise.

## 2.3 Receiver

The data processing at the receiving side begins with the energy detection block. If the received signal enhances the threshold value, a logical “1” will be recorded. For another case, a logical “0” will be recorded. The comparator threshold value can be determined according to signal-to-noise ratio [18]. In our paper, the comparator threshold value is equal half of impulse amplitude. Then, preamble search in an input signal follows. The block of correlation processing performs this function. After signal detection, a stream of logical ones and zeros is fed into the decoder block to get useful data bits.

# 3 Orthogonal Sequences

## 3.1 Definition and Properties of OOC

The optical orthogonal code is a family of  $\{0, 1\}$  sequences. OOC sequences have to satisfy the following two properties to provide orthogonality among different users:

- Each sequence should be distinguished from a time-shifted version of itself (auto-correlation constraint).
- Each sequence should be distinguished from a possibly time-shifted version of another sequence (cross-correlation constraint).

These sequences are described by following parameters: number of elements in sequence, weight of the sequence, auto-correlation constraint, and cross-correlation constraint. When last two parameters are equal 1, OOC sequence will have the best correlation characteristics. Thus, each sequence can be expressed as

$$S_u = s_{u,0}, s_{u,1}, \dots, s_{u,F-1}, \quad (4)$$

where  $u = 1, 2, \dots, N$ ;  $N$  is the number of users;  $F$  is the number of elements in a sequence; and  $s_{u,j} \in \{0, 1\}$ .

The relation between the auto-correlation and the auto-correlation constraint is determined as

$$\theta_{uu}(\tau) = \sum_{j=0}^{F-1} s_{u,j} \cdot s_{u,((j+\tau)\text{mod}F)} \leq \lambda_a (\tau \neq 0), \quad (5)$$

where  $\lambda_a$  is the auto-correlation constant;  $\tau$  is the shift in term of elements of sequence; and  $\theta_{uu}(0)$  is equal to the weight of the sequence  $K$ .

The relation between the cross-correlation and the cross-correlation constraint is given as

$$\theta_{uw}(\tau) = \sum_{j=0}^{F-1} s_{u,j} \cdot s_{w,((j+\tau)\text{mod}F)} \leq \lambda_c, \quad (6)$$

where  $\lambda_c$  is the cross-correlation constant.

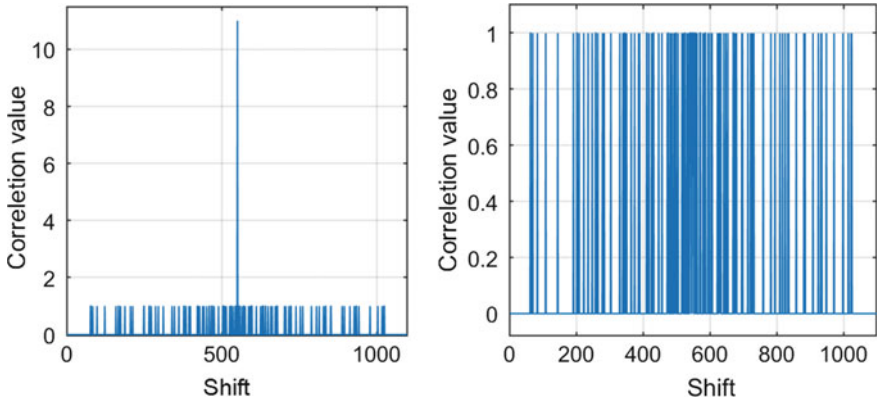
The optimal length of a sequence can be described by the following equation:

$$F_{\text{opt}} = N \cdot K \cdot (K - 1) + 1. \quad (7)$$

The example of auto-correlation and cross-correlation function of two different OOC sequences with length 551 elements is represented in Fig. 2.

### 3.2 Definition and Properties of M-Sequences

M-sequences are pseudo-random binary sequences, which are completely controlled by the tap sequence. It has approximately the same number of ones and zeros. A tap sequence defines which bits in the current state will be combined to determine the

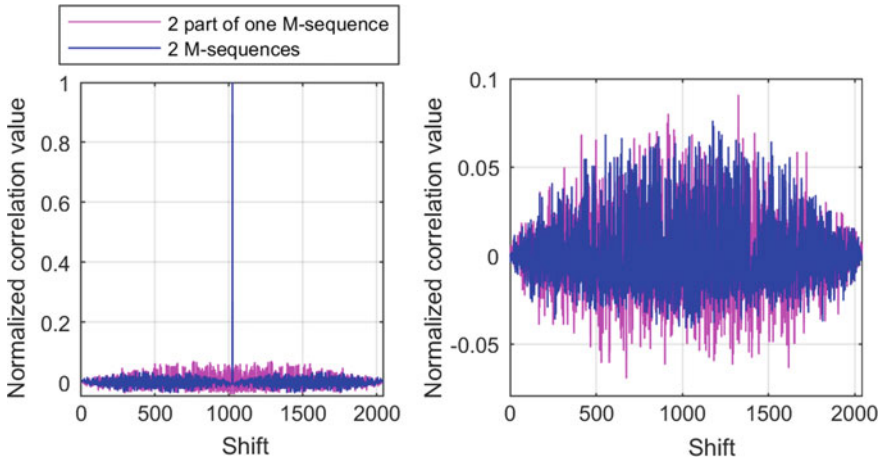


**Fig. 2** Plot for the auto-correlation function of two OOC sequences at the left side and the cross-correlation function at the right side

input bit for the next state. The combination is generally performed using module-2 addition.

We assess the auto-correlation and cross-correlation functions of M-sequences forming different ways to choose more effective one. M-sequence can be formed utilizing two polynomials or dividing one sequence to several parts.

It can be seen from Fig. 3 that lobes of the auto-correlation and cross-correlation function are lower for case of using two different polynomials for M-sequences forming. We use this type of creating sequences in the future investigation to keep



**Fig. 3** Plot for the auto-correlation function normalized on the length of the preamble of two M-sequences on the left side and the cross-correlation function normalized on the length of the preamble at the right side

high performance characteristics. Other sequences, for example, Walsh codes, do not consider, due to insufficient correlation properties.

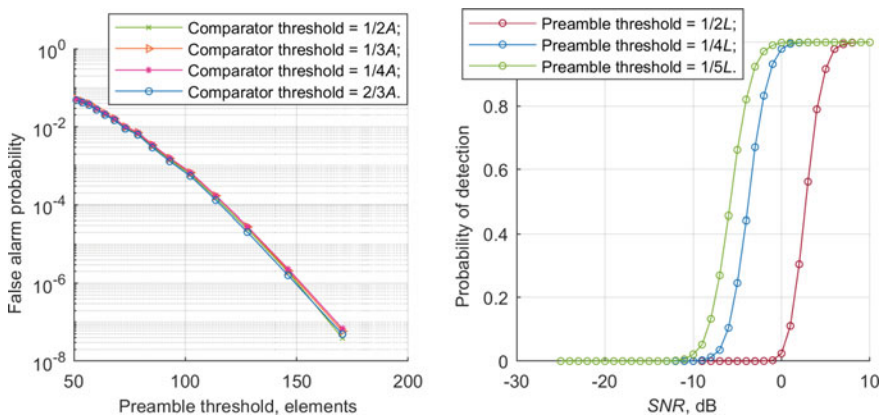
### 4 Results and Discussion

In this section, we summarize the results on the performance of the UWB WSN system using M-sequences and OOC as preambles. We use the transmission of the packet that contains 24 information bits. The number of experiments was chosen 100,000 times to take into account statistics (Fig. 4).

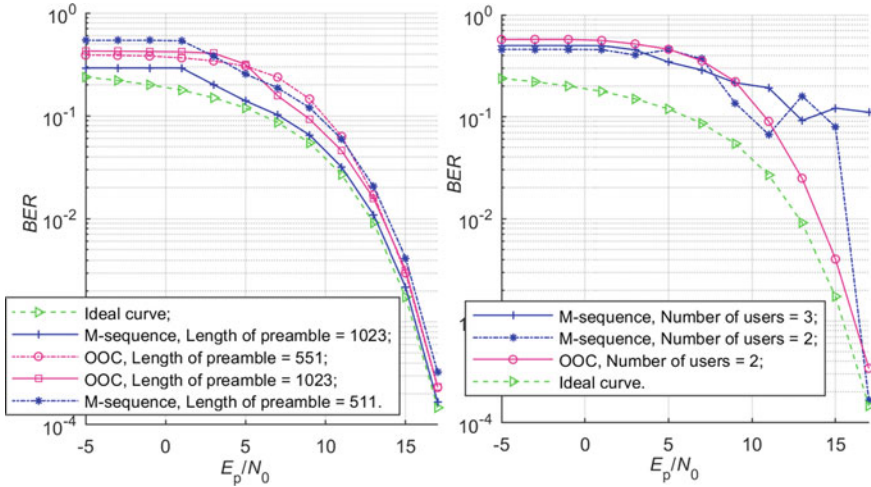
Firstly, we determined the best decision of preamble threshold value utilizing the Bayesian approach. According to the dependence of false alarm probability on signal-to-noise ratio, the preamble threshold value enhanced one fifth of preamble length provides the false alarm probability equaled zero. The choice of a comparator threshold does not affect false alarm probability. However, preamble threshold value has to be as small as possible corresponding to the probability of detection curve.

Secondly, we compared M-sequences and OOC using the BER estimation using different length of sequences. The ideal curve of transmission without a preamble is shown. As can be seen from Fig. 5, M-sequences ensure the best performance characteristics for case of transmission by only one user. However, OOC sequences are more stable to length changing.

Figure 5 depicts the case of asynchronous transmission by different number of users. For this purpose, random shift intervals are added between signals from different users in time domain. According to the figure, if we increase number of users, the performance of scheme will become better for case of using OOC sequences compared with M-sequences. OOC sequences contain small number of pulses that



**Fig. 4** Plot for value of false alarm probability versus preamble threshold value at the left side and plot for value of probability of detection versus SNR value at the right side. *A*—the impulse amplitude; *L*—the preamble length



**Fig. 5** Plot for value of *BER* versus energy per pulse to noise power spectral density ratio for only one transmitting user using M-sequences and OOC sequences at the left side and plot for value of *BER* versus energy per pulse to noise power spectral density ratio for the case of asynchronous transmission by different number of users at the right side

lets to increase number of users that makes this scheme more effective in UWB WSN systems, which use the OOK modulation and an energy detector at the receiving side. OOC sequences can be applied in systems utilizing Aloha access control protocol.

## 5 Conclusion

In this paper, we presented the modeling of an UWB WSN system considering problem of multiple access and synchronization scheme. We determined that the best decision of preamble threshold value is one fifth of preamble length. The results show that M-sequences ensure the best performance characteristics for case of transmission by only one user, but OOC sequences are more stable to preamble length changing. In addition, if number of users is high or Aloha access control protocol is used, OOC sequences will demonstrate the best characteristics. Thus, we determined an effective choice of sequences used for preambles in UWB WSN systems, which use OOK modulation and energy detector at the receiving side.



## References

1. Kvashenkina, O.E., Gabdullin P.G., Arkhipov, A.V.: SmartFoil: a novel assembly technology for electronic circuit boards in multifunctional units. In: IEEE International Conference on Electrical Engineering and Photonics (EExPolytech), pp. 202–206 (2018).
2. Pergushev A., Sorotsky, V.: Signal distortion decreasing in envelope tracking power amplifiers. In: IEEE International Conference on Electrical Engineering and Photonics (EExPolytech), pp. 44–47 (2018).
3. Trubin, P., Savchenko E., Velichko, E.: Development of polarimetric sensor for identification system. In: IEEE International Conference on Electrical Engineering and Photonics (EExPolytech), pp. 279–282 (2018).
4. Savchenko, E.A., Velichko, E.N., Aksenov E.T., Nepomnyashchaya, E.K.: Combined method for laser selection, positioning and analysis of micron and submicron cells and particles. In: International Conference Laser Optics (ICLO), pp. 539–539 (2018).
5. Korotkov, A., Loboda, V., Dzyubanenko S., Bakulin, E.: Fabrication and testing of MEMS technology based thermoelectric generator. In: 7th Electronic System-Integration Technology Conference (ESTC), pp. 1–4 (2018).
6. Korotkov, A.S., Loboda, V.V., Makarov, S.B., Feldhoff, A.: Modeling thermoelectric generators using the ANSYS software platform: methodology practical applications and prospects. *Russ. Microelectron.* **46**(2), 131–138 (2017)
7. Korotkov, A., Loboda, V., Feldhoff, A., Groeneveld, D.: Simulation of thermoelectric generators and its results experimental verification. In: Proceedings of IEEE International Symposium on Signals Circuits and Systems, pp. 13–14 (2017).
8. Sadovaya, Y.S., Gelgor, A.I.: Synthesis of signals with a low-level of out-of-band emission and peak-to-average power ratio. In: IEEE International Conference on Electrical, pp. 103–106 (2018).
9. Zavjalov, S.V., Volvenko, S.V., Makarov, S.B.: A method for increasing the spectral and energy efficiency SEFDM signals. *IEEE Commun. Lett.* **20**(12), 2382–2385 (2016)
10. Cassioli, D., Win, M., Vatalaro, F., Molisch, A.: Low complexity rake receivers in ultra-wideband channels. *IEEE Trans. Wireless Commun.* **6**(4), 1265–1275 (2009)
11. Nagaraj, Rassam, F.: Improved non-coherent UWB receiver for implantable biomedical devices *IEEE Trans. Biomed. Eng.* **63**(10) (2016).
12. Decarli, N., Giorgetti, A., Dardari, D., Chiani, M., Win, M.Z.: Stop-and-go receivers for non-coherent impulse communications. *IEEE Trans. Wireless Commun.* **13**(9), 4821–4835 (2014)
13. Tang, Y., Ruan, Y.: Research on a novel synchronization and detection scheme used in energy detection UWB receiver. In: 11th International Symposium on Communications & Information Technologies (ISCIT), pp. 109–113 (2011).
14. Troesch, F., Wittneben, A.: An ultra wideband transmitted reference scheme gaining from intersymbol interference. In: Asilomar Conference on Signals, Systems, and Computers, pp. 1070–1074 (2007).
15. Witrals, K., Pausini, M.: Statistical analysis of transmitted-reference UWB systems on multipath channels. In: IEEE ICUWB, Waltham, MA (2006).
16. Volvenko, S.V., Zavjalov, S.V., Gruzdev, A.S., Vasiljev, D.S.: Experimental ultra wideband wireless sensor network for data collection. In: DSPA, pp. 345–351 (2017).
17. Volvenko, S.V., Ge, D., Zavjalov, S.V., Gruzdev, A.S., Rashich, A.V., Svechnikov, E.L.: Experimental wireless ultra wideband sensor network for data collection. In: Progress in Electromagnetics Research Symposium—Spring (PIERS), pp. 965–970 (2017).
18. Zavjalov, S.V., Tropkina, I.A., Mikhailov, A.S.: Effective choice of parameters of IR-UWB sensor network system. *J. Phys.: Conf. Ser.* **1236**(1) (2019).

# The Deep Survival Forest and Elastic-Net-Cox Cascade Models as Extensions of the Deep Forest



Lev Utkin , Andrei Konstantinov , Anna Meldo , Victoria Sokolova , and Frank Coolen 

**Abstract** Two new survival models, the deep survival forest and the Elastic-Net-Cox Cascade, are presented in the paper. They can be regarded as a combination of random survival forests and the Elastic-Net-Cox models with the deep forest (DF) proposed by Zhou and Feng. The main ideas to construct the models are to replace the original random forests incorporated into the DF with the corresponding survival analysis models. A stacking algorithm implemented in the deep survival forest and the Elastic-Net-Cox Cascade, which can be regarded as a link between the DF levels, uses quantiles of the random time-to-event and the mean time-to-event computed from the estimated survival functions at every level of the DF. Numerical examples with real data illustrate the proposed models.

**Keywords** Random forest · Survival tree · Deep learning · Stacking · Survival analysis · Cox model

## 1 Introduction

One of the important peculiarities of many systems dealing with diagnosis statements and making decisions about a corresponding treatment is availability of censored data, which are usually considered in the framework of survival analysis [1]. Censored

---

L. Utkin (✉) · A. Konstantinov  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [lev.utkin@gmail.com](mailto:lev.utkin@gmail.com)

A. Meldo  
Clinical Research Center of Specialized Types of Medical Care (Oncological), St. Petersburg, Russia

V. Sokolova  
St. Petersburg State Forest Technical University, St. Petersburg, Russia

F. Coolen  
Durham University, Durham, UK

data usually take place in studying time-to-event where only partial (censored) information instead of precise measurement of survival time is available for some subjects. The importance of survival analysis demonstrates significant growth nowadays due to the increasing role of machine learning models and algorithms dealing with censored data.

There are three main groups of the survival analysis models. The first group consists of parametric models, which use some well-known probability distributions in order to model real processes such that the parameters of the probability distributions become to be the main goal of survival analysis. These models are the most efficient under condition that the probability distributions are known. However, this condition is usually violated. In contrast to the parametric models, semi-parametric models, which make up the second group, make no assumption about the distribution of time-to-event, but do make assumptions about how covariates change survival experience. One of the well-known semi-parametric models is the Cox proportional hazards model [2] where the baseline hazard is allowed to vary with time. This is the most popular model in survival analysis. Being a semi-parametric model, the Cox model does not require knowledge of the probability distribution of time-to-event. The proportional hazards assumption in the Cox model means that different subjects have hazard functions that are proportional, i.e., the ratio of the hazard functions for two subjects with different covariates is a constant and does not vary with time [3]. This is an important assumption which, however, restricts the model application. As a result, the Cox model is based on using the linear proportional hazards condition.

It should be noted that the Cox model is a useful and quite interesting technique which effectively allows us to process survival data and to estimate how different features or covariates impact on the corresponding hazard function. Therefore, one can find a huge amount of work devoted to the Cox model and its modifications in the literature [4–7].

The third group of survival models contains nonparametric or distribution-free models, which are used when no theoretical distribution adequately fits the data. Moreover, they do not assume a specific relationship between covariates and the survival time like the Cox model. One of the well-known and popular nonparametric survival models is the Kaplan–Meier model, which is based on individual survival times and assumes that censoring is independent of survival time, that is, the reason an observation is censored is unrelated to the cause of failure.

In spite of many advantages of the Cox model and its popularity, it cannot be applied to analyze applications when survival data have a high dimensionality, when the relationship between covariates and the time-to-event is nonlinear, or when the amount of observations is rather small. In order to overcome these obstacles and conditions, a lot of approaches have been proposed. It is important to point out that these approaches are concerned with semi-parametric models, for example, the Cox model, as well as nonparametric models. There are various Lasso modifications (adaptive Lasso, group Lasso, etc.) of the Cox model [8–11]. In fact, these modifications introduce the regularization into the Cox model in order to restrict the set of model parameters and to adapt the Cox model to approaches used in machine learning frameworks. The aforementioned modifications opened a set of new survival models.

One of these models is the so-called Elastic-Net-Cox model which is based on the extension of the Cox model with the linear combination of the L1-norm penalty and L2-norm penalty [12]. However, all these models do not overcome the assumption of the linear relationship between covariates and the time-to-event. Therefore, a lot of new models have been developed in order to avoid the assumption of linearity. Part of the models is based on using neural networks as a function which links covariates and the survival time [13–17]. Other models use the support vector machine approach for dealing with survival data [18–21]. Strongly speaking, it should be noted that most proposed models cannot be viewed as extensions of the Cox model. Some of the models belong to the nonparametric models.

A lot of models have been proposed which are based on using the so-called survival decision trees and random survival forests (RSF). The corresponding decision trees and random forests (RFs) [22] differ from the original well-known similar machine learning models by splitting criteria applied to decision trees, which use peculiarities of the censored data analysis [23–26]. It turned out that random survival forests can be regarded as one of the best models for survival analysis, especially when the training set is small or when a lot of censored examples are available in the training set.

One of the important questions when we deal with survival models is a measure for evaluation of the quality of models and for tuning them. The problem is that traditional measures of predictive performance cannot be directly used due to censoring, and they have to be modified in order to take into account the censoring.

There are several measures for evaluating survival analysis, including the concordance index (C-index) [27], the L1-loss (the average absolute value of the difference between the true time-to-event and the predicted time), the Brier score (the mean squared error between the  $\{0, 1\}$ -event status at some predefined time and the predicted survival probability at the same time), etc. We use the C-index for evaluating survival analysis and the model quality. This measure estimates how good the model is at ranking survival times.

One can see from the above that a lot of machine learning models are available for performing survival analysis. However, a lot of new effective machine learning models have been developed in the last years. It is necessary to adapt these models to solving the survival analysis problems. One of the interesting models developed in the last years is the so-called deep forest (DF) [28]. This is an ensemble-based model which includes a set of the RFs organized in a special form of levels of a forest cascade, similarly to layers in neural networks. In contrast to neural networks, the DF contains many RFs instead of neurons. Therefore, the problem to adapt the DF to survival analysis problems arises, and it is solved below. First of all, it is proposed to replace the RFs with RSFs. This is an obvious replacement. One of the important peculiarities of the DF is a scheme of the feature vector forming for its use at the next levels of the DF cascade. When we solve a classification problem by means of the DF, an output vector at a level of the original DF cascade is formed by concatenating the original feature vector of a training or testing example with the class probability vector (augmented features) obtained by averaging the class probability vectors of all trees. In other words, results of classification at a level of the forest cascade are

used for training and building RF at the next level. This important peculiarity of the DF establishes a connection between levels of the forest cascade.

The main aim of the presented work is to develop a deep survival forest (DSF), which can be regarded as an extension of the DF in the case of censored data for survival analysis. Unfortunately, the approach for forming the connection between levels of the forest cascade by means of the concatenation of augmented features with original feature vectors cannot be used directly for the DSF because outputs of RSFs differ from the original RFs. Therefore, we propose a modification of the DF algorithm by using outputs of the RSFs. The augmented features for concatenation with the original feature vector are proposed to consist of two parts. The first part of the augmented features consists of a set of quantiles of the time-to-event. The second part consists of the mean time-to-event computed by using the obtained SF. By using the above ideas, we construct the DSF which solves the survival analysis problem. Moreover, we construct a cascade structure which is similar to the DSF, but the Elastic-Net-Cox models are used instead of the RSFs. This structure is called the Elastic-Net-Cox Cascade (ENCC).

Two main strategies of the DSF construction were studied and analyzed. The first one uses accumulation of augmented features with every level of the forest cascade such that the length of the feature vector increases with levels. The second strategy uses only accumulated features computed at the last level, and the augmented features of all previous levels are forgotten. As a result, the length of the feature vector is not changed at levels of the forest cascade. It should be noted that the second strategy has outperforming results in comparison with the first one. Therefore, we give the numerical results only for the second strategy.

This paper is organized as follows. Section 2 provides some definitions of survival analysis including the survival and cumulative hazard functions, the Cox model and the Kaplan–Meier estimator. An extension of the Cox model using some types of regularization is given in Sect. 3. RSFs are introduced in Sect. 4. The DF proposed by Zhou and Feng [28] is described in Sect. 5. Architectures of the DSF and the ENCC are considered in Sect. 5. Section 6 provides results of numerical experiments. Concluding remarks are made in Sect. 7.

## 2 Some Basic Definitions of Survival Analysis

For simplicity, we will use the term patient to indicate a subject of interest. A training set  $D$  usually consists of  $n$  triplets  $(\mathbf{x}_i, \delta_i, T_i)$ ,  $i = 1, \dots, n$ , where every triplet characterizes a patient such that  $\mathbf{x}_i = (x_{i1}, \dots, x_{im}) \in \mathbf{X} \subseteq \mathbf{R}^m$  is the vector of the patient parameters or features;  $T_i \in \mathbf{R}_+$  is time-to-event of the patient;  $\delta_i = 1$  corresponds to an uncensored observation; and  $\delta_i = 0$  indicates a censored observation. We aim to estimate the time to the event  $T$  for a new patient having a feature vector  $\mathbf{x}$ .

Key concepts of survival analysis are the survival function (SF) and the cumulative hazard function (CHF). The SF  $S(t)$  is the probability of surviving up to time  $t$ . Suppose we have the ordered times-to-event for patients:  $T_1 \leq T_2 \leq \dots \leq T_n$ . The

estimated SF can be expressed as follows (the Kaplan–Meier estimator):

$$\hat{S}(t) = \prod_{T_i < t} \frac{n_i - d_i}{n_i}$$

where  $n_i$  is the total number of individuals at risk (alive and not censored) just prior to  $T_i$ ;  $d_i$  is the total number of events happening until time  $T_i$ .

The CHF  $H(t)$  is defined as the integral of the hazard function  $h(t)$  which is the rate of event at time  $t$  given that no event occurred before time  $t$  [1].

An important model in survival analysis is the Cox proportional hazards model [2]. In accordance with the model, the hazard function  $h(t|\mathbf{x})$  at time  $t$  given the feature vector  $\mathbf{x}$  is defined as follows:

$$h(t|\mathbf{x}) = h_0(t) \exp(\mathbf{x}\mathbf{b}^T),$$

where  $h_0(t)$  is a baseline hazard function;  $\mathbf{b} = (b_1, \dots, b_m)$  is an unknown vector of regression coefficients.

To find parameters of the Cox model, the partial likelihood is used of the form:

$$L(\mathbf{b}) = \prod_{j=1}^n \left[ \frac{\exp(\mathbf{x}_j \mathbf{b}^T)}{\sum_{i \in R_j} \exp(\mathbf{x}_i \mathbf{b}^T)} \right]^{\delta_j}.$$

Here  $R_j$  is the set of patients who are at risk at time  $t_j$ .

Another important concept in survival analysis is a measure for comparison of different survival models, called the C-index [27]. This is a probability that the event times of a pair of patients are correctly ranked. Let  $t_1, \dots, t_n$  be predefined time points. Then there holds

$$C = \frac{1}{M} \sum_{i:\delta_i=1} \sum_{j:t_i < t_j} \mathbf{1}[\hat{S}(t_i|\mathbf{x}_i) > \hat{S}(t_j|\mathbf{x}_j)].$$

Here  $M$  is the number of all comparable or admissible pairs (a pair is not admissible if the events are both right-censored or if the earliest time in the pair is censored);  $\mathbf{1}[a]$  is the indicator function taking the value 1 if  $a$  is true, and 0 otherwise.

### 3 The Elastic Net Method in Survival Analysis

In order to restrict the set of model parameters and to deal with high-dimensional data when the number of features in the given data is almost equal to or even exceeds the number of training examples, regularized Cox models have been introduced [10, 11, 29]. These models can be regarded as new survival models, but they preserve

the linear relationship between features and the time-to-event. One of the models, called the Lasso-Cox, is an extension of the well-known Lasso model [9] on survival analysis. The Lasso-Cox is based on incorporating the L1-norm penalty into the partial likelihood  $L(\mathbf{b})$ . This penalty is of the form  $\lambda \sum_{k=1}^m |b_k|$ , where  $\lambda$  is a hyperparameter which controls the strength of the regularization. In the same way, the Cox model can be extended by using the L2-norm penalty  $\lambda \sum_{k=1}^m b_k^2$  (Ridge-Cox) and by using the linear combination of the L1-norm penalty and L2-norm penalty with a coefficient  $\alpha \in [0, 1]$  (Elastic-Net-Cox) [12]. The Elastic-Net-Cox model uses the penalty.

$$\lambda \left( \alpha \sum_{k=1}^m |b_k| + (1 - \alpha) \sum_{k=1}^m b_k^2 \right).$$

Our aim is to incorporate the Elastic-Net-Cox model into the cascade structure like the DSF and to compare it with the DSF based on applying RSFs. The corresponding cascade structure will be called the Elastic-Net-Cox Cascade (ENCC).

## 4 Random Survival Forests and Deep Forests

Let us give some definitions of the RSF. Every decision tree in the RSF differs from the original decision trees by splitting rules. There are several splitting rules, for example, the log-rank rule, the approximate log-rank splitting rule, and the conservation of events splitting rule. Their detailed descriptions can be found in [6, 30]. An algorithm of constructing the RSF is described by Ishwaran et al. [30] in detail. The output of every tree is the CHF estimate defined by means of the Nelson–Aalen estimator. The ensemble CHF estimate for a patient is obtained by averaging CHFs of all trees.

Before considering the DSF and the ENCC, we briefly introduce the DF [28]. The DF architecture is a cascade which consists of a set of levels such that each level receives feature information processed by its preceding level, and outputs its result to the next level [28]. The important idea underlying the DF is a class probability distribution computed for every decision tree and each feature vector, and producing a RF class vector. In order to use the results of classification at some level of the forest cascade, the RF class vectors are concatenated with the original vector to be the input to the next level of the cascade. An architecture of the DF proposed by Zhou and Feng [28] is shown in Fig. 1.

Our aim is to modify the DF structure in order to adapt it to survival analysis.

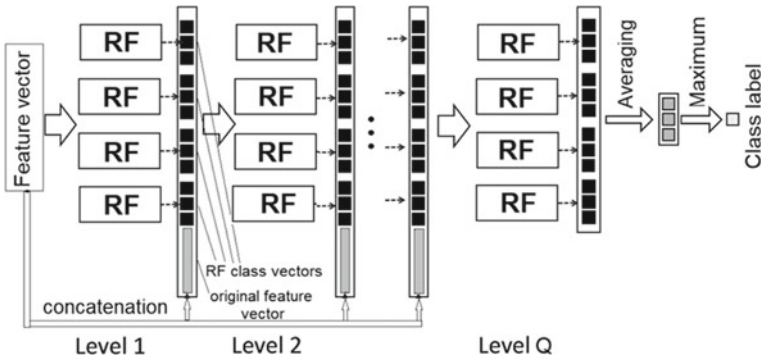


Fig. 1 DF architecture

### 5 The DSF and the ENCC

In order to adapt the DF to survival analysis, we propose to implement the following ideas. First of all, RFs should be replaced with RSFs. However, this replacement raises a question which is very important for implementing the DSF: How to construct the augmented feature instead of the class probability vectors used in the DF? We have to choose some representation of every RSF output for an efficient implementation of the stacking algorithm. This representation should be useful for the next levels and should show how exactly an example has been analyzed at the previous level.

Note that the SF  $S(t)$  is the most informative representation of the results. Therefore, we use the SF to represent it as the augmented features. Unfortunately, we cannot apply the whole SF because it may have too many jumps. Moreover, it may mask the original feature vector if this vector has a small number of elements. A reasonable way to overcome this difficulty is to consider quantiles of the random time to event, which can be obtained from the SF  $\hat{S}_f(t|\mathbf{x}_i)$  corresponding to the  $i$ -th patient. Let us take  $v-1$  quantiles which have the values  $t_i(p_1), \dots, t_i(p_{v-1})$  such that

$$t_i(p_k) = \inf\{t : p_k \leq 1 - \hat{S}_f(t|\mathbf{x}_i)\}, \quad p_k = k/v, \quad k = 1, \dots, v - 1.$$

Now we can concatenate  $v-1$  quantiles with the original feature vector. The value  $v$  is a tuning parameter. It should be noted that the quantiles of the time to event in the DSF play the same role as the class probability distributions in the DF.

The next feature we propose to use as an augmented feature is the mean time to the event  $a_i$  of the  $i$ -th patient, which can be simply computed from the SF  $\hat{S}_f(t|\mathbf{x}_i)$  by integrating. The motivation for choosing the mean value as an augmented feature for use at the next forest cascade levels is to reduce some bias in statistical characteristics of the corresponding random variables.

Finally, the vector  $\mathbf{A}^{(k,q)}$  of augmented features for the  $i$ -th patient, computed at the  $q$ -th level of the forest cascade by using the output of the  $k$ -th RSF, can be written



as

$$\mathbf{A}^{(k,q)} = \left( t_i^{(k,q)}(p_1), \dots, t_i^{(k,q)}(p_{v-1}), a_i^{(k,q)} \right).$$

Then the whole feature vector for training or testing at level  $q + 1$  is determined as follows:

$$\mathbf{x}_i^{(q+1)} \leftarrow (\mathbf{x}_i, \mathbf{A}^{(1,q)}, \dots, \mathbf{A}^{(K,q)}).$$

Here  $K$  is the number of RSFs at a level. We suppose for simplicity that all levels of the forest cascade have the same numbers of RSFs.

An important question raises with respect to the method for determining the concatenated feature vector for the next level. We have to define how to use the augmented features  $\mathbf{A}^{(k,q)}$ . In order to answer on this question, we propose to apply two main strategies and then to study them.

The first strategy defines the vector  $\mathbf{x}_i^{(q+1)}$  as concatenation of the original vector and the augmented features obtained at the previous  $q$ -th level. As a result, lengths of the vectors  $\mathbf{x}_i^{(q+1)}$  are identical for all levels, and they do not depend on levels. We remember only the previous level of the forest cascade.

The second strategy defines the vector  $\mathbf{x}_i^{(q+1)}$  as concatenation of the original vector and the augmented features obtained at all previous levels. In this case, we remember all previous levels of the forest cascade. The vector  $\mathbf{x}_i^{(q+1)}$  can also be defined as concatenation of the vector  $\mathbf{x}_i^{(q)}$  from the previous  $q$ -th level and the augmented features obtained at the  $q$ -th level, i.e., there holds

$$\mathbf{x}_i^{(q+1)} \leftarrow \left( \mathbf{x}_i^{(q)}, \mathbf{A}^{(1,q)}, \dots, \mathbf{A}^{(K,q)} \right).$$

Every strategy has some pros and cons. On the one hand, the second strategy is more informative in comparison with the first one because it exploits results obtained at all levels. On the other hand, the second strategy leads to complicated calculations especially at the last levels of the forest cascade. Moreover, worse learning results at the first levels may negatively impact on the final results. Similar arguments in favor, or against, of the first strategy can be provided. An optimal choice of strategy can be determined only by studying a certain dataset.

Figure 2 illustrates one of the possible implementations of the DSF architecture. The considered DSF uses the first strategy of using augmented features. One can see from Fig. 2 that every level of the forest cascade contains three RSFs. So, three vectors of augmented features  $\mathbf{A}^{(1,q)}$ ,  $\mathbf{A}^{(2,q)}$ ,  $\mathbf{A}^{(3,q)}$  are used at every level. The output of the last level consists of three CHF in Fig. 2. The output of the whole DSF is a CHF computed by averaging CHFs obtained at the last level. It can be seen from Figs. 1 and 2 that the architectures of the DF and the DSF are similar. However, they are different because they consist of quite different components.

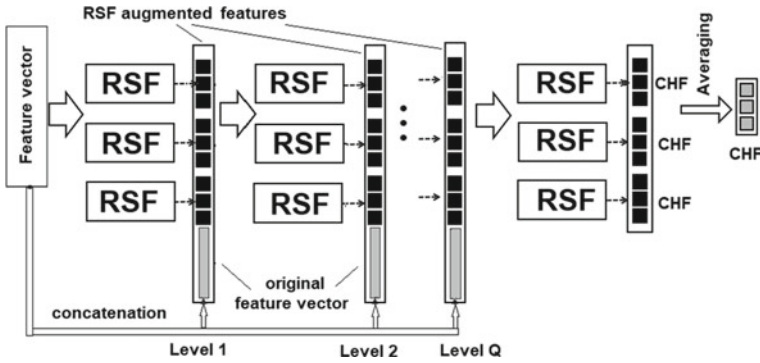


Fig. 2 DSF architecture

The ENCC has the same architecture, but every RSF is replaced with the Elastic-Net-Cox model.

## 6 Numerical Experiments

In order to investigate the proposed DSF and ENCC models, we use the following publicly available datasets:

The chronic myelogenous leukemia survival (CML) dataset consists of 507 observations (7 features). The dataset can be obtained via the “multcomp” R package.

The lupus nephritis dataset (LND) contains data on 87 patients (3 features). The dataset is from <https://www.stat.rice.edu/~sneeley/STAT553/Datasets/survivaldata.txt>.

The heart transplant dataset (HTD) contains data on 69 patients (2 features) receiving heart transplants. The dataset can be obtained from <https://lib.stat.cmu.edu/datasets/stanford>.

The gastric cancer dataset (GCD) contains data on survival of 90 patients (4 features) with locally advanced, non-resectable gastric carcinoma. The dataset can be obtained via the “coxphw” R package.

The DSF and the ENCC are implemented using Python. Cross-validation is applied with 100 repetitions for evaluating the C-index such that 80% of data are used for training and 20% are for testing. Moreover, we used 8 quantiles as augmented features, the accumulation of augmented features with every level of the forest cascade, one RSF or Elastic-Net-Cox model at every level, and 100 decision trees in every RSF.

Figure 3 illustrates the C-index as a function of the number of cascade levels for the DSF and ENCC models obtained for the dataset CML. Numerical results corresponding to the DSF are depicted by the dash-and-dot line, results obtained for the ENCC are represented by the solid line. It can be seen from Fig. 3 that there is

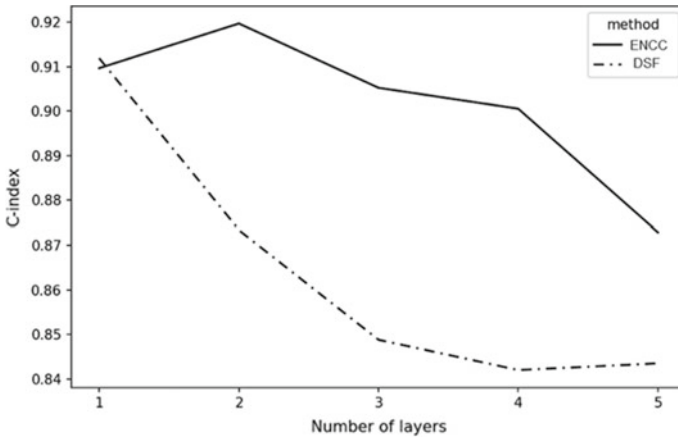


Fig. 3 Comparison of the DSF and ENCC models on the dataset CML

some improvement of the ENCC with the number of cascade levels. However, we also observe that the C-index of the DSF is significantly reduced.

Figure 4 shows the same dependencies, but for the dataset GCD. We again see from Fig. 4 that the ENCC outperforms the DSF for the number of levels larger than 2. We also see that the C-index of the ENCC is increased with the number of cascade levels.

Interesting results can be seen from Fig. 5 where the models are used to analyze the dataset LND. The DSF clearly outperforms the ENCC. Moreover, we observe the growth of the C-index of the DSF at the second level of the forest cascade.

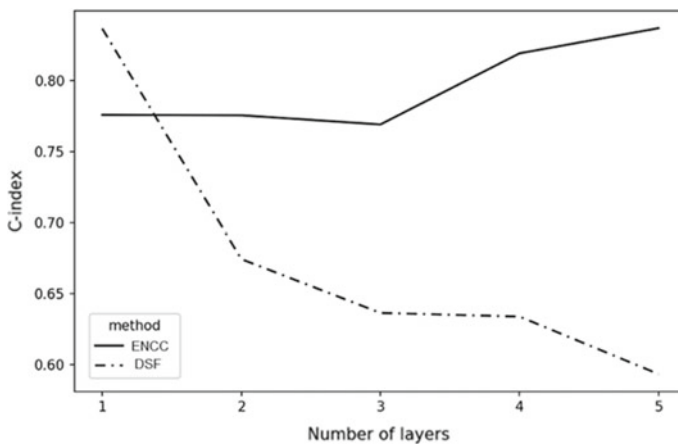


Fig. 4 Comparison of the DSF and ENCC models on the dataset GCD

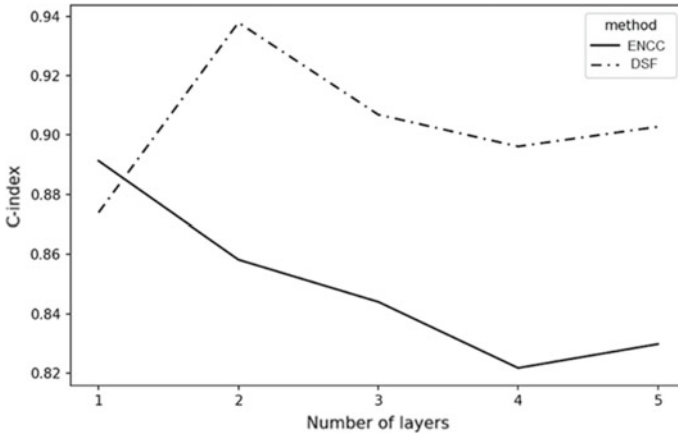


Fig. 5 Comparison of the DSF and ENCC models on the dataset LND

At the same time, we have to point out that the C-index may be reduced with levels of the cascade. For example, results for the HTD dataset shown in Fig. 6 demonstrate this reduction for the DSF as well as for the ENCC.

These numerical experiments imply that efficiency of the studied models strongly depends on the corresponding dataset and cannot be predicted before experiments.

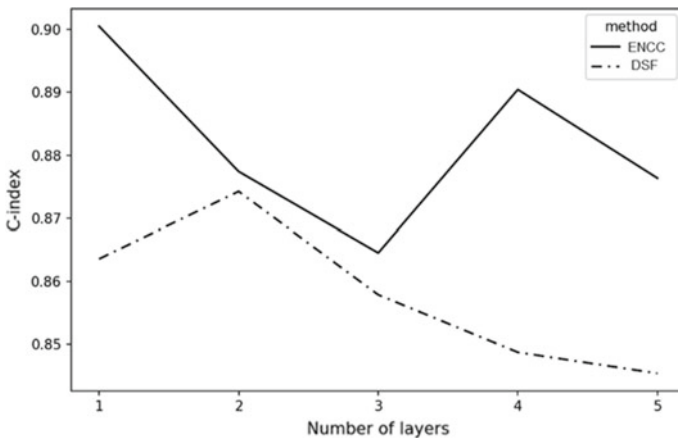


Fig. 6 Comparison of the DSF and ENCC models on the dataset HTD

## 7 Conclusion

Two new survival models have been proposed in this paper. The main idea underlying these models is to modify the DF by replacing the RFs with the RSFs and the Elastic-Net-Cox models. Moreover, a special stacking algorithm has been proposed in order to connect pairs of cascade levels. The proposed models contribute into survival analysis by improving the original RSFs and the Elastic-Net-Cox models which are considered as weak regressors.

Two main directions for further research can be pointed out. First, the models can be improved by using adaptive weighted schemes which can be viewed as extension of the adaptive weighted deep forest [31], where weights are assigned to training and testing examples in accordance with the results. The weights allow us to control the training and testing processes. Second, the models can be improved by implementing the feature selection procedure at every level of the cascade. This idea allows to reduce the training time significantly and may increase the survival analysis accuracy.

**Acknowledgements** The reported study was funded by RFBR, project number 18-29-03250.

## References

1. Hosmer, D., Lemeshow, S., May, S.: *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley, New Jersey (2008)
2. Cox, D.: Regression models and life-tables. *J. R. Stat. Soc., Ser. B (Methodological)* **34**, 187–220 (1972)
3. Lee, E., Wang, J.: *Statistical Methods for Survival Data Analysis*. Wiley, New Jersey (2003)
4. Kumar, D., Klefsjo, B.: Proportional hazards model: a review. *Reliab. Eng. Syst. Saf.* **44**, 177–188 (1994)
5. Scheike, T., Zhang, M.: Extensions and applications of the Cox-Aalen survival model. *Biometrics* **59**, 1036–1045 (2003)
6. Wang, P., Li, Y., Reddy, C.: Machine learning for survival analysis: a survey. [arXiv:1708.04649](https://arxiv.org/abs/1708.04649) (2017)
7. Zhou, D., Mital, D., Srinivasan, S., Shibata, M.: PenalisedCox regression models for survival data. *Int. J. Med. Eng. Inf.* **9**, 1–19 (2017)
8. Kim, J., Sohn, I., Jung, S.H., Kim, S., Park, C.: Analysis of survival data with group lasso. *Commun. Stat.—Simul. Comput.* **41**, 1593–1605 (2012)
9. Tibshirani, R.: The lasso method for variable selection in the cox model. *Stat. Med.* **16**, 385–395 (1997)
10. Zhang, H., Lu, W.: Adaptive Lasso for Cox’s proportional hazards model. *Biometrika* **94**, 691–703 (2007)
11. Witten, D., Tibshirani, R.: Survival analysis with high-dimensional covariates. *Stat. Methods Med. Res.* **19**, 29–51 (2010)
12. Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for Cox’s proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011)
13. Biganzoli, E., Boracchi, P., Marubini, E.: A general framework for neural network models on censored survival data. *Neural Netw.* **15**, 209–218 (2002)
14. Eleuteri, A., Tagliaferri, R., Milano, L., Placido, S.D., Laurentiis, M.D.: A novel neural network-based survival analysis model. *Neural Netw.* **16**, 855–864 (2003)

15. Faraggi, D., Simon, R.: A neural network model for survival data. *Stat. Med.* **14**, 73–82 (1995)
16. Luck, M., Sylvain, T., Cardinal, H., Lodi, A., Bengio, Y.: Deep learning for patient-specific kidney graft survival analysis. [arXiv:1705.10245](https://arxiv.org/abs/1705.10245) (2017)
17. Nezhad, M., Sadati, N., Yang, K., Zhu, D.: A deep active survival analysis approach for precision treatment recommendations: application of prostate cancer. [arXiv:1804.03280v1](https://arxiv.org/abs/1804.03280v1) (2018)
18. Belle, V.V., Pelckmans, K., Suykens, J., Huffel, S.V.: Survival SVM: a practical scalable algorithm. In: ESANN, pp. 89–94 (2008)
19. Khan, F., Zubek, V.: Support vector regression for censored data (SVRc): A novel tool for survival analysis. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 863–868. IEEE (2008)
20. Polsterl, S., Navab, N., Katouzian, A.: An efficient training algorithm for kernel survival support vector machines. [arXiv:1611.07054v](https://arxiv.org/abs/1611.07054v) (2016)
21. Kiaee, F., Sheikhzadeh, H., Mahabadi, S.: Relevance vector machine for survival analysis. *IEEE Trans. Neural Netw. Learn. Syst.* **27**, 648–660 (2015)
22. Breiman, L.: Random Forests. *Mach. Learn.* **45**, 5–32 (2001)
23. Bou-Hamad, I., Larocque, D., Ben-Ameur, H.: Discrete-time survival trees and forests with time-varying covariates: application to bankruptcy data. *Stat. Model.* **11**, 429–446 (2011)
24. Hu, C., Steingrimsson, J.: Personalized risk prediction in clinical oncology research: applications and practical issues using survival trees and random forests. *J. Biopharm. Stat.* **28**, 333–349 (2018)
25. Ishwaran, H., Blackstone, E., Pothier, C., Lauer, M.: Relative risk forests for exercise heart rate recovery as a predictor of mortality. *J. Am. Stat. Assoc.* **99**, 591–600 (2004)
26. 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)
27. Harrell, F., Califf, R., Pryor, D., Lee, K., Rosati, R.: Evaluating the yield of medical tests. *J. Am. Med. Assoc.* **247**, 2543–2546 (1982)
28. Zhou, Z.H., Feng, J.: Deep forest: towards an alternative to deep neural networks. [arXiv:1702.08835v2](https://arxiv.org/abs/1702.08835v2) (2017)
29. Tibshirani, R.: Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B (Methodological)* **58**, 267–288 (1996)
30. Ishwaran, H., Kogalur, U., Blackstone, E., Lauer, M.: Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008)
31. Utkin, L., Konstantinov, A., Meldo, A., Ryabinin, M., Chukanov, V.: A deep forest improvement by using weighted schemes. In: Proceedings of the 24th Conference of Open Innovations Association FRUCT, Moscow, Russia, pp. 451–456. IEEE (2019).

# An Explanation Method for Siamese Neural Networks



Lev Utkin, Maxim Kovalev, and Ernest Kasimov

**Abstract** A new method for explaining the Siamese neural network is proposed. It uses the following main ideas. First, the explained feature vector is compared with the prototype of the corresponding class computed at the embedding level (the Siamese neural network output). The important features at this level are determined as features which are close to the same features of the prototype. Second, an autoencoder is trained in a special way in order to take into account the embedding level of the Siamese network, and its decoder part is used for reconstructing input data with the corresponding changes. Numerical experiments with the well-known dataset MNIST illustrate the propose method.

**Keywords** Interpretable model · Explainable intellect · Siamese neural network · Prototype · Embedding · Autoencoder

## 1 Introduction

Deep models play an important role in making prediction for many applications. However, a lot of machine learning techniques are not explainable, they are black boxes and do not explain their predictions. This may be a problem for applying the models to various field, for example, to medicine. Therefore, a lot of explanation models have been developed, which can be viewed as special meta-models for explaining the deep model predictions [1, 2].

A lot of machine learning models are regarded as black boxes, i.e., it is assumed that we do not know any details of the black-box model, for example, its structure, parameters, etc., but its input and the corresponding output are known and can be used for training the explanation model. A well-known method is the Local Interpretable Model-agnostic Explanations (LIME) [3]. According to the LIME, the explanation may be derived locally from randomly generated synthetic neighbor examples. There are also modifications of the LIME, for example, ALIME [4], NormLIME [5],

---

L. Utkin (✉) · M. Kovalev · E. Kasimov  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [lev.utkin@gmail.com](mailto:lev.utkin@gmail.com)

DLIME [6]. Another well-known method is the SHAP [7] and its modifications [8, 9]. It should be noted that there are also alternative methods, for example, influence functions [10], a multiple hypothesis testing framework [11], counterfactual explanations [12].

Some explanation methods use a prototype technique which selects representative instances (prototypes) from the training data, for instance, from examples belonging to the same class [13, 14].

Another interesting idea realized in many explanation methods is the perturbation technique [15, 16]. These methods assume that contribution of a feature can be determined by measuring how prediction score changes when the feature is altered [7]. At the same time, the perturbation technique may be computationally hard when perturbed inputs have a lot of features, for example, pictures.

We have to point out a number of interesting survey papers devoted to explainable methods [17–19], which cover many questions related to the methods.

We consider an approach which is agnostic to the black-box model. This means that we do not know or do not use any details of the black-box model. Only its input and the corresponding output are used for training the explanation model.

To the best of our knowledge, there are no appropriate algorithms for explaining the Siamese neural network (SNN). Therefore, we propose a method to explain the SNN [20, 21] as the black-box model. The SNN consists of two identical neural subnets sharing the same set of weight. The SNN aims to compare a pair of feature vectors in terms of their semantic similarity or dissimilarity. It realizes a nonlinear embedding of data with the objective to bring together similar examples and to move apart dissimilar examples. SNNs have been applied to various problems, including image recognition and verification, visual tracking, novelty and anomaly detection, one-shot and few-shot learning [21–27].

Problems for explaining the SNN stem from two main reasons. First, the input examples for the SNN are semantically similar or dissimilar, and direct distances between them may not have a sense. Second, there is no an inverse map between the embeddings and the corresponding input examples, i.e., we do not know subsets of features in the input vector corresponding to some features of the embedded vector.

We try to solve these problems by applying the following ideas. First, we find prototypes of all classes at the embedding level and select features having the smallest Euclidean distance between the embedding of the explained example and the prototype. Second, we train an autoencoder with a special loss function which takes into account embeddings obtained by means of the SNN. The decoder part of the autoencoder is used to reconstruct the introduced perturbations to observe features of the reconstructed example with largest changes. These features are nothing else, but the explanation of interest.

The paper is organized as follows. A description of the SNN and its peculiarities are given in Sect. 2. Two important concepts of explanation methods, the perturbation technique and prototypes, are considered in the same section. The proposed explanation method, which is the aim of the paper, is provided in Sect. 3. Numerical experiments illustrating the proposed method on the basis of the well-known MNIST dataset are studied in Sect. 4. Concluding remarks are provided in Sect. 5.



## 2 Siamese Neural Networks, Perturbations, and Prototypes

Let  $\{(\mathbf{x}_i, y_i), i = 1, \dots, n\}$  be a dataset consisting of  $n$  feature vectors  $\mathbf{x}_i \in \mathbf{R}^m$  of size  $m$  with labels  $y_i \in \{1, 2, \dots, C\}$ . Let us construct a new training set  $S = \{(\mathbf{x}_i, \mathbf{x}_j, z_{ij}), (i, j) \in K\}$  consisting of pairs of examples  $\mathbf{x}_i$  and  $\mathbf{x}_j$  with binary labels  $z_{ij} \in \{0, 1\}$  assigned to them. If both feature vectors  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are semantically similar, i.e., they belong to the same class, then  $z_{ij}$  is 0. If the vectors are semantically dissimilar, i.e., they correspond to different classes, then  $z_{ij}$  is 1. So, the training set  $S$  can be divided into two subsets: a similar or positive set with  $z_{ij} = 0$  and a dissimilar or negative set with  $z_{ij} = 1$ .

Suppose new feature representations of the input examples  $\mathbf{x}_i$  and  $\mathbf{x}_j$  as the SNN outputs are  $\mathbf{h}_i \in \mathbf{R}^D$  and  $\mathbf{h}_j \in \mathbf{R}^D$ , respectively. The SNN realizes a map  $f$  such that  $\mathbf{h}_i = f(\mathbf{x}_i)$ , which tries to make the Euclidean distance  $d(\mathbf{h}_i, \mathbf{h}_j)$  as small (large) as possible for the similar (dissimilar) pair of objects. A standard architecture of the SNN is shown in Fig. 1.

It should be noted that there are many specific loss functions for training the SNN [28]. We use a function called the contrastive loss function. It is defined as

$$l(\mathbf{x}_i, \mathbf{x}_j, z_{ij}) = (1 - z_{ij})\|\mathbf{h}_i - \mathbf{h}_j\|_2^2 + z_{ij} \max(0, \tau - \|\mathbf{h}_i - \mathbf{h}_j\|_2^2).$$

Here  $\tau$  is a predefined threshold. Hence, the total error function for minimizing is defined as

$$L(W) = \sum_{i,j} l(\mathbf{x}_i, \mathbf{x}_j, z_{ij}) + \mu R(W).$$

Here  $R(W)$  is a regularization term added to improve generalization of the neural network;  $W$  is the matrix of the neural subnet parameters;  $\mu$  is a hyper-parameter which controls the strength of the regularization. The above problem is usually solved by using a gradient descent scheme.

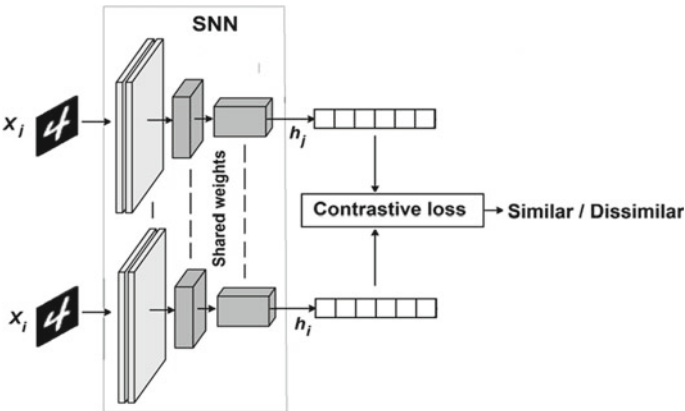


Fig. 1 An architecture of the SNN

Let us consider definitions of perturbations and prototypes, which will be used in the proposed explanation method. It has been mentioned that some explanation methods are based on applying perturbation schemes which explicitly test the explained model's response to local perturbations. The intuition behind the technique is that the more a model's response depends on a feature, the more predictions change with the corresponding feature changes. The perturbation scheme  $\rho$  perturbs all features of  $\mathbf{x}$  into  $\hat{\mathbf{x}}$  as follows:  $\rho(\mathbf{x}, \delta) = \hat{\mathbf{x}} = \mathbf{x} + \delta$ . Here  $\delta$  is the perturbation vector. In many cases, finding the optimal perturbing scheme for all instances is intractable due to a possible large dimensionality of input examples. Therefore, various techniques are available to simplify this procedure.

Following the work of Snell et al. [29], a prototype is a data example that is representative of a subset of data, for example, a set of examples from a class. If we have  $D$ -dimensional representation of every  $\mathbf{x}_i$  through an embedding function  $f : \mathbf{R}^m \rightarrow \mathbf{R}^D$ , then the prototype  $\mathbf{c}_k \in \mathbf{R}^D$  of class  $k$  is defined as [29]:

$$\mathbf{c}_k = \frac{1}{n_k} \sum_{i:y_i=k} f(\mathbf{x}_i) = \sum_{i:y_i=k} \mathbf{h}_i.$$

### 3 The Proposed Method for the SNN Explanation

The proposed method for explaining the SNN can be represented by means of an algorithm consisting of two parts. The first part aims to train the additional autoencoder with a special loss function. This autoencoder can be called as an explainable autoencoder. The goal of the second part is to perturb the embedding vector at the SNN output and to use the decoder of the trained autoencoder in order to reconstruct the perturbed embeddings and to observe the features which are changed.

Suppose we have a trained SNN as a black box. For every input vector  $\mathbf{x}_i$ , we have the corresponding embedding vectors  $\mathbf{h}_i$  such that  $\mathbf{h}_i = f(\mathbf{x}_i)$ . The main idea to incorporate the additional autoencoder is the following.

Suppose we have an embedding vector  $\mathbf{h}$  with a set of important features. However, these features do not explain why the considered explained example belongs to a class, say to class  $k$ . In order to answer this question, it is necessary to find an inverse mapping from  $\mathbf{h}$  to  $\mathbf{x}$ , i.e., the vector  $\mathbf{x}$  has to be reconstructed from  $\mathbf{h}$ . The reconstruction can be carried out by means of a neural network with input values  $\mathbf{h}_i$  and output values  $\mathbf{x}_i$ . Our numerical experiments have demonstrated that it is difficult to train a reconstruction neural network especially when the dimension of vectors  $\mathbf{x}$  is very large and the number of training examples is small due to possible overfitting of the network. It is simpler to train an autoencoder and then to use its trained decoder part for reconstruction. In order to exploit the decoder for reconstruction of vector  $\mathbf{h}$ , the autoencoder has to be trained in a special way. First of all, the length  $D$  of its code (the hidden representation) has to coincide with the length of vector  $\mathbf{h}$ . The loss

function should take into account the proximity of vectors  $\mathbf{h}_i$  and the corresponding vectors of the autoencoder hidden representation.

Suppose that the input examples for the proposed incorporated autoencoder are vectors  $\mathbf{x}_i$ , then we expect to get reconstructed vectors  $\tilde{\mathbf{x}}_i$  as its outputs. The corresponding loss function  $L_{\text{recon\_a}}$  for training the autoencoder is defined as follows:

$$L_{\text{recon\_a}}(W) = \sum_{i=1}^n \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|_2^2.$$

We do not write a regularization term because it will be used later. However, we cannot apply the autoencoder in its standard form because we need to have the vectors  $\mathbf{h}_i^*$  in the hidden layer coinciding with the vectors  $\mathbf{h}_i$  obtained by means of the SNN. Therefore, we propose to change the loss function for training the autoencoder by adding the loss function  $L_{\text{close}}$  in the following way:

$$\begin{aligned} L_{\text{autoen}}(W) &= \gamma L_{\text{recon\_a}}(W) + \mu L_{\text{close}}(W) + \lambda R(W) \\ &= \gamma \sum_{i=1}^n \|\mathbf{x}_i - \tilde{\mathbf{x}}_i\|_2^2 + \mu \sum_{i=1}^n \|\mathbf{h}_i - \tilde{\mathbf{h}}_i\|_2^2 + \lambda R(W). \end{aligned}$$

Here  $R(W)$  is a regularization term,  $\lambda$  is a hyper-parameter which controls the strength of the regularization;  $W$  is the set of the neural network weights;  $\gamma$  and  $\mu$  are weights that control the interaction of the loss terms;  $\tilde{\mathbf{h}}_i$  are the autoencoder hidden representation vectors.

We can now use the decoder part for reconstruction of the perturbed embeddings that is for implementing the second part of the algorithm. This trick allows us to significantly simplify the training process and to get acceptable vector reconstructions. It should be noted that an architecture of the encoder differs from the architecture of a subnetwork of the SNN because we consider the SNN as a black box whose architecture is unknown. A scheme of the explanation algorithm first part is shown in Fig. 2. It can be seen from the training scheme that the autoencoder is trained by using embedding vectors from one of the SNN subnetworks.

Now we consider the second part of the explanation algorithm under condition that there is available the trained decoder for reconstruction. A schematic representation of the part is shown in Fig. 3. It is based on using prototypes and perturbations. By having embedding vectors  $\mathbf{h}_i$  for all training examples  $\mathbf{x}_i$ , we can compute prototypes  $\mathbf{c}_k \in \mathbf{R}^D$  for every class  $k$  as it is shown in Sect. 2 based on embedded vectors  $\mathbf{h}_i$  such that  $y_i = k$ . Without loss of generality, we suppose that an explained example  $\mathbf{x}$  belongs to class  $k$ . It is obvious in this case, that the explained example and the prototype are semantically similar (of course if the SNN is correctly classified the example). This implies that the Euclidean distance between the embedded vector  $\mathbf{h}$  of the explained example and the prototype  $d(\mathbf{h}, \mathbf{c}_k)$  should be smaller than the

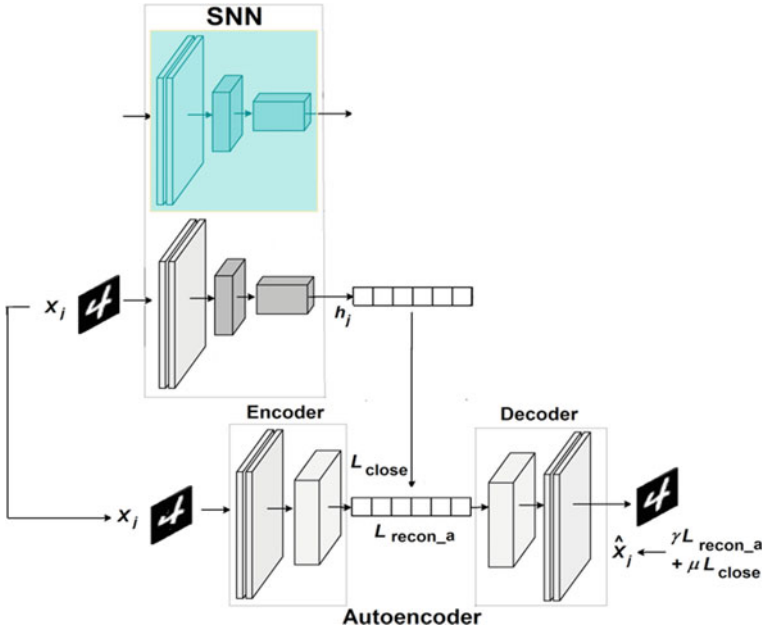


Fig. 2 Autoencoder training part of the explanation algorithm

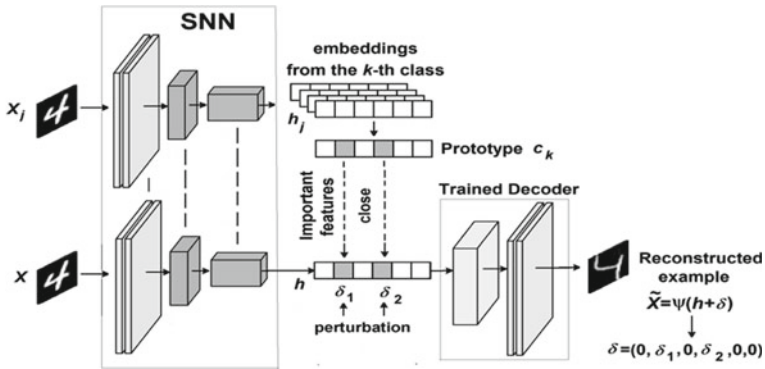


Fig. 3 Second part of the explanation algorithm

Euclidean distance between the embedded vector of the explained example and the prototypes of other classes  $d(\mathbf{h}, \mathbf{c}_i), i \neq k$ .

So, we have vectors  $\mathbf{h}$  and  $\mathbf{c}_k$  consisting of  $D$  features. It is obvious that features in  $\mathbf{h}$ , which are close to the corresponding features in  $\mathbf{c}_k$ , can be viewed as important features. These important features define the fact that the explained example belongs

to class  $k$ . Therefore, they should be selected. Let us introduce the number of important features  $s < D$  such that the index set  $J \subseteq \{1, \dots, D\}$  consists of  $s$  indices corresponding to smallest distances between  $h_i$  and  $c_i^{(k)}$ ,  $i = 1, \dots, D$ . Denote ordered  $s$  features of  $\mathbf{h}$  as  $h_{(i)}$ ,  $i = 1, \dots, s$ . Then by perturbing the embedded vector  $\mathbf{h}$  (features  $h_{(1)}, \dots, h_{(s)}$ ), we can construct a training set which consists vectors  $\mathbf{h} + \delta_i$ . Here  $\delta_i$  is the perturbation vector such that indices of its non-zero elements are from the index set  $J$ , other elements are equal to zero.

In order to study how the important features of the hidden representation impact on the original vector  $\mathbf{x}$ , we use the trained decoder to reconstruct vectors from  $\mathbf{h} + \delta_i$  and to investigate how features of the reconstructed vector  $\tilde{\mathbf{x}} \in \mathbf{R}^m$  are changed. In sum, we have the embedding vector  $\mathbf{h}$ , the reconstruction  $\tilde{\mathbf{x}}$ , the index set  $J$  of important features of  $\mathbf{h}$ . The trained decoder implements a function  $\psi : \mathbf{R}^D \rightarrow \mathbf{R}^m$ , i.e.,  $\tilde{\mathbf{x}} = \psi(\mathbf{h})$ . In order to determine the important features of  $\tilde{\mathbf{x}}$ , we perturb the important features of  $\mathbf{h}$  and to observe which features of  $\tilde{\mathbf{x}}$  have the largest changes. We use the random perturbation of important features of  $\mathbf{h}$  when every feature is added with a positive random value. Moreover, the pre-trained decoder is again trained by using only the SNN output embeddings.

Let  $q$  be the number of important features  $\tilde{\mathbf{x}}$ . By generating the random vector  $\delta$  many times, say  $N$  times, and observing changes of  $\tilde{\mathbf{x}}$ , we can compute mean realize changes of all features. Then features, having  $q$  largest changes of the reconstructed example, explain the considered example.

Let us return to the scheme in Fig. 3. It can be seen from the scheme that one subnetwork in the SNN is conditionally used for getting the vector  $\mathbf{h}$ . Another subnetwork provides a set of embedding vectors in order to compute the prototype  $\mathbf{c}_k$ . The important features (two features) in  $\mathbf{h}$  are shown by dashed cells. They are close to the same features in the prototype. The reconstruction network (decoder) is shown on the right side of the picture. The perturbed vector is fed to the decoder in order to get the reconstructed vector which depends on perturbations.

Perturbation vectors are randomly generated with respect to the normal distribution with zero expectations and small variances of  $s$  features in accordance with the index set  $J$ . We take only positive perturbations because changes of features should be closer to the prototype.

## 4 Numerical Experiments

The proposed explanation method is studied by applying the MNIST dataset which is a commonly used large dataset of  $28 \times 28$  pixel handwritten digit images [30]. It has a training set of 60,000 examples, and a test set of 10,000 examples. The digits are size-normalized and centered in a fixed-size image. The dataset is available at <https://yann.lecun.com/exdb/mnist/>.

The length of the hidden representation layer is 10, i.e., the vector  $\mathbf{h}$  consists of 10 features. The autoencoder implementation uses ReLU as activation function for all layers except for the last layer where a sigmoid activation function is used.

Perturbations are sampled from the normal distribution with zero mean and standard deviation  $0.1 \min\{|h_1 - c_{k1}|, \dots, |h_D - c_{kD}|\}$ , where  $h_i$  is the  $i$ -th component of vector  $\mathbf{h}$ ,  $c_i^k$  is the  $i$ -th component of the prototype  $\mathbf{c}_k$ . The number of perturbations is 5000.

An architecture of the autoencoder is given in Table 1. The architecture contains an encoder (the first and second columns) and an equivalent decoder (the third and fourth columns). The encoder comprises convolution layers (Conv), max pooling operations (Pooling), flatten layers (Flatten) which flatten a matrix input to a simple vector, dense layers (Dense) which are a fully connected layer. The decoder block has additionally deconvolution layers (UpSampling), reshape layers (Reshape) which change the dimensions of its input without changing its data.

We show below quadruples of pictures such that the first picture in every quadruple is an original image of a digit, the second picture is the reconstructed digit, the third picture is the original image and the corresponding mask of explanation features, the fourth picture is the explanation feature in the form of the mask. The explanation features can be regarded as correct if they clearly show difference of the considered digit belonging to the classified class from digits belonging to other classes.

Four examples of the correct explanation of digits from MNIST are shown in Figs. 4, 5, 6 and 7. It can be seen from the pictures that all original digits are perfectly reconstructed by the trained decoder. However, the quality of explanation depends on the reconstructed images. Figure 8 illustrates an example of the incorrect explanation when the reconstructed image significantly differs from the original image. This implies that the autoencoder is not perfectly trained or its architecture does not allow us to efficiently reconstruct all images from the testing set. Another interesting case is when the digits are incorrectly classified by the SNN. This case is demonstrated in Fig. 9, where the original digit 4 is classified by the SNN as the digit 9. As a result, the explainer selects features which actually indicate the digit 9 instead of 4. In fact, this

**Table 1** Autoencoder architecture

Encoder		Decoder	
Layer	Output	Layer	Output
Input	$28 \times 28 \times 1$	Input	20
Conv1	$28 \times 28 \times 16$	Dense1	40
Pooling1	$14 \times 14 \times 16$	Dense2	128
Conv2	$14 \times 14 \times 8$	Reshape	$4 \times 4 \times 8$
Pooling2	$7 \times 7 \times 8$	UpSampling1	$8 \times 8 \times 8$
Conv3	$7 \times 7 \times 8$	Conv1	$7 \times 7 \times 8$
Pooling3	$4 \times 4 \times 8$	UpSampling2	$14 \times 14 \times 8$
Flatten	128	Conv2	$14 \times 14 \times 16$
Dense1	40	UpSampling3	$28 \times 28 \times 16$
Dense2	20	Conv3	$28 \times 28 \times 1$



Fig. 4 Explanation of the digit 1

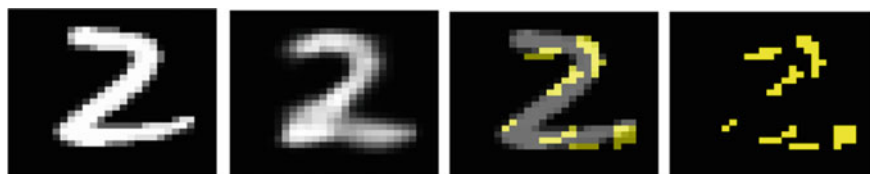


Fig. 5 Explanation of the digit 2

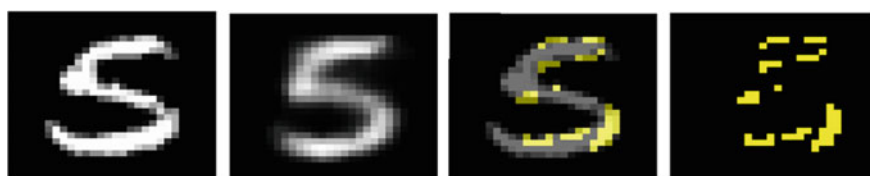


Fig. 6 Explanation of the digit 5



Fig. 7 Explanation of the digit 9

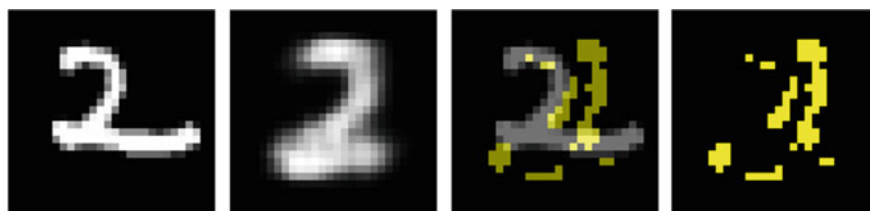
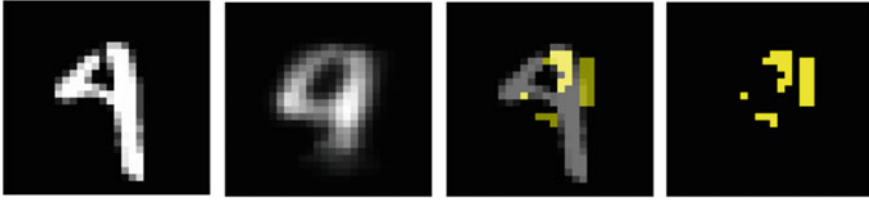


Fig. 8 Incorrect explanation of the digit 2



**Fig. 9** Digit 4 is incorrectly classified to 9

case shows that the proposed method correctly explains, but the explanation depends on the correctness of the black-box model classification.

## 5 Conclusion

A new method for explaining the SNN has been presented in the paper. The main ideas underlying the method are comparison of the explained example with a prototype at the embedding level and reconstruction of the embedding feature vectors by means of a separately trained special autoencoder in order to analyze the impact of the embedding vector perturbations on the reconstructed features. The proposed method can be applied to various problems which use SNNs.

It is important to note that the SNN can be regarded as a part of a general distance metric learning approach. Therefore, applications of the proposed explanation method can be extended on various models of the distance metric learning. One of the interesting directions for the extension is the novelty and anomaly detection because this problem has a huge amount of applications.

A bottleneck of the proposed model is the autoencoder which has to be trained by using the dataset. The problem arises when the analyzed dataset is rather small. Ways for solving the problem can be regarded as a direction for further research. Another problem is that the method is based on the random perturbations. At the same time, there are a lot of interesting works, for example, [31] or [32], where perturbations are determined in an optimal way by solving the corresponding optimization problems. The use of this approach to modifying the proposed method is another direction for further research.

**Acknowledgements** This work is supported by the Russian Science Foundation under grant 18-11-00078. with the phrase: The reported study was funded by RFBR, project number 20-01-00154.



## References

1. 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** (2019), Article 93, 1–42
2. Molnar, C.: *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Published Online. <https://christophm.github.io/interpretable-ml-book/> (2019)
3. 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)
4. Shankaranarayana, S., Runje, D.: Alime: autoencoder based approach for local interpretability. [arXiv:1909.02437](https://arxiv.org/abs/1909.02437) (2019)
5. Ahern, I., Noack, A., Guzman-Nateras, L., Dou, D., Li, B., Huan, J.: NormLIME: A new feature importance metric for explaining deep neural networks. [arXiv:1909.04200](https://arxiv.org/abs/1909.04200) (2019)
6. Zafar, M., Khan, N.: DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. [arXiv:1906.10263](https://arxiv.org/abs/1906.10263) (2019)
7. Strumbel, E., Kononenko, I.: An efficient explanation of individual classifications using game theory. *J. Mach. Learn. Res.* **11**, 1–18 (2010)
8. Aas, K., Jullum, M., Loland, A.: Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. [arXiv:1903.10464](https://arxiv.org/abs/1903.10464) (2019)
9. Ancona, M., Oztireli, C., Gros, M.: Explaining deep neural networks with a polynomial time algorithm for Shapley values approximation. [arXiv:1903.10992](https://arxiv.org/abs/1903.10992) (2019)
10. Koh, P., Liang, P.: Understanding black-box predictions via influence functions. In: *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, pp. 1885–1894 (2017)
11. Burns, C., Thomason, J., Tansey, W.: Interpreting black box models with statistical guarantees. [arXiv:1904.00045](https://arxiv.org/abs/1904.00045) (2019)
12. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv. J. Law Technol.* **31**, 841–887 (2017)
13. Bien, J., Tibshirani, R.: Prototype selection for interpretable classification. *The Ann. Appl. Stat.* **5**, 2403–2424 (2011)
14. Kim, B., Rudin, C., Shah, J.: The Bayesian case model: a generative approach for case-based reasoning and prototype classification. In: *Advances in Neural Information Processing Systems*, pp. 1952–1960 (2014)
15. Fong, R., Vedaldi, A.: Explanations for attributing deep neural network predictions. In: *Explainable AI*, vol. 11700 of LNCS, pp. 149–167. Springer, Cham (2019)
16. Vu, M., Nguyen, T., Phan, N., R. Gera, M.T.: Evaluating explainers via perturbation. [arXiv:1906.02032v1](https://arxiv.org/abs/1906.02032v1) (2019)
17. Gilpin, L., Bau, D., Yuan, B., Bajwa, A., Specter, M., Kagal, L.: Explaining explanations: An overview of interpretability of machine learning. in 2018. In: *IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pp. 80–89. IEEE (2018)
18. Mohseni, S., Zarei, N., Ragan, E.: A survey of evaluation methods and measures for interpretable machine learning. [arXiv:1811.11839](https://arxiv.org/abs/1811.11839) (2018)
19. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1**, 206–215 (2019)
20. Bromley, J., Bentz, J., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Sackinger, E., Shah, R.: Signature verification using a Siamese time delay neural network. *Int. J. Pattern Recognit Artif Intell.* **7**, 737–744 (1993)
21. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, 539–546. IEEE (2005)
22. Koch, G., Zemel, R., Salakhutdinov, R.: Siamese neural networks for one-shot image recognition. In: *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, pp. 1–8. Lille, France (2015)
23. Hu, J., Lu, J., Tan, Y.P.: Discriminative deep metric learning for face verification in the wild. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1875–1882. IEEE (2014)

24. Zhang, C., Liu, W., Ma, H., Fu, H.: Siamese neural network based gait recognition for human identification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2832–2836. IEEE (2016)
25. Jiang, C., Xiao, J., Xie, Y., Tillo, T., Huang, K.: Siamese network ensemble for visual tracking. *Neurocomputing* **275**, 2892–2903 (2018)
26. Masana, M., Ruiz, I., Serrat, J., van de Weijer, J., Lopez, A.: Metric learning for novelty and anomaly detection. [arXiv:1808.05492](https://arxiv.org/abs/1808.05492) (2018)
27. Utkin, L., Lukashin, A., Popov, S., Zaborovsky, V., Podolskaja, A.: A Siamese autoencoder preserving distances for anomaly detection in multi-robot systems. In: 2017 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO), Prague, Czech Republic, pp. 39–44. IEEE (2017)
28. Bellet, A., Habrard, A., Sebban, M.: A survey on metric learning for feature vectors and structured data. *arXiv preprint* [arXiv:1306.6709](https://arxiv.org/abs/1306.6709) (2013)
29. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. [arXiv:1703.05175v2](https://arxiv.org/abs/1703.05175v2) (2017)
30. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998)
31. Dhurandhar, A., Pedapati, T., Balakrishnan, A., Chen, P.Y., Shanmugam, K., Puri, R.: Model agnostic contrastive explanations for structured data. [arXiv:1906.00117](https://arxiv.org/abs/1906.00117) (2019)
32. Looveren, A.V., Klaise, J.: Interpretable counterfactual explanations guided by prototypes. [arXiv:1907.02584](https://arxiv.org/abs/1907.02584) (2019)

# Hierarchical Multi-agent System for Production Control Using KPI Reconciliation



Vladislav Kovalevsky , Vadim Onufriev , and Anton Dybov 

**Abstract** In this work, the task of production key performance indicators' values reconciliation is set, that takes into account their hierarchical structure and interrelationships. The structural scheme of a hierarchical multi-agent system for production control was developed where horizontal (sibling) and vertical (multilevel) connections between agents are shown. Then the possible situations of sibling and multilevel interactions are described, such as changing of a task by the controlling agent, sending notifications about the impossibility of a maintaining current operation mode, and others. The agents' data exchange algorithms, which are used in order to optimize key performance indicators, are shown. The implementation of developed algorithms using client-server architecture is shown, which also includes at the bottom level data exchange between the agents and programmable logic controllers. The single bytes command system for agents' interactions is described.

**Keywords** Cyber-Physical systems · Intelligent control systems · Multi-agent systems · Digital twin · Industry 4.0

## 1 Introduction

The modern trends of industry development are Industry 4.0 [1–3], IIoT [4], and Digital Twins [5–8]. According to these concepts, traditional centralized industry monitoring and control systems should be replaced with decentralized multi-agent systems where at the low level work emergency response equipment and programmable logic controllers and above them work independent agents.

In modern industry increases the role of unified system of factory production equipment development [9] so that all plant capacities ultimately work to achieve

---

V. Kovalevsky (✉) · V. Onufriev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [vladkov@spbstu.ru](mailto:vladkov@spbstu.ru)

A. Dybov  
Technical University of Berlin, Berlin, Germany

common goals, which are quantified in KPI (key performance indicators) [10]. If it comes about control of enterprise at whole and KPI describe work of all enterprise at high level then this KPI called high level KPI.

In subdivisions of the enterprise and at the shop floor, there are also KPI, but they are KPI of the lower level [11]. At the bottom level of this hierarchy of indicators lay parameters of shop floor equipment, their parts, and operation modes. They converted to KPI using neural network models of a bottom level. It is obvious that KPI of the bottom level should be in concordance with KPI of the higher levels, KPI of a shop floor, subdivision, or even the highest KPI of the whole enterprise. Hence, in the KPI hierarchy indicators of the higher levels define the required values of the KPI in lower levels and in the same time depend on them.

It is assumed in this paper that mathematical dependencies between indicators of different levels already given, for example, they can be calculated using approach described in the earlier work [8]. Multi-agent approach requires to define agents' interaction algorithms in order to make decisions through networking [12, 13]. The complexity of designing an end-to-end enterprise control system is relatively high, thus in most cases only decision support systems that help human operators make decisions are developed [14, 15].

## 2 The Hierarchical Multi-agent Control Systems

Various research teams have studied the task of multi-agent approach implementation in industrial settings. Various papers suggest using this approach for diverse aspects of manufacturing process. In particular, the following use cases for multi-agent approach in industrial settings can be distinguished: (1) production planning, (2) fault diagnostics, (3) production control, (4) flexible and scalable reconfiguration of assembly lines, (5) virtual enterprises, supply, and outsourcing management. In many cases, multi-agent control systems consist not of fully autonomous single-level agents, but of hierarchy of agents, where top level agent receives only data that he needs for his work, and coordinates the work of dependent agents but does not dictate them specific actions to perform.

Paper [16] describes hierarchical distributed manufacturing control system, where each level of production hierarchy is controlled by intelligent agents of this level. Another example of hierarchical multi-agent system is given in paper [17]. There MAS is used to control multi-microgrid system that consists of large number of local power generator and supply devices. Paper [18] deals with control of ship cooling system. The control system in this case consists of three level hierarchy of agents, where top level agent decomposes tasks and gives these decomposed subtasks to his subordinate lower level agents, and lowest level agents control directly the hardware. Paper [19] delves into usage of multi-agent approach in production shop floor. Here MAS consists of agent that allocates production tasks, agents that control production equipment and auxiliary agents in between.

When developing interactions between agents in hierarchical control system, it is important to take into consideration several possible scenarios in which the need for data exchange arises. For instance in situation when at the level of strategic planning the production task was changed, and this change entails recalculation of KPI at lower levels. Or opposite situation of down-top interaction, that is similar to diagnostics scenario that described in [8], when system continuously checks its own state comparing it with KPI values in its knowledge base, and should the need arise sends requests to higher levels. But in the stated work the system was centralized, all data was accumulated in single center that was responsible for control decisions making.

The above-mentioned papers show various designs and possible applications of hierarchical multi-agent control system. But either they describe only one-way data flow from agent to agent, or the exchanged data is too detailed, or there is need for human operator in the middle.

The goal of this work is, after analysis of current works in the subject area and their shortcomings, to develop autonomous control system that by using multi-agent approach and key performance indicators of each level of production hierarchy enables flexible change of production plans and their adjustment in case of a production task or a shop floor situation change.

For this purpose, next tasks should be addressed: develop unified scheme of control agents distribution among processes and indicators through all levels of production hierarchy; develop appropriate algorithmic models of agents'; develop KPI data exchange protocols; and suggest set of commands to exchange in the given scenarios.

In order to develop hierarchical system that consists of agents that control every level of production hierarchy, it needs to decompose production processes, break them down to subprocesses and so on till the bottom level of production hierarchy. Then for every such subprocess its KPI and KPI limits should be identified, and after that neural network models of KPI interrelations between adjacent levels should be calculated, for example, like it is suggested in works [8, 20]. After that agents that responsible for every process in the hierarchy and its KPI are developed, using the information that was calculated in the previous step. So every agent in the resulting system is connected to some subprocess and controls it.

Taking into account the above-mentioned steps we suggest the next scheme of distribution of control agents among processes and indicators in production hierarchy, using IDEF0 approach (Fig. 1).

At the bottom level of the hierarchy in Fig. 1 agents are PLCs that are controlling processes of the lowest level.

After the hierarchical structure of agents distribution among processes is defined the next step is to develop algorithms of data exchange between agents, taking into account that this interaction can be done by agents of the same level and as well by agents of adjacent levels.

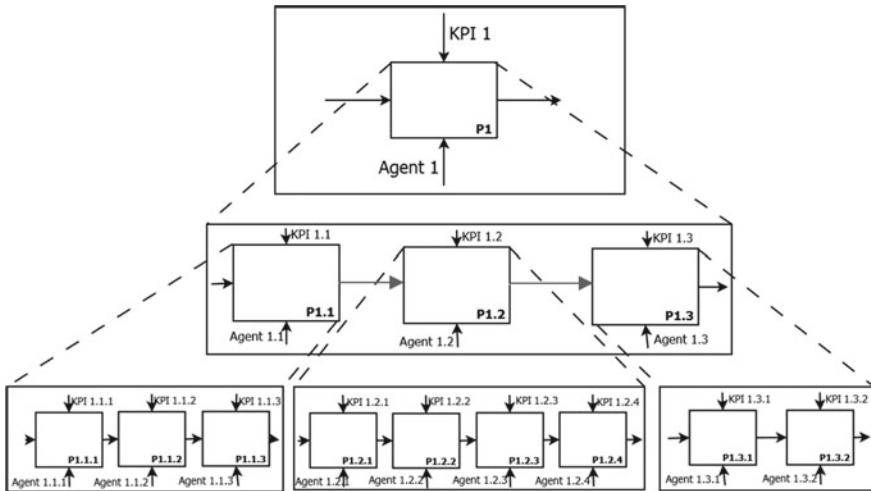


Fig. 1 Hierarchical scheme of control agents distribution

### 3 Algorithms of Agents' Interaction

#### 3.1 Algorithms of Interaction Between Agents in Adjacent Levels

Every agent stores in its knowledge base information about required values of its own KPI and about their limit values, about required values of KPI of its subordinate agents and their limit values, and about analytic relationship between these KPIs that shows how change in KPI of every subordinate agent impacts KPI of the higher-level agent and vice versa.

In this regard let us consider situation, when agent of the higher level sends request about necessity of KPI change to its subordinate agent. This can happen, for instance, when from top (from the strategic planning agent) down (to the agent that controls enterprise section) arrives request to increase speed of production (change KPI of this agent that depends on work of subordinate agents that control enterprise section equipment). After receiving this request agent should determine which changes should make every subordinate to it agent so that the whole system will provide the requested new value of agent's KPI. For this purpose, agent uses information from its knowledge base and neural network model that describes interrelation between its KPI and KPIs of agents that are subordinate to it. Then using information about limit values of KPI of its subordinate agents agent determines whether they are able to provide the new requested values of indicators. If according to its information they are unable to provide the requested values of KPI, then the agent sends corresponding response to the higher level of hierarchy.

If the change of the KPIs is possible, then the agent sends new KPI values to its subordinate agents and requests from them to change KPI. Every subordinate agent should in its turn also to check whether it with all its subordinate structure can make the requested change of KPI. If the change is not possible, then the agent sends refuse to the higher-level agent.

These steps are repeated at every level of production hierarchy. But while higher-level agents check the potential possibilities to change KPI, the lowest level agents, that are connected to PLC, check the new KPI values only against technical information received from PLC, and its own neural network models that show interrelation between these technical information and KPIs that are coming from higher-level KPI.

The scheme of interaction between agents at adjacent levels when the change is initiated at the top is shown in Fig. 2.

At the same time, it should be remembered that critical notifications that can influence the production at whole and initiate change of the production task can also be initiated at lower levels and go upwards. To this class of situations belongs situation of sending notification to higher-level agent about impossibility to keep further the current values of KPI, or the opposite notification about possibility to increase KPI values.

However, other agents of the same level in hierarchy should be also notified about these or other changes, because agents within one level are closely interconnected: indicators of such agents notably influence their joint work.

### ***3.2 Algorithms of Interaction Between Agents of the Same Level***

Let us consider situation when as a result of raw material degradation the equipment unit requires more raw material in order to keep the same KPI value. In this case agent of this equipment, unit sends request to the agent of the previous equipment unit and asks for its KPI change (e.g., delivery of more raw material).

Agent that receives such request checks in its turn possibility of its KPI change (and for this purpose sends request to its subordinate agents) and if the KPI change is possible, then it sends request about KPI change upwards to higher level of production hierarchy. If the higher-level agent permits KPI change, then it calculates new KPI values, stores them in its knowledge base, and sends confirmation to the subordinate agent. The subordinate agent changes its KPI and sends confirmation to the agent of the next equipment unit in the production chain.

But if the agent of the previous equipment unit that received request to change KPI is unable to perform this change of KPI, or if it received rejection from the higher-level agent, then it responses with rejection to the next agent that has asked for KPI change, and this agent sends notification upward to higher-level agent about impossibility to pertain current KPI value and about necessity to recalculate KPI

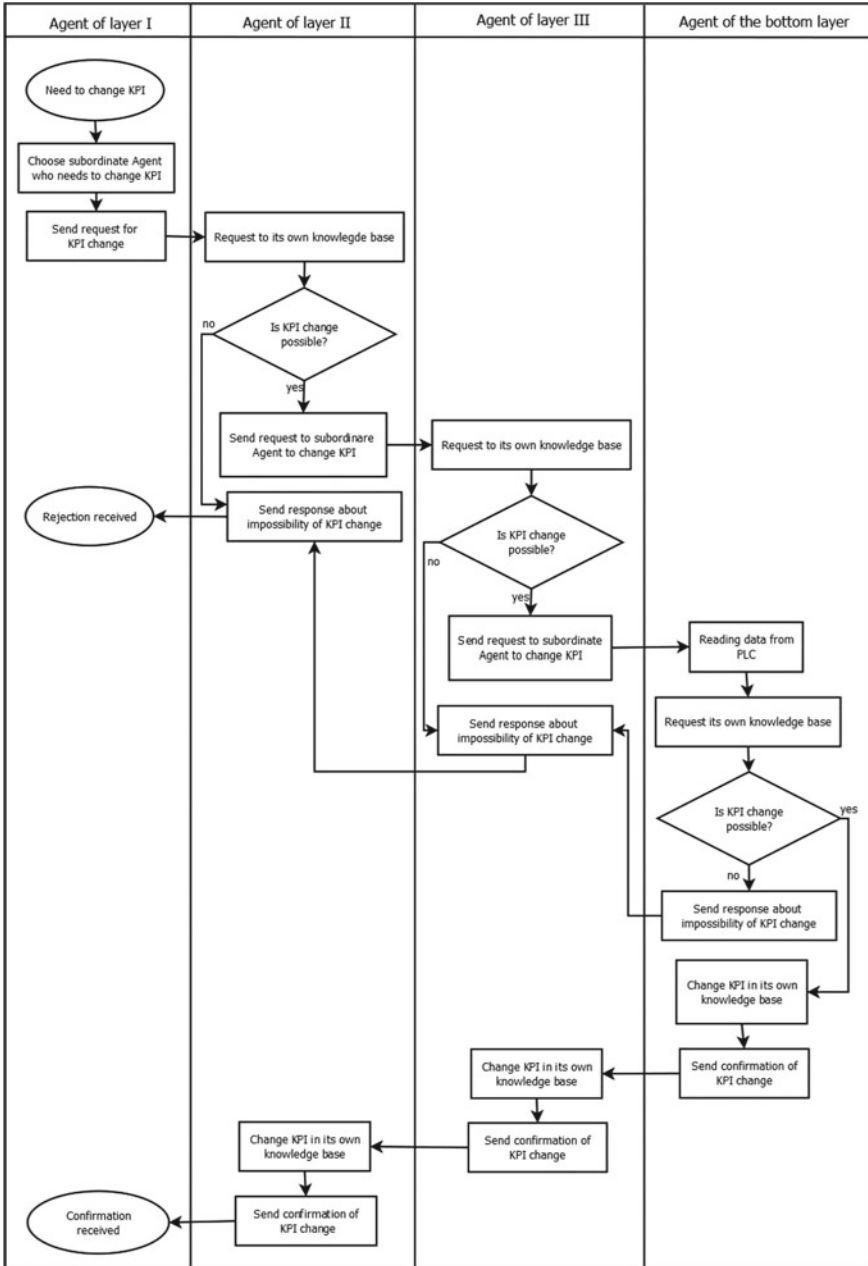


Fig. 2 Algorithm of adjacent level agents' interaction



values of other equipment units in this production chain. The scheme of interaction between agents is shown in Fig. 3.

Now let us consider implementation of these algorithms of data exchange between agents.

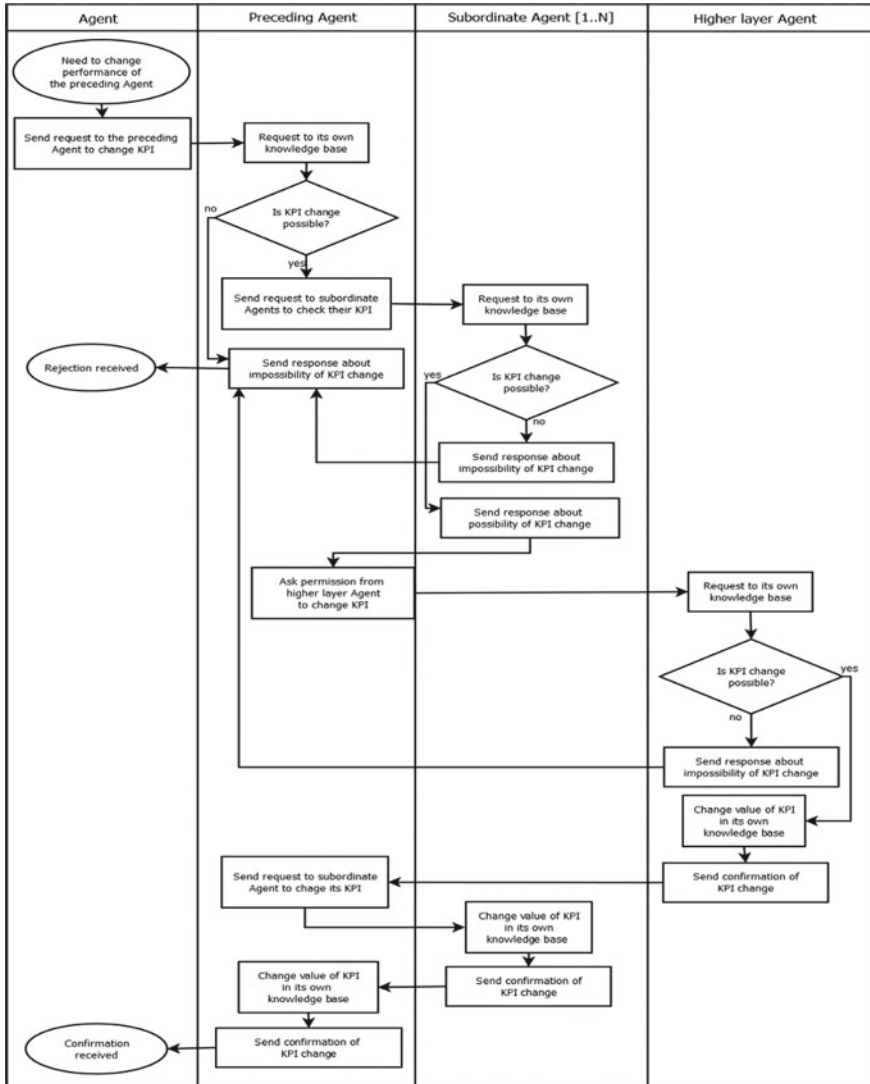


Fig. 3 Algorithm of interaction between agents of the same level

## 4 Implementation of Data Exchange

### 4.1 Implementation of Data Exchange Between Agents

The mechanism of data exchange between agents of the same or different levels was implemented using TCP/IP protocol. The data exchange in this case follows the client-server scheme. The server opens connection on some port, continuously reads requests that come to this port, and sends to clients responses. The client, knowing IP-address of the server and the port to which to connect, connects to the server and starts sending requests and receiving responses. The data exchange takes the form of sending and receiving of different sets of bytes.

Every agent knows IP-addresses of its subordinate agents, address of the higher-level agent, and the addresses of the adjacent agents in the production chain. Every agent has the server part that responsible for receiving requests from other agents, as well as the client part that enables sending of such requests.

The scheme of the client-server interaction with remote computer and the general design of the agent and its interaction with PLC and PC are shown in Fig. 4.

Now let us look at the developed system of data exchange where agent's requests and responses take the form of a certain sequences of bytes.

The first byte describes the goal of the request:

01—request to lower level agent to change KPI with the check of the possibility of the change;

02—response (confirmation/rejection) to the request to change KPI;

03—request to the adjacent agent to change KPI;

04—request to the higher-level agent for permission to change KPI;

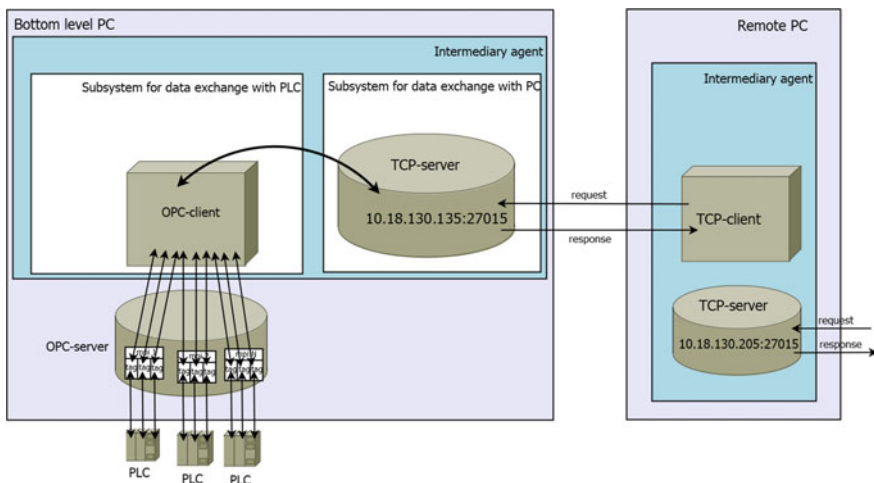


Fig. 4 Structure scheme of the lowest level agent

- 05—critical notification about maximal possible maintained value of KPI;
- 06—command to change KPI without checks and confirmations.

Then the next byte describes the object on which the action should be performed (in the case of request) or the object on which the recipient will be notified (in the case of notification).

This object is described by an identifier that is unique within the system and can use different number of bytes depending on the amount of objects in the system. In our case, the identifier represents single-byte integer, for example:

- 78—temperature
- 163—expenditure
- 250—the speed of the conveyor belt.

Then follows one or several bytes that contains information about object which identifier was specified earlier. In case of request to change KPI these bytes contain the new value of the indicator, and in case of response—information about confirmation or rejection.

Let us give an example of data packets exchange between agents in the mentioned above case, when the equipment unit needs more raw material due to the loss in its quality and the agent of the equipment unit sends **request to the agent of the previous unit** and asks for its KPI change.

At the first step, the agent sends request to the agent of the previous equipment unit to change its KPI:

03	250	70
----	-----	----

The agent of the previous unit sends requests to its subordinate agents to change their KPI:

01	85	17	...	01	163	23
----	----	----	-----	----	-----	----

The subordinate agents send upwards responses

02	85	1	...	02	163	1
----	----	---	-----	----	-----	---

After confirmation from its subordinate agents, the agent of the previous unit sends request to change KPI to the higher-level agent to whom it itself subordinates.

04	250	70
----	-----	----

After confirmation from the higher-level agent, the agent sends commands to its subordinate agents to change their KPI

06	85	17
----	----	----

...

06	163	23
----	-----	----

and notifies the agent of the next equipment unit that the KPI value was changed

02	250	1
----	-----	---

#### ***4.2 Implementation of Data Exchange Between Agent and PLC***

To date various research teams have developed big number of different frameworks, software libraries, and other tools for development of multi-agent systems. The majority of such libraries were developed using Java language (e.g., AnyLogic, Cougaar, JADE), that enables development that is more flexible. But these tools are more suitable for simulation of different interacting systems, and not for a use in real projects. When there are high requirements to performance and memory usage, then the C++ language is more suitable [21].

Therefore, a special software agent responsible for data exchange in the same level of production hierarchy as well as between different levels according to the suggested algorithms was developed using C++ language. Such an agent is located in every node in production hierarchy and is responsible for sending and receiving data to other nodes on the same or on different levels. This agent includes two subsystems: the subsystem for data exchange with other agents and a subsystem for data exchange with PLC.

In this work, subsystem for data exchange with PLC was implemented in two variations, and when the lowest level agent is created, it is possible to choose the most suitable for the specific case option for data exchange with PLC. The chosen subsystem defines how the data exchange with PLC will be carried out but the work of the agent at whole remains the same no matter which method of data exchange was chosen.

In one case, agent uses for the subsystem of data exchange with PLC a free library libNoDave that was developed for data exchange with PLC of Siemens Company [22].

In another case, subsystem for data exchange with PLC uses OPC (OLE for Process Control) technology that implies existence of OPC-server and OPC-client. Use of this technology is unified because in this case there is no need to take in account implementation details of specific type of PLC, because in this scheme OPC-server

is responsible for connection to PLC, and it hides details of PLC implementation. In our case, application developed in environment SIMATIC WinCC Flexible was used as OPC-Server to connect to Siemens Simatic controllers.

The structure scheme of the lowest level agent that uses OPC technology for data exchange with PLC is shown in Fig. 4.

The lowest level agent implemented this way is able to exchange technological information with PLC as well as send requests and receive responses about KPI change from other agents.

In our work, higher-level agents run on single board computers Raspberry Pi 3b, and the lowest level agents run on PC that connected directly to PLC.

For testing of suggested algorithms, the agents of the lowest level were implemented as software modules that run on PC with the following specification: Intel Pentium Dual CPU E2200 2.20 GHz, 3 Gb. RAM, working on OS Windows 7, and having Wi-Fi module to data exchange with other agents. The higher-level agents are implemented on single board computers Raspberry Pi 3b that also contain Wi-Fi modules for data exchange. Knowledge base of each agent that stores the required information about KPI and their limit values is implemented as files in N3/TURTLE format (knowledge representation language that is part of the RDF environment). The information about interconnections between different KPIs is represented as neural network models that are stored as matrix values.

## 5 Conclusion

In this work, the hierarchical system of control agents distribution is described, based of which multi-agent algorithms of key performance indicators reconciliation are suggested and scenarios of agents' interaction are described where the network of agents decides autonomously about further actions; the system of commands that agents working by described algorithms send to each other is defined; hardware and software tools for implementation of data exchange between agents are proposed and tested.

The developed algorithms and their implementation aimed to coordinate work of software agents that take part in the process of enterprise control in all levels from PLC up to ERP. Not all possible situations that can arise during agents' interactions are described, but the given information will be enough for adaptation of the proposed methods for arbitrary production control system.

The works assume that every agent has information about KPIs limit values of its subordinate agents, but the problem of calculation and transfer of this information still to be decided and for that purpose the models of dependency between different KPIs need to be calculated.

The next step is to implement prototype of the whole system on microprocessor boards and in this case the algorithms probably will require modification.

## References

1. Bassi, L.: Industry 4.0: hope, hype or revolution? In: 2017 IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI). pp. 1–6., Modena (2017). <https://doi.org/10.1109/RTSI.2017.8065927>
2. Kannengiesser, U., Müller, H.: Multi-level, viewpoint-oriented engineering of cyber-physical production systems: an approach based on Industry 4.0, system architecture and semantic web standards. In: 44th Euromicro Conference on Software Engineering and Advanced Applications (SEAA). pp. 331–334. Prague (2018). <https://doi.org/10.1109/SEAA.2018.00061>
3. Cheng, G.J., Liu, L.T., Qiang, X.J., Liu, Y.: Industry 4.0 development and application of intelligent manufacturing. In: 2016 International Conference on Information System and Artificial Intelligence (ISAI). pp. 407–410. Hong Kong (2016). <https://doi.org/10.1109/ISAI.2016.0092>
4. Mumtaz, S., Alsahaily, A., Pang, Z., Rayes, A., Tsang, K.F., Rodriguez, J.: Massive Internet of Things for industrial applications: addressing wireless IIoT connectivity challenges and ecosystem fragmentation. *IEEE Ind. Electron. Mag.* **11**, 28–33 (2017). <https://doi.org/10.1109/MIE.2016.2618724>
5. Vachalek, J., Bartalsky, L., Rovny, O., Sismisova, D., Morhac, M., Loksik, M.: The digital twin of an industrial production line within the industry 4.0 concept. In: 21st International Conference on Process Control (PC), pp. 258–262. Strbske Pleso (2017). <https://doi.org/10.1109/PC.2017.7976223>
6. Yusen, X., Bondaletova, N.F., Kovalev, V.I., Komrakov, A.V.: Digital twin concept in managing industrial capital construction projects life cycle. In: 11th International Conference “Management of Large-Scale System Development” (MLSD). pp. 1–3. Moscow (2018). <https://doi.org/10.1109/MLSD.2018.8551867>
7. Tao, F., Zhang, M.: Digital twin shop-floor: a new shop-floor paradigm towards smart manufacturing. *IEEE Access* **5**, 20418–20427 (2017). <https://doi.org/10.1109/ACCESS.2017.2756069>
8. Kostenko, D., Kudryashov, N., Maystrishin, M., Onufriev, V., Potekhin, V., Vasiliev, A.: Digital twin applications: diagnostics, optimisation and prediction. In: Katalinic, B. (ed.) 29th International DAAAM Symposium. pp. 0574–0581. Vienna (2018). <https://doi.org/10.2507/29th.daaam.proceedings.083>
9. Eckhardt, A., Muller, S., Leurs, L.: An evaluation of the applicability of OPC UA publish subscribe on factory automation use cases. In: IEEE International Conference on Emerging Technologies and Factory Automation (ETFA). pp. 1071–1074. Turin (2018). <https://doi.org/10.1109/ETFA.2018.8502445>
10. Gligorea, R.: Key Performance Indicators in the Oil & Gas Industry—BP Company, <https://www.performancemagazine.org/key-performance-indicators-oil-bp/>, last accessed 2019/12/11
11. Xiong, G., Qin, T., Wang, F., Hu, L., Shi, Q.: Design and improvement of KPI system for materials management in power group enterprise. In: Proceedings of 2010 IEEE International Conference on Service Operations and Logistics, and Informatics, SOLI 2010, pp. 171–176 (2010). <https://doi.org/10.1109/SOLI.2010.5551585>
12. Leitão, P., Barbosa, J., Vrba, P., Skobelev, P., Tsarev, A., Kazanskaia, D.: Multi-agent system approach for the strategic planning in ramp-up production of small lots. In: Proceedings—2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC). pp. 4743–4748. Manchester (2013). <https://doi.org/10.1109/SMC.2013.807>
13. Budaev, D., Amelin, K., Voschuk, G., Skobelev, P., Amelina, N.: Real-time task scheduling for multi-agent control system of UAV’s group based on network-centric technology. In: International Conference on Control, Decision and Information Technologies, (CoDIT). pp. 378–381. St. Julian’s (2016). <https://doi.org/10.1109/CoDIT.2016.7593592>
14. Aksyonov, K.A., Bykov, E.A., Smoliy, E.F., Aksyonova, O.P., Kai, W.: Efficient decision support for control and management processes of industrial enterprises with BPsim.DSS. In: Chinese Control and Decision Conference (CCDC), pp. 261–265. Mianyang (2011). <https://doi.org/10.1109/CCDC.2011.5968183>

15. Nadouri, S., Ouhammou, Y., Sahnoun, Z., Hadjali, A.: Towards a multi-agent approach for distributed decision support systems. In: 27th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE), pp. 78–84. Institute of Electrical and Electronics Engineers Inc., Paris (2018). <https://doi.org/10.1109/WETICE.2018.00021>
16. Zhang, L., Zhang, Y.: Research on hierarchical distributed coordination control in process industry based on multi-agent system. In: 2010 International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), pp. 96–100. Changsha City (2010). <https://doi.org/10.1109/ICMTMA.2010.683>
17. Zheng, G., Li, N.: Multi-agent based control system for multi-microgrids. In: International Conference on Computational Intelligence and Software Engineering (CiSE), Wuhan (2010). <https://doi.org/10.1109/CISE.2010.5676818>
18. Shao, Y., Wu, Y., Chen, Y.: Design and research of multi-agent control system for central cooling system. In: 3rd International Conference on Cloud Computing and Intelligence Systems (CCIS), pp. 218–221. Shenzhen (2014). <https://doi.org/10.1109/CCIS.2014.7175732>
19. Wei, C., Jing, J., Mei, C., Shaoting, G.: Design of workshop production management control system based on multi-agent. In: 11th Conference on Industrial Electronics and Applications (ICIEA). Hefei (2016). <https://doi.org/10.1109/ICIEA.2016.7603622>
20. Lang, A., Stanley, K.O.: NeuroEvolutionary meta-optimization. In: International Joint Conference on Neural Networks (2013). <https://doi.org/10.1109/IJCNN.2013.6707097>
21. Theiss, S., Vasyutynskyy, V., Kabitzsch, K.: AMES - A Resource-Efficient Platform for Industrial Agents. In: IEEE International Workshop on Factory Communication Systems. pp. 405–413., Dresden (2008)
22. LibNoDave—A free communication library for Simatic S7 PLCs, <http://libnodave.sourceforge.net>. Last accessed 2019/12/11

# Semi-supervised Learning for Medical Image Segmentation



Mikhail Kots , Mikhail Pozigun , Andrei Konstantinov ,  
and Viacheslav Chukanov 

**Abstract** Semi-supervised learning is a combination of conventional supervised methods with weakly supervised learning. A recent development in neural networks allows to achieve high-quality results but the training requires a large amount of annotated examples. This hinders the applicability of deep learning to some problems, especially medical imaging. In this paper, we present a semi-supervised learning approach based on convolutional neural networks (CNN) for medical image segmentation. A network is trained on a combination of fully labeled samples that have segmentation masks available and weakly labeled samples that only have class labels. We performed experiments that compare the results of the semi-supervised model with the baseline supervised method. Experiment results show the superiority of suggested methods on a low amount of fully annotated samples for lung nodules CT images.

**Keywords** Semi-supervised learning · Neural networks · Medical imaging · Machine learning

## 1 Introduction

Recently, deep learning-based methods have become the most popular solution for various pattern recognition and computer vision tasks. These algorithms allow achieving high-quality results via supervised learning, but they usually require a lot of annotated samples. This creates a significant issue when applying deep learning to problems where the production of annotated training samples is expensive or time-consuming. One of the most notable examples is medical image segmentation, where segmentation masks for training have to be prepared manually by experts.

A number of publications suggested different solutions to this problem. A common strategy for image segmentation is to train deep models with more simple labels such as class-level labels or unlabeled images. These methods usually referred to as weak

---

M. Kots (✉) · M. Pozigun · A. Konstantinov · V. Chukanov  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [mkots96@gmail.com](mailto:mkots96@gmail.com)



supervision. However, several researches stated that some algorithms could largely benefit from using some fully annotated samples.

Adaptation of conventional supervised methods for semi-supervised training proved to be a challenging problem, especially in a medical imaging scope due to a high diversity of object's features in them. Several papers proposed various approaches to increase segmentation quality by using unlabeled samples [1]. These methods are commonly called semi-supervised learning [2].

One of the earliest groups of methods is self-training. They suggest an iterative procedure to incorporate unlabeled samples. First, a model is pre-trained with labeled data. Then it is used to classify unlabeled data. After that, the most reliable unlabeled samples are added to the training set with predicted labels and the classifier is retrained. This process is repeated until no more samples can be added. The suggested approach was used in numerous applications, for example, Gu et al. utilized it for label propagation in retinal images [3]. However, these methods have a few downsides since it requires to find a criterion to choose reliable samples and they are prone to strengthen classification errors from early iterations.

Another notable approach for semi-supervised learning is co-training [4] which improves self-training by eliminating the need for criterion selection. Co-training assumes that data could be described by two independent and sufficient subsets of features. Separate classifiers are used for each subset. Predictions of one classifier are used as new labeled data for another classifier during the iterative training process. Zhou et al. [5] propose to use the co-training approach for 3D CT scans segmentation by utilizing features from three planes. One model generates pseudo-labels of unlabeled data combining predictions from three views and enlarging the dataset for another model training. Alternation of training and pseudo-labels generations is repeated several times.

An alternative approach is to use weakly labeled samples instead of unlabeled. This method was used by Mlynarski et al. [6] for brain tumor segmentation. The main idea is to use subnetwork that performs image classification and train this subnetwork jointly with segmentation subnetwork sharing weights between most of the layers. It uses several fully connected layers for classification subnetwork.

There is a group of methods under the name of multi-instance learning, which is based solely on weak supervision. Classic problem statement implies that the training set is divided into groups called bags with labels and the algorithm learns to predict instance label by training on bags labels. There are numerous publications on multi-instance learning problems solved by classical machine learning algorithms [7–9]. In the case of an image segmentation task, images are treated as bags and pixels—as instances. These methods can also benefit from using fully labeled data to improve the training process [10].

One of the earliest deep learning approaches for the multi-instance learning problem proposed in [11]. This paper focused on solving the task of image classification and object localization using a pre-trained CNN model. The global max-pooling is used to determine the best-scoring object position within the image.

Several alternative global pooling operations have been suggested [3]. The most notable example is the weighted average of samples proposed by [12]. Weights are

determined by a neural network. Suggested weighted pooling could provide more informative bag representation for the bag-level classifier. In [13] joint learning of separate classifier and localizer parts is presented. The proposed approach does not use any pre-trained model. Combined loss function which is a weighted sum of losses of classifier and localizer is used in simultaneous training of these parts. This method helps to set proper initial values for localizer and then find good local optimum during training.

Jia et al. [14] present a fully convolutional neural network (FCNN) model for image segmentation. In this algorithm classification labels are derived as a weighted sum of the pixel-level predicted probability maps from FCNN layers. Generalized mean has been used for the calculation of image-level predictions to improve gradient flow. Besides image labels, a rough estimation of the segmented area has been exploited as an additional input to enhance learning capability. Wang et al. [15] propose a series of neural networks for bag classification without calculating instance probability. Residual connections, as well as additional fully connected layers for predicting instance scores followed by one of three different types of pooling layers, have been proposed for boosting classification performance.

Let us denote the training set as a union of labeled images and weakly labeled images  $X = X_S \cup X_W$ . Weakly labeled images have only class-level labels provided  $X_W = \{(x_i, c_i) | x_i \in R^n, c_i \in \{0..K\}\}$ , while labeled images provided with pixel-wise annotation  $X_S = \{(x_i, y_i) | x_i \in R^n, y_i \in \{0..K\}^n\}$ , where  $K$  is a number of classes. In this work, we only going to consider binary segmentation, so  $c_i \in \{0, 1\}$  and  $y_i \in \{0, 1\}^n$ . We assume that the number of labeled samples is very little, so any supervised learning algorithm will not be effective, and weakly labeled samples are a larger part of the dataset. The goal is to develop a model that trains on both types of samples and demonstrates superior segmentation quality comparing to the baseline model which is trained only on labeled data.

This paper presents a semi-supervised learning method based on a convolutional neural network for medical image segmentation. We propose a network architecture based on the multi-instance learning approach as well as end-to-end training procedure which allows training jointly on any combination of fully labeled and weakly labeled samples. We also provide experimental results that prove that the suggested method is applicable to lung nodule segmentation.

## 2 The Suggested Approach

Our baseline for image segmentation is a modification of U-Net. This network is based on FCNN and features symmetric upsampling and downsampling parts with skip connections, which concatenate convolution outputs of the same shape. This allows it to train on fewer training images and output high-quality segmentation masks. U-Net has a little number of parameters compared to other models, capable of training on a low amount of data [16] and has numerous successful applications in medical imaging [17]. The only modification we introduced to the baseline model

is batch normalization [18] after each convolution to stabilize and speed up network convergence.

In order to bridge weakly labeled data with conventional neural networks, we used the global pooling layer [15]. It aggregates features across low-dimensional feature maps into a single number, which represents the presence or absence of a certain feature in a certain part of the image. The global pooling layer is followed by  $1 \times 1$  convolution to perform weighted averaging over all features.

Several publications proposed different versions of global pooling operation [11, 12, 15]. In this method, global average pooling was used as it was proved superior for the segmentation task [19].

Most publications suggest applying global pooling to the last layer of the baseline model [6, 14, 15]. However, using global pooling over low-dimensional representation might be more beneficial. We suggest using it after the last convolution of the network’s downsampling part. The proposed model has two outputs: predicted segmentation mask and class label. Depending on input samples being fully labeled or weakly labeled, the network uses either of these outputs to compute loss function value. The network architecture is presented in Fig. 1.

The network is trained on both labeled and weakly labeled images in turns. On each iteration, the training algorithm draws a batch of fully labeled and weakly labeled samples and trains on both subsequently. Cross entropy is used as a loss function on both outputs.

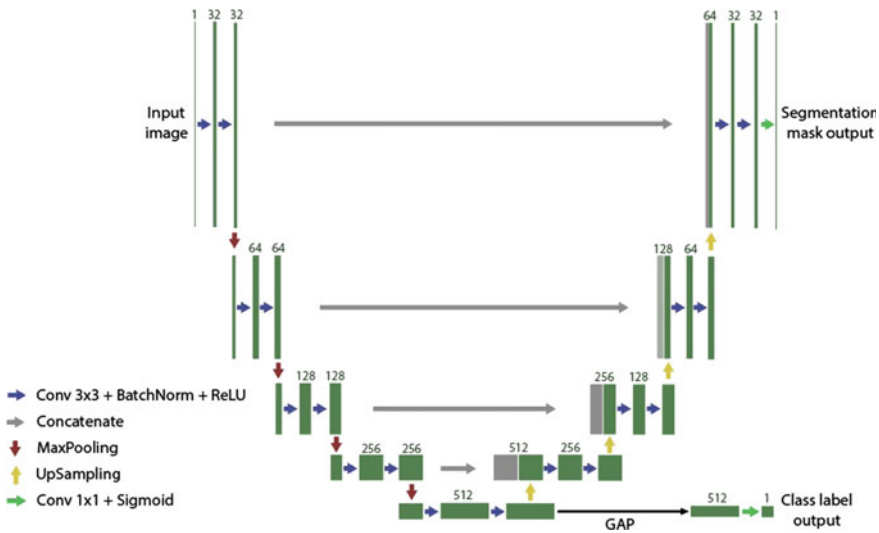


Fig. 1 The network architecture

### 3 Experiments

Medical images for these experiments are provided by medical segmentation decathlon challenge 2018 which consists of several CT and MRI image sets with different labeled regions of interest (ROI). All 3D volumes were split into 2D slices across the  $z$ -axis resulting in  $512 \times 512$  images. The following three types of images were picked for the experiments:

1. CT liver images. Abdomen images with the labeled liver. For testing purposes, 10,000 slices were chosen for cross-validation with an equal balance between slices with or without ROI.
2. CT spleen images. Abdomen images with the labeled spleen. Testing images were chosen in the same way as the liver.
3. CT lung nodules images. Chest images with labeled lung nodules. Dataset was balanced between classes to account for a low number of slices with ROI which resulted in  $\sim 5000$  samples.

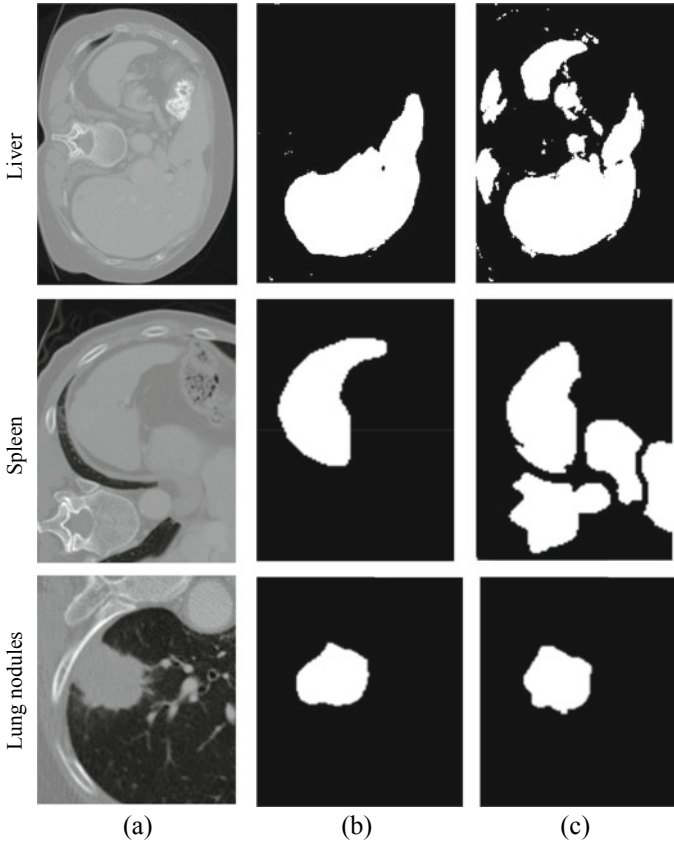
All images in the dataset have full annotation available but for these experiments, we substituted segmentation masks with class labels for part of the samples. For each sample class label  $c_i = 1$  if at least one pixel from segmentation mask belongs to ROI,  $c_i = 0$  otherwise.

The first experiment was aimed to determine algorithm performance on different datasets. Results were obtained via 5–2 cross-validation and exactly 40% of training data was used as fully labeled samples, the rest was weakly labeled samples. To estimate the segmentation quality, we used average DICE across all validation samples. For each ROI set, we measured the segmentation quality of the baseline model trained in a supervised fashion on a given amount of fully labeled images and compared them to the semi-supervised model. Both models were trained for 100 epochs using Adam optimizer with  $\beta_1 = 0.9$ ,  $\beta_2 = 0.999$ , learning rate =  $1 \times 10^{-5}$  and batch size = 4. The results are present in Table 1. Examples of segmentation are presented in Fig. 2.

The second experiment was aimed to discover an influence of ratio between labeled and weakly labeled samples on segmentation quality. This experiment was performed on lung nodule ROI and percent of fully labeled training data ranged from 10 to 70%. The rest of the parameters were set up exactly like the previous experiment. The results are present in Table 2.

**Table 1** Segmentation quality on different ROI

ROI	Baseline	Semi-supervised
Liver	<b>0.74</b>	0.52
Spleen	<b>0.69</b>	0.56
Lung nodules	0.62	<b>0.65</b>



**Fig. 2** Segmentation quality on different ROI. **a** Source image, **b** baseline results, **c** semi-supervised results

**Table 2** Segmentation quality for different labeled samples amount

% of labeled samples (# slices)	Baseline	Semi-supervised
10% (~400)	0.09	<b>0.26</b>
40% (~1600)	0.62	<b>0.65</b>
70% (~2800)	0.68	<b>0.69</b>

## 4 Discussion

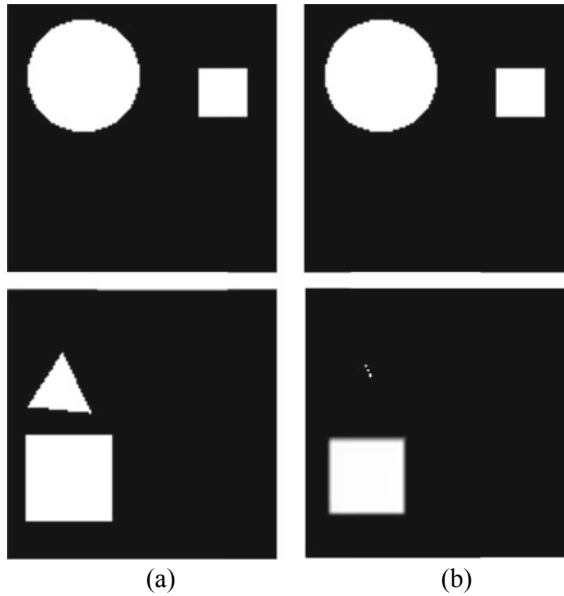
In this paper, we have developed the CNN-based semi-supervised learning approach. The results of the first experiments show that the proposed method is not equally applicable to different types of ROI. During this experiment method achieved quality increase compared to baseline only for lung nodules. Despite several publications

reporting good quality segmentation on general-purpose image dataset [11, 12, 15, 20], results do not seem to always transition well to medical images. Authors believe that this outcome has to do with the nature of medical data. In the presented use case weak annotations do not provide any information about ROI. Instead, the network learns distinctive features from dissimilarities of objects and the background of given samples. Medical images tend to have a similar outline thus ROI constantly presented in the slice with other objects which the semi-supervised approach recognizes as a part of ROI. To demonstrate this problem, we performed an experiment on synthetic data with simple geometric figures. The dataset consists of binary images with a random combination of circle, square, and triangle. All figures have a random position and size. The task is to segment circle and it is always presented in the image together with the square. Dataset consists of 1000 samples. For this experiment only weakly labeled samples were used. The sample prediction of the model is presented in Fig. 2. The segmentation mask always contains labeled square even considering that some of the negative samples in the dataset contained square. Moreover, during validation squares on images without circle were also labeled very confidently. This indicates that the network learned square features as a part of the object. All the above explains poor performance on some medical datasets, however, in case of abnormalities as ROI such as lung nodules suggested method performs considerably better.

The second experiment proves that the suggested method is always beneficial for different ratios between labeled and weakly labeled samples although it becomes less effective with more training samples as the baseline model reaches the limit of its generalization ability. We also noticed that this approach does not perform to its full potential in case of a very little amount of labeled samples where the baseline method vastly overfits. In this case additional weakly labeled samples improve the quality, but the resulting quality is still considerably low.

## 5 Conclusion

In this paper, we presented the semi-supervised learning approach based on CNN and joint end-to-end training procedure. Experimental results show that the suggested method demonstrates a significant quality improvement in comparison with the baseline model on lung nodules CT images. The proposed method has been proven to train on a low amount of fully labeled data. Solution for segmentation task for arbitrary medical dataset seems to not be feasible for this method and quality improvement in case of a low amount of fully labeled samples most likely will be the topic of further research (Fig. 3).



**Fig. 3** Results of a synthetic experiment. **a** input image, **b** output

**Acknowledgements** This work was supported by Philips Research.

## References

1. Bagherzadeh, J., Hasan, A.: A review of various semi-supervised learning models with a deep learning and memory approach. *Iran J. Comput. Sci.* **2**(2), 65–80 (2019)
2. Zhu, X., Goldberg, A.: Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.* **3**(1), 1–130 (2009)
3. Gu, L., Zheng, Y., Bise, R., Sato, I., Imanishi, N., Aiso, S.: Semi-supervised learning for biomedical image segmentation via forest oriented super pixels (voxels). In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 702–710. Springer, Cham (2017).
4. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pp. 92–100. ACM (1998).
5. Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E.K., Yuille, A.L.: Semi-supervised multi-organ segmentation via deep multi-planar co-training. *arXiv preprint [arXiv:1804.02586](https://arxiv.org/abs/1804.02586)* (2018).
6. Mlynarski, P., Delingette, H., Criminisi, A., Ayache, N.: Deep learning with mixed supervision for brain tumor segmentation. *J. Med. Imag.* **6**(3), 034002 (2019)
7. Carbonneau, M.A., Cheplygina, V., Granger, E., Gagnon, G.: Multiple instance learning: a survey of problem characteristics and applications. *Pattern Recogn.* **77**, 329–353 (2018)
8. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: *Advances in Neural Information Processing Systems*, 577–584 (2003).

9. Zhang, Q., Goldman, S.A.: EM-DD: An improved multiple-instance learning technique. In: *Advances in Neural Information Processing Systems*, pp. 1073–1080 (2002).
10. Chen, H., Wang, Y., Wang, G., Qiao, Y.: Lstd: a low-shot transfer detector for object detection. In: *32nd AAAI Conference on Artificial Intelligence* (2018).
11. Oquab, M., Bottou, L., Laptev, I., Sivic, J.: Is object localization for free?—Weakly-supervised learning with convolutional neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 685–694 (2015).
12. Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. *arXiv preprint [arXiv:1802.04712](https://arxiv.org/abs/1802.04712)* (2018).
13. Hwang, S., Kim, H.E.: Self-transfer learning for weakly supervised lesion localization. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 239–246. Springer, Cham (2016).
14. Jia, Z., Huang, X., Chang, E., Xu, Y.: Constrained deep weak supervision for histopathology image segmentation. *IEEE Trans. Med. Imaging* **36**(11), 2376–2388 (2017)
15. Wang, X., Yan, Y., Tang, P., Bai, X., Liu, W.: Revisiting multiple instance neural networks. *Pattern Recogn.* **74**, 15–24 (2018)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, Cham (2015).
17. Çiçek, O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 424–432. Springer, Cham (2016)
18. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *32nd International Conference on Machine Learning*, pp. 448–456 (2015).
19. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Med. Image Anal.* **54**, 280–296 (2019)
20. Kraus, O.Z., Ba, J.L., Frey, B.J.: Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* **32**(12), i52–i59 (2016)



# Developing a New Generation of Reconfigurable Heterogeneous Distributed High Performance Computing System



Alexander Antonov , Vladimir Zaborovskij , and Ivan Kisilev 

**Abstract** Restrictions of current architectures of Heterogeneous High-Performance Computing (HPC) systems lead to Reconfigurable Heterogeneous HPC (RH HPC) which are able to adapt to a particular solved task on the hardware level. Highly disruptive technologies like Artificial Intelligence (AI), Internet of Things (IoT), and Machine Learning (ML) are expected to lead not only fundamental shift in classical multi-cores and multi-treads-based approaches to high-performance computing, but also open new direction in designing systems with dynamically reconfigurable runtime environment. The state-of-the-art integrated circuits allow realizing the general architecture of RH HPC at different levels of the distributed HPC: from the level of the supercomputer, to the level of user computers and the remote, built-in units intended for data acquisition, management, and control. The article describes proposed architecture for new generation of Reconfigurable Heterogeneous Distributed HPC (RHD HPC) system, including architectures of each sub-parts and highlights already developed components for such RHD HPC.

**Keywords** Heterogeneous High-Performance Computing · Hardware Reconfigurable · Field-Programmable Gate Array · Machine Learning · DC-Cloud · EDGE

## 1 Introduction

Current High-Performance Computing (HPC) systems such as Supercomputers [1], Data centers, Cloud services are becoming larger and larger; consume more and more power [2]. This trend leads to increasing of required space, power consumption, and cost of deploying and maintenance service for HPC systems.

Due to rapidly increasing requirements: to performance, measured in Floating-point Operations Per Second (FLOPS); to Power Efficiency (FLOPS/W); to Calculation Efficiency (Real FLOPS/Peak FLOPS); to Size Efficiency (Real FLOPS/square),

---

A. Antonov (✉) · V. Zaborovskij · I. Kisilev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [antonov@eda-lab.ftk.spbstu.ru](mailto:antonov@eda-lab.ftk.spbstu.ru)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021  
N. Voinov et al. (eds.), *Proceedings of International Scientific Conference on Telecommunications, Computing and Control*, Smart Innovation, Systems and Technologies 220, [https://doi.org/10.1007/978-981-33-6632-9\\_22](https://doi.org/10.1007/978-981-33-6632-9_22)

255

modern HPC systems are evolving toward heterogeneous computing [3]. Current heterogeneous HPCs mostly use General Purpose Graphics Processing Units (GPGPUs) [4] and Application-Specific Integrated Circuit (ASIC) as accelerators to perform efficient solving of Artificial Intelligence (AI), Machine Learning (ML), Internet of Things (IoT), and Big Data analysis tasks [5]. The next step in evolving HPC systems is reconfigurable computing [6].

Reconfigurable computing technology is based on Field-Programmable Gate Array (FPGA) [7, 8]. FPGA is an integrated circuit (IC) that can change its internal structure according to the task being solved. FPGA consists of programmable logic cells that can perform any logic/memory functions and programmable matrix (interconnection matrix) that can connect all logic cells together to implement complex functions. FPGA is programmed, or configured, by binary file, called configuration file, that configure logic cells and interconnection matrix. Configuration file sets up the logic cells and the interconnection matrix such that FPGA can implement the task being solved. Modern FPGA contains not only logic cells and interconnection matrix but also Digital Signal Processing (DSP) blocks, Random Access Memory (RAM) blocks, High Bandwidth Memory (HBM) based on embedded DDR memory blocks, hardware implemented controllers and transceivers for external: DDR memory, PCIe interface, 100G Ethernet.

State-of-the-art FPGA could be configured on-the-fly. It means that FPGA could be configured for solving new task during execution of current task. FPGA could be partially configured. It means that a part of FPGA could be configured for solving new task while the rest of FPGA continues to solve current task. Finally, FPGA could be configured and partially configured through PCIe and Ethernet for solving a particular task. It means that FPGA-based PCIe Accelerator deployed on a Host or FPGA-based remote accelerator connected to a Host by high-speed channel, for example, 100G Ethernet, could be on-the-fly dynamically configured, or re-configured, to solve a particular task with efficiency of hardware implementation.

## 2 Materials and Method

Compared to existing heterogeneous HPC systems [9–13] which consist of a Multi-Processor Unit (MPU), or clusters of MPUs, and GPGPU-based accelerators, Reconfigurable Heterogeneous HPCs (RH HPC), by using reconfigurable FPGA-based accelerators, are able to meet the requirements of particular tasks, such as: data structures, calculation algorithms, real-time requirement and allow to solve particular tasks more efficiently [14–16] in terms of Power Efficiency (FLOPS/W), Calculation Efficiency (Real FLOPS/Peak FLOPS) and Size Efficiency (Real FLOPS/square).

Current understanding of System-on-Chip (SoC) is: FPGA, often referenced as Logic Part of SoC; Multi-core processor, often referenced as Processor Part of SoC, with GPU accelerator; a lot of embedded peripheral components [16], such as PCIe, USB3.0, MAC Ethernet, SATA, DDR4, SPI/QSPI, NAND memory, SD Card which are deployed on silicon of a single IC. It means that state-of-the-art SoC

has reconfigurable heterogeneous architecture and could be treated as tiny RH HPC systems.

State-of-the-art FPGA, SoCs, and of-the-shelf devices allow to use the Reconfigurable Heterogeneous (RH) architecture for building Supercomputers, Data centers, and Cloud services (DC-Cloud RH HPC), office computers (Premises RH HPC) and remote high-performance computing systems (Edge RH HPC).

In the chapter, we review proposed architectures of Reconfigurable Heterogeneous Distributed High-Performance Computing (RHD HPC) System and architecture of developed and deployed PCIe-based reconfigurable accelerator.

### 3 Results

#### 3.1 *The Proposed Architecture of Reconfigurable Heterogeneous Distributed HPC System*

The proposed architecture of reconfigurable heterogeneous distributed HPC system (see Fig. 1) is based on our previous researches dealing with HPC architectures based on OpenCL standard [17, 18].

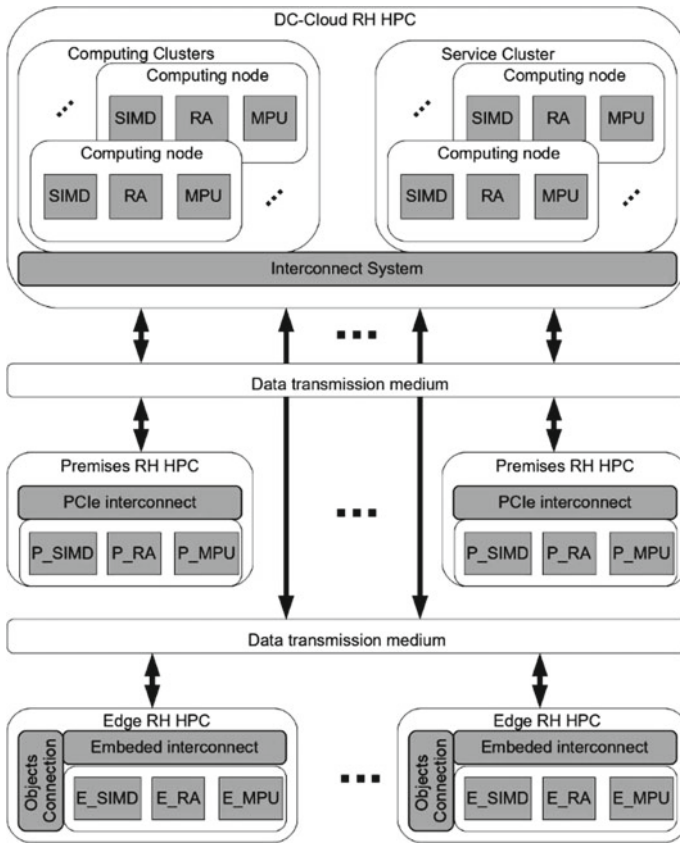
DC-Cloud RH HPC (see Fig. 1) consists of some Computing Clusters and one Service Cluster. Each of the clusters consists of some identical computing nodes with MPU, Reconfigurable FPGA-based Accelerator (RA), and Single Instruction Multiple Data (SIMD) accelerator, which was previously pointed as GPGPU. The Computing Clusters are intended for solving the particular computational tasks. The Service Cluster implements:

- performance evaluation of all Computing Clusters, remotely connected Premises RH HPC and Edge RH HPCs;
- optimal task distribution between available computational resources. Optimization criterion could be pointed by user or assigned by AI, deployed on the Service Cluster, automatically.

Premises RH HPC (see Fig. 1) consists of Premises MPU (P\_MPU), Premises Reconfigurable FPGA-based Accelerator (P\_RA), and Premises Single Instruction Multiple Data (P\_SIMD) accelerator. PCIe3.0  $\times$  16 (PCIe4.0  $\times$  16) interface provides interconnection between P\_MPU, P\_RA, and P\_SIMD. Premises RH HPC could be:

- identical with the Computing Node of DC-Cloud RH HPC. In this case, P\_MPU, P\_RA, and P\_SIMD are identical to MPU, RA, and SIMD, respectively;
- specialized for solving particular tasks.

Edge RH HPC (see Fig. 1) consists of Embedded MPU (E\_MPU), Embedded Reconfigurable FPGA-based Accelerator (E\_RA), and Embedded Single Instruction Multiple Data (E\_SIMD) accelerator. PCIe3.0  $\times$  16 (PCIe4.0  $\times$  16) interface



**Fig. 1** The proposed architecture of reconfigurable heterogeneous distributed HPC system

provides interconnection between P\_MPU, P\_RA, and P\_SIMD. The connection, pointed on Fig. 1 as Embedded Interconnection, can be realized by:

- PCIe interface if all E\_MPU, E\_RA, E\_SIMD or some of them are separate devices
- interconnection matrix if all E\_MPU, E\_RA, E\_SIMD are deployed inside FPGA.

Since Edge RH HPC is intended for interaction with sensors and actuators, their important element is Object Connection block, highlighted in Fig. 1. The Object Connection block can contain analog-to-digital converters (ADCs), digital-to-analog converters (DACs), digital inputs/outputs (DIO), and other means of interaction with the particular object.

It is assumed that data transmission medium, pointed in Fig. 1, can be an arbitrary combination of wired (with speed of 1–100 GBIT/s) and wireless (e.g., Bluetooth, Wi-Fi, LTE, 5G) connections. The choice of which is depended by the characteristics of the solving tasks and the remote objects parameters.

### 3.2 The Proposed Architecture of the Computing Node

The proposed architecture of the Computing Node (see Fig. 2) has been extracted from the performance demands of Machine Learning tasks [19].

The proposed architecture of the Computing Node contains:

- Two Central Processing Units (CPUs). Each CPU itself is a multiprocessing unit containing several, up to several dozen, computing cores, and some number of embedded controllers for high-speed connection with external dynamic memory, PCIe interface, 1–100G Ethernet connections, etc. CPUs should have a direct connection and implement non-uniform memory access (NUMA).
- Dynamic Random Access Memory (DRAM) blocks, which, at the physical level, are DDR4 memory modules. DRAM is the local memory for each processor, the width, performance, and volume of which depends on the purpose, particular solving tasks, desirable performance, and power consumption of the Computing Node.
- Some number of SIMD accelerators. Each SIMD accelerator should have independent connection with each CPU in the Computing Node. The independent connection could be as simple as PCIe3.0 (PCIe4.0) × 16\8 interface [20], or as advanced as Open Coherent Accelerator Processor Interface (OpenCAPI) or Cache Coherent Interconnect for Accelerators (CCIX) interface or Compute eXpress Link (CXL).
- Some number of RA accelerators. Each RA accelerator should have independent connection with each CPU in the Computing Node. The independent connection

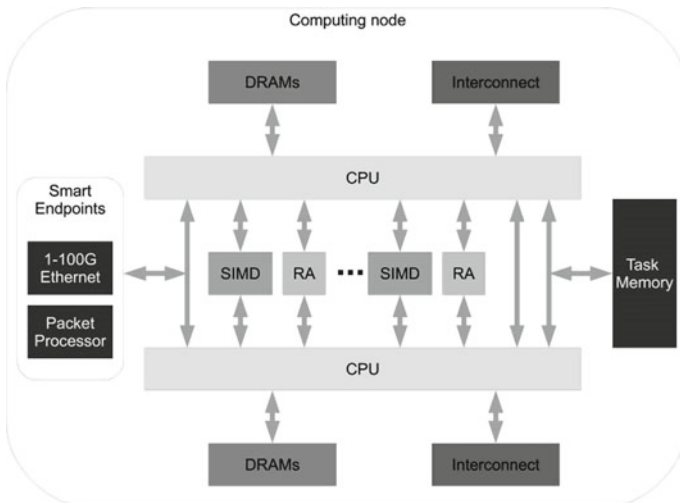


Fig. 2 The proposed architecture of the Computing Node

could be as simple as PCIe3.0 (PCIe4.0)  $\times$  16/8 interface, or as advanced as OpenCAPI [10] or CCIX interface.

- Task Memory, which is Random Access Memory (RAM) with capacity from 16GByte and performance comparable with DDR4 memory. The Task Memory should be connected with each CPU could access to Task Memory through PCIe3.0 (PCIe4.0) interfaces or by OpenCAPI/CCIX, by implementing Uniform Memory Access (UMA) at the tasks level.
- Interconnect which is Interconnect blocks, which are a local parts of the Interconnect System, see Fig. 1. The Interconnect blocks provide high-speed wired connections between Computing Nodes in the Computing and Service Clusters and between Clusters. Each Interconnect block should contain one or some 100G connectors/controllers.
- Smart Endpoints, which are intelligent units for direct, without using CPUs, channel with outside world. Each Smart Endpoint consists of 1–100G Ethernet block and Packet Processors. The 1–100G Ethernet block could contain one or more Ethernet(SFP/QSFP connectors and PHysical Layer (PHY) controller. Packet Processor is intended for solving security issues and intelligent data processing, like particular data extraction for further processing, with efficiency of hardware implementation.

Proposed architecture of the Computing Node is treating as universal architecture for building Computing Clusters, Service Clusters and as a core architecture for Premises RH HPC.

### 3.3 *The Proposed Architecture of the EDGE RH HPC*

The proposed architecture of the Edge RH HPC (see Fig. 3) is based on our experience in developing and deploying of high-performance Systems-on-Chip.

The proposed architecture of the Edge RH HPC contains:

- Multi-Core CPU, which is a main processing unit.
- E\_SIMD accelerator, which is tightly coupled with Multi-Core CPU. It could be implemented as a separate Integration Circuit (IC) or as embedded GPGPU unit inside SoC device.
- E\_RA accelerator, which could be implemented as a separate IC or as embedded unit, deployed on Logic Part of SoC device.
- DRAM blocks, which, at the physical level, are DDR4 memory modules. DRAMs are the local memory for Logic Part and Processor Part of SoC device. The width, performance, and volume are functions of the purposes, particular solving tasks, desirable performance, and power consumption of the Edge RH HPC.
- Packet Processor, which is intended for solving security issues and intelligent data processing, like particular data extraction for further processing, with efficiency of hardware implementation. The Packet Processor should be implemented as a separate IC or deployed on Logic Part of SoC device.

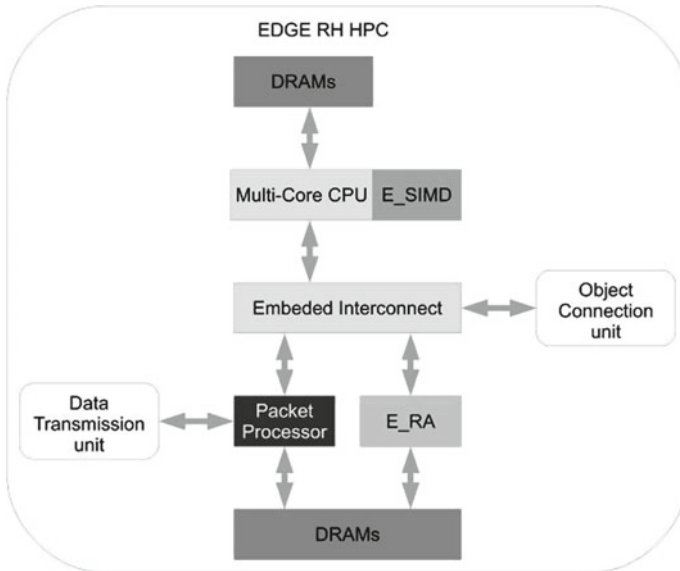


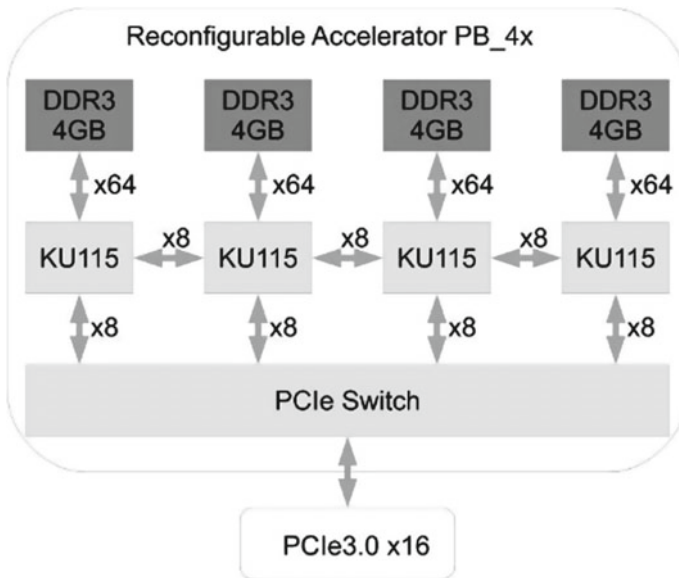
Fig. 3 The proposed architecture of the Edge RH HPC

- Embedded Interconnect, which is interconnection between all units of Edge RH HPC. The Embedded Interconnect could be implemented as a PCIe3.0 (PCIe 4.0), NV-Link, etc. interface or deployed on Logic Part of SoC device.
- Object Connection Unit, which provides interaction with particular object connected with Edge RH HPC. Object Connection Unit could include, but is not limited by, ADCs; DACs; transceivers; analog and digital sensors; GPS, GLONASS, Baidu, Galileo devices, etc.
- Data transmission unit, which is a local, remote, part of Data transmission medium (see Fig. 1). The data transmission unit should provide an arbitrary combination of wired (with speed of 1–100 GBIT/s) and wireless (e.g., Bluetooth, Wi-Fi, LTE, 5G) connections. The choice of particular media for data transmission is depended by the characteristics of the solving tasks and the remote objects parameters.

### 3.4 Developed and Deployed Reconfigurable Accelerator for DC-Cloud RH HPC

We developed and deployed in Supercomputer Center ‘Polytechnic’ [12] the reconfigurable accelerator (PB<sub>4</sub>×) [13], based on Xilinx Kintex UltraScale FPGA. The structure of the reconfigurable accelerator PB<sub>4</sub>× (see Fig. 4) was developed for providing the highest level of parallelism for reconfigurable accelerators.

The reconfigurable accelerator PB<sub>4</sub>× consists of:



**Fig. 4** The structure of developed reconfigurable accelerator PB\_4

- Four KU115 devices, each is Xilinx Kintex UltraScale KU115 FPGA. It is the largest Kintex UltraScale device available.
- Four DDR3 memory blocks. Each block, having 4 GByte capacity, is connected by 64 data bus to its own KU115 device.
- PCIe Switch, which provides non-blocking connection each of KU115 devices.

The architecture of the accelerator PB\_4 $\times$  provides the ability to connect any KU115 to any other. Throughput between KU115 and bandwidth of data channels between KU115 and DDR3 memory are balanced:

- Throughput, in any direction, between the PCIe Switch and each of KU115:  $8 \text{ lanes} \times 8 \text{ Gb/s} = 64 \text{ Gb/s}$ , provided by PCIe3.0  $\times 8$  interface.
- Throughput, in any direction, between KU115:  $8 \text{ lanes} \times 16 \text{ Gb/s} = 128 \text{ Gb/s}$ , provided by Aurora GTH  $16 \text{ Gb/s} \times 8$  interface.
- Bandwidth of data channels between KU115 and DDR3 memory:  $64 \text{ bits} \times 1600 \text{ m/s} \sim 100 \text{ Gb/s}$ .

The use of the PCIe Switch and the availability of high-speed communication channels between each KU115 allow a flexibility to change the configuration of the reconfigurable accelerator. An operating system (OS) can see PB\_4 $\times$  as:

- Four independent reconfigurable accelerator. Each of which is connected to the PCIe bus by eight channels of PCIe3.0 is implemented on Xilinx Kintex UltraScale KU115 FPGA, and has 4GByte of DDR3 memory





**Fig. 5** Top view of implemented reconfigurable accelerator PB\_4x

- One huge reconfigurable accelerator. In such configuration, a solved task will be implemented on all KU115, connected together by high-speed channels. Such huge reconfigurable accelerator will have 16 Gbit DDR3 memory divided into four independent channels.

The reconfigurable accelerator PB\_4x was implemented as PCIe3.0- $\times$ 16 [14] expansion card (see Fig. 5).

To integrate the reconfigurable accelerator into Xilinx SDAccel environment we developed [12]:

- The hardware platform, which is hardware design for each KU115 enabling integration of the reconfigurable accelerator with the host computer.
- The set of drivers enabling integration of the reconfigurable accelerator with CentOS  $\times$  64 deployed on the host computer.

By the time, a Computation Node with the developed reconfigurable accelerator is deployed in Supercomputer Center ‘Polytechnic.’

## 4 Discussion

The nearest future researches will deal with performance evaluation of developed and deployed reconfigurable accelerator. We expect to get significant leap in performance for the tasks related with Deep Neural Networks inference in Supercomputer Center.

The second direction for future researches is implementation and performance investigation of HPC EDGE unit developed in accordance with proposed architecture.

## 5 Conclusions

The chapter describes the developed hardware architecture of Reconfigurable Heterogeneous Distributed High-Performance Computing (RHD HPC) Systems, which integrates a wide set of distributed heterogeneous computing nodes. By meeting a set of requirements, such as multi-components, distributed, intelligent computer systems may be applied for wide range context-aware high-performance computing tasks.

**Acknowledgements** The authors are grateful to the Supercomputer Center ‘Polytechnic’ for the help in gaining access to the resources of the supercomputer.

## References

1. Top500 Efficiency, Power. [Online]. Available: <https://www.top500.org/statistics/efficiency-power-cores/>. Last accessed 10 Oct 2019
2. Mantovani, F., Calore, E.: Performance and power analysis of hpc workloads on heterogeneous multi-node clusters. *J. Low Power Electron. Appl.* **8**, 13. <https://doi.org/10.3390/jlpea8020013>
3. Ashraf, M.U., Alburai Eassa, F., Ahmad Albeshri, A., Algarni, A.: Performance and power efficient massive parallel computational model for HPC Heterogeneous exascale systems. *IEEE Access* **PP**(99):1–1 (2018). <https://doi.org/10.1109/ACCESS.2018.2823299>
4. Anzt, H., Ribizel, T., Flegar, G., Chow, E., Dongarra, J.: A parallel threshold ILU for GPUs. In: 33rd IEEE International Parallel and Distributed Processing Symposium, Rio de Janeiro, Brazil, 20–24 May 2019
5. Supercomputer Mixes Streams with CPU, GPU, and FPGA. [Online]. Available: <https://www.nextplatform.com/2019/04/18/supercomputer-mixes-streams-with-cpu-gpu-and-fpga/>. Last accessed 10 Oct 2019
6. Kobayashi, R., Oobata, Y., Fujita, N., Yamaguchi, Y., Boku, T.: OpenCL-ready high speed FPGA network for reconfigurable high performance computing, pp. 192–201 (2018). <https://doi.org/10.1145/3149457.3149479>
7. Xilinx FPGA. [Online]. Available: <https://www.xilinx.com/>. Last accessed 10 Oct 2019
8. Intel FPGA. [Online]. Available: <https://www.intel.com/content/www/us/en/products/programmable.html>. Last accessed 10 Oct 2019
9. IBM PowerPC9. [Online]. Available: <https://www.ibm.com/it-infrastructure/power/power9>. Last accessed 10 Oct 2019
10. OpenCAPI. [Online]. Available: <https://opencapi.org/>. Last accessed 2019/10/10
11. NVIDIA Tesla V100. [Online]. Available: <https://www.nvidia.com/en-us/data-center/tesla-v100/>. Last accessed 10 Oct 2019
12. Intel Xeon. [Online]. Available: <https://www.intel.com/content/www/us/en/products/docs/processors/xeon/2nd-gen-xeon-scalable-processors-brief.html>. Last accessed 10 Oct 2019
13. Supercomputer Center ‘Polytechnic’. [Online]. Available: <https://www.top500.org/system/178469>. Last accessed 10 Oct 2019
14. Dongarra, J.J., Gottlieb, S., Kramer, W.T.C.: Race to exascale. *Comput. Sci. Eng.* **21**(1), 4–5 (2019). <https://doi.org/10.1109/MCSE.2018.2882574>
15. Robert, Y., Le Fèvre, V., Hori, A., Bouteiller, A., Dongarra, J., Bosilca, G., Hérault, T.: Comparing the performance of rigid, moldable and grid-shaped applications on failure-prone HPC platforms. *Parallel Comput.* <https://doi.org/10.1016/j.parco.2019.02.002>

16. Wong, K., Tomov, S., Dongarra, J.: Hands-on research and training in high-performance data sciences, data analytics, and machine learning for emerging environments. In: ISC 2019, Frankfurt, 16 June 2019
17. Antonov, A.P., Filippov, A.S., Kiselev, I.O.: Design of reconfigurable computer supporting OpenCL standard. St. Petersburg State Polytech Univ J Comput Sci Telecommun Control Syst 11(4):108–118. <https://doi.org/10.18721/JCSTCS.11408>
18. Antonov, A.P., Zaborovskiy V.S.: Heterogeneous OpenCL based high performance computing system. Tech. Sci. 8:6–18 (2018)
19. Utkin, L.V., Zhuk, Y.A., Zaborovsky, V.S.: An anomalous behavior detection of a robot system by using a hierarchical Siamese neural network. In: 2017 XX IEEE International Conference on Soft Computing and Measurements (SCM), pp. 630–634. IEEE (2017)
20. PCIe standard. [Online]. Available: <https://pcisig.com/specifications>. Last accessed 10 Oct 2019

# Usage of a BART Algorithm and Cognitive Services to Research Collaboration Platforms



Sergey Saradgishvili  and Ilia Voronkov 

**Abstract** The ability to predict behavior in complex systems has always interested scientists. With the development of science, there is a gradual complication of systems and data processing methods in them. In this publication, work in the system is achieved by emulating interaction at various levels and nodes through collaboration platforms. The key idea of the whole area is the ability to predict the behavior of the system node based on experience and data obtained at previous stages of work. Improvement of such approaches in the future can have a serious impact on the process of improvement and the evolutionary transition to systems that are currently impossible to imagine. This transition remains impossible until humanity has learned to work effectively in the current collaboration platforms. The paper considers an algorithm for processing the obtained data and its extension using existing cognitive services for the analysis of texts. In the future, the algorithm may be expanded to work with visual information.

**Keywords** Collaboration · BART · CRISP-DM · Azure

## 1 Introduction

### 1.1 Research Workflow

The research process is reduced to the systematization of data obtained using various research methods (observation, experiment, and others). In this work, authors try to apply a similar algorithm to specific systems, which in the study are called collaboration platforms. Such systems may include information services, which basically contain the idea of a group of users working on one or several objects in this system. The simplest example of the implementation of such a system can be an ordinary dialogue, only it occurs by means of messaging on electronic devices, and it is

---

S. Saradgishvili · I. Voronkov (✉)  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [iliaftk@outlook.com](mailto:iliaftk@outlook.com)

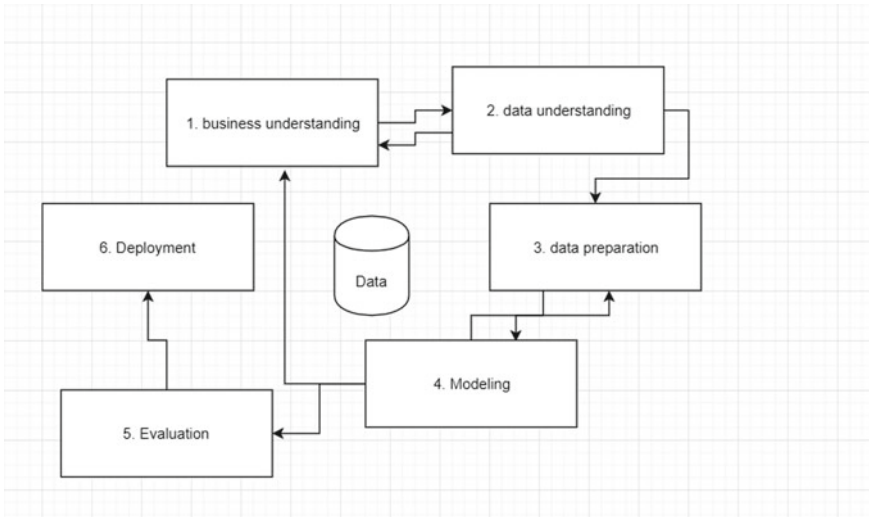
possible to predict the behavior of the participants in the dialogue using an analysis of previously received messages. In this work, the authors sought to emulate the work of the enterprise (company, plant) using modern information resources:

- (1) Office 365 provides work with e-mail (as the most common way of correspondence in companies and provides the ability to use the chat system (Skype for business, Teams). The authors deliberately turned off the use of media files to simplify the work.
- (2) Emulation of the operation of an industrial device (conveyor) [1].
- (3) Document Management System based on the SharePoint online package [2].
- (4) SAP modules with Azure blockchain [3].

These components of the system emulated problems, worked with documents (real users), and processes for maintaining contracts. Based on the results obtained, time series analysis was applied using the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology [4]. The effectiveness of using this approach is found as a module of the difference in the value of the estimate at the forecasting stage with the value at each stage of the iteration.

## ***1.2 Modeling and Investigations Approach in Collaboration Systems***

Most often, researchers focus on highlighting interactions between users of the system to fully describe and understand the processes involved [5, 6]. The topic of research on the impact of work in shared access systems is very popular in scientific papers. This emphasis can be explained by the study of the impact of collaboration between professional groups on issues [7–11]. The implementation of the technical component of the solution for the creation, support, and application of work algorithms in systems is also widely used in articles [12, 13]. The technology that is used in this work to form an enterprise model is also used to develop training systems, where the factor of perception of educational material in the system plays a key role [14]. Despite the presence of works on various topics, most of them focus on a specific issue based on the purpose of the study: the technical structure or the impact on the work of the team. This work sets its task, using the experience of previous years, to look at the process of forming interactions from a new level: the system at the same time is limited in itself and at the same time can be affected from outside.



**Fig. 1** Schematic implementation CRISP-DM

## 2 Modeling Interactions in a Collaboration Environment

### 2.1 Usage of CRISP-DM

We use CRISP-DM to implement a model for collecting and analyzing information that is generated by users in collaboration platforms. Figure 1 presents 6 steps that are in the cycle of this approach.

This method represents a continuous cycle of work on data, during which  $N$  is performed several times, the error in the predicted data tends to zero based on the information obtained in the previous step. This method is one of the most common practices for modeling the data mining process.

### 2.2 Regression Tree

A regression tree is a class of regression models that allows you to divide the input space of factor variables into segments. Subsequently, the whole chain of the regression model can be supplemented and processed for each of the nodes representing the regression function in an intuitive visual form [15, 16]. Let us introduce a definition of how the interaction model in a collaboration environment relates to a process expression in the form of a tree:

1. The internal node of the tree is a description of the rules for partitioning the space of explanatory variables. (e.g., An element of the “message” system is an

**Table 1** Types of decision trees

Tree name	Description	Example in systems
For classification	The result of the prediction is the data ownership class	Office 365 message
For regression	The result in this case is the predicted value of the target function	Media File in SharePoint

indivisible unit for selecting data for a finite segment of users. One author can have many messages, while each message has a single author in a simplified model.)

2. Tree Leaves—Own Model of Local Regression.
3. Branch—transition conditions between nodes.

To apply the regression tree, we will build connections in the collaboration system on the principle of dividing and isolating a set of information that does not intersect with other classes of the system. We will carry out the division sequentially until there remains the possibility of isolating a new class from the essence of the entity. Each model can be defined and limited by itself. Types of decision trees for collaboration are presented in Table 1.

### 2.3 *BART Algorithm*

Consider a Bayesian approach to evaluating nonparametric functions using regression trees. This algorithm allows us to generalize the regression tree that was allocated above and time series for iterations using the CRISP-DM method. BART is a combination (C&RT) [17] of the algorithm and the standard autoregressive integrated moving average (ARIMA) models and their components (AR, MA The SETAR and ASTAR models are linear models of homogeneous models (Inst message, office 365 Exchange file) that build several adaptive regression splines (MARS) based on time series in a single iteration of the processing complex. BART has two main differences from the SETAR and ASTAR models:

1. Error estimates for BART models can differ from each other both for each node and for each iteration of the cycle.
2. BART is characterized by a gap between the models of autoregression.

To convert the model according to the time series, we will use the conversion method, where the resulting variable  $CD_t$  corresponds to the sum of the previous value  $CD_{t-1}$  and the delayed value  $CD_{t-p}$  adjusted for the Sentiment coefficient  $\beta$ , which in turn, it is a cumulative estimate for each node of the system over time  $t$  obtained using Azure Data Text functions. Release the rules for dividing input data into segments (Fig. 2).

Most algorithms use recursive separation of the data on which training takes place. BART uses an iterative construction method. Also, in our example, we add

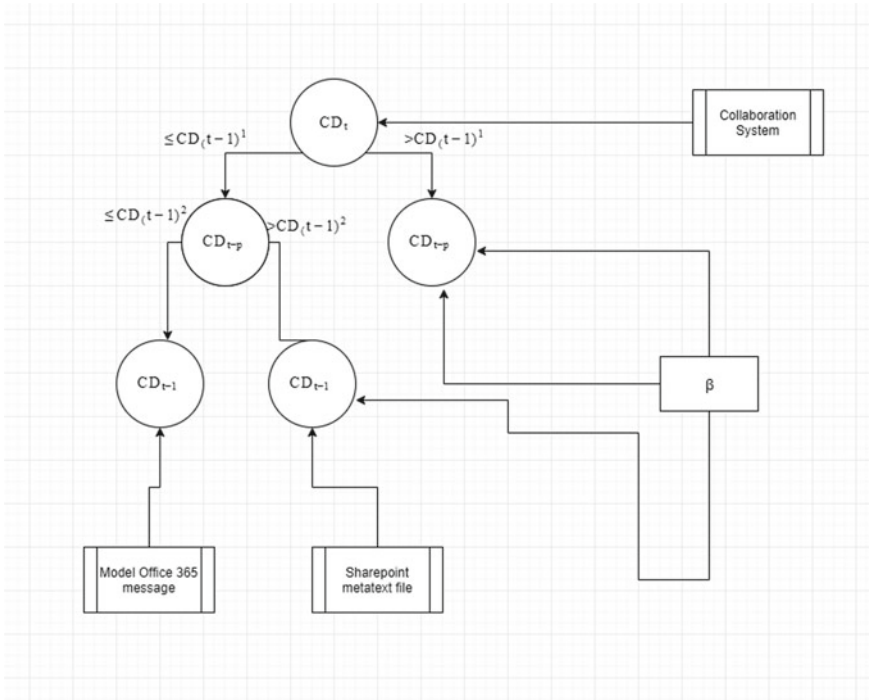


Fig. 2 Autoregressive tree building diagram

the Sentiment  $\beta$  coefficient at each step of the calculation, which is in the range of values [0.1] and is cumulative. This coefficient is calculated for each individual and influences the result in each node. 0 is the correspondence of an absolute negative comment in the system (message in correspondence, with pronounced indignation or discontent). 1 is a positive value. During the simulation, no 0 to 1 were met.

Consider a data processing algorithm:

*Step 1:* Build a regression tree for the root value (the entire Collaboration system considered). The construction of a regression tree begins from a single value (root node), which is defined as the Median (ME, second quartile  $Q_{50\%}$ ) of the entire time series  $CD_i^1$  and is calculated the Eq. (1).

$$ME = Q_{50\%} = \frac{1}{2}(CD_i^{\min} + CD_i^{\max}) + \beta \tag{1}$$

In this expression, the median is the sum of a real number with a probability of exceeding an arbitrary size equal to 0.5 and a non-negative value  $\beta$ , which tends to 1 with an ideal system of interaction in the system.

*Step 2:* For each unprocessed node we find the best partition. The partition itself will be selected based on a predefined rule.



The split rule selection mechanism is like the C&RT algorithm. The difference lies in the rule for selecting the criterion for evaluating the termination of cleavage. During testing, we used the information criterion for a better separation based on the entropy indicator, because it prefers options with less complexity of the tree. This algorithm will determine an entropy information gain. When using BART, any vertex in the tree (except the root) has two children. The final chain is built from tree nodes from top to bottom and an informational assessment of the node predictor occurs, dividing the time series into a subset of experiments.

*Entropy criterion.* The value for the sample  $CD_i^j$  is calculated by the formula (2):

$$H^{(P,N)} = H(P/(P+N) + N/(P+N)) \quad (2)$$

After receiving information about the current node and its predecessors, the entropy value is calculated by the following relation (3):

$$H_\varphi(P, N, p, n) = (p+n/P+N)H^1(p, n) + (P+N-(p+n)/P+N)H^{(P-p, N-n)} \quad (3)$$

$P$  is the number of objects in the subset  $C$ ,  $p$  is the number of objects, all objects that correspond to  $p \in P \in nN$  are the total number of nodes in the system at each iteration step  $n \in N$ .

The change in entropy is calculated by the formula (4) which shows the amount of information corresponding to class  $C$  and not corresponding to this separation:

$$IGain_C(\varphi, CD_1^i) = H^{(P,N)} - H^{(P,N,p,n)} + \beta \quad (4)$$

For the early Q information criterion, we use the extended Bayesian information criterion [18] (5):

$$EBIC = J * (\ln n + 2 \ln p) + n \ln n(SSE/n) \quad (5)$$

SSE is the sum of squares of the residuals of the model;  $J$  is the number of model parameters;  $n$  is the training number of examples;  $p$  is the mathematical ratio of the number of vertices in the tree and the unifying criteria. In a first approximation, we will use multiplication. We will calculate the EBIC value until the next value is less than the previous one.

*Step 3:* Continue dividing the model and consider the change in entropy until the EBIC value at step  $n$  is greater than at step  $n - 1$  and is a real number. As soon as these conditions are no longer fulfilled, the algorithm completes its work. The tree is considered formed.

We will study the effectiveness of the BART +  $\beta$  approach with traditional ARIMA (autoregressive integrated moving average) algorithms. We will build a series of experiments on the leaves of a tree according to the rule (6):

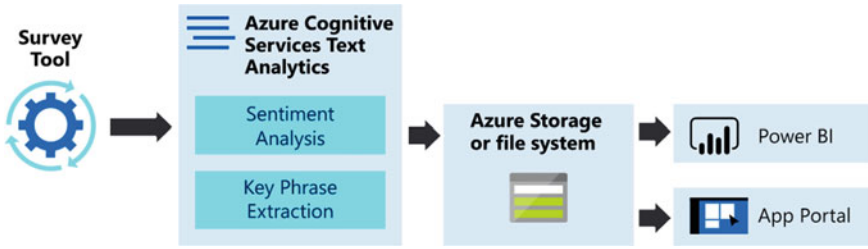


Fig. 3 Azure Cognitive Services as service

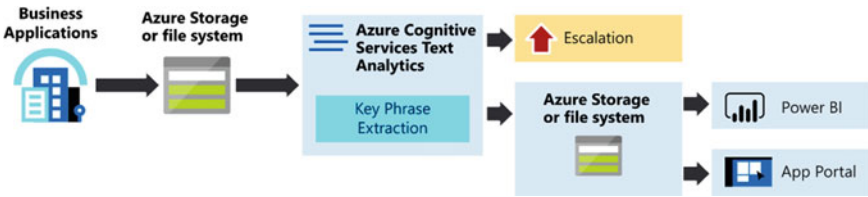


Fig. 4 Azure Cognitive Services as service for escalation

$$F(L)(1^d + d(n - 1)L^{d-1})X_t = \mu + \theta(L)\varepsilon_t, \varepsilon_t \sim N(0, \sigma^2) \tag{6}$$

$X_t$ —time series,  $s, L$ —is the lag operator,  $F(L)$ —is the polynomial degree  $p$  from  $L$ ,  $\mu$ —is the average process value,  $d$ —is the order of process integration  $X_t$ . If  $d = 0$ , then process  $X_t$  can be described by ARIFMA ( $p, q$ ) or ARIMA ( $p, 0, q$ ). During the simulation, we also conducted studies using the classical configuration parameters for the algorithms: ARIMA (1, 0, 1) and ARFIMA (1,  $d$ , 1).

*Step 4:* We will describe methods of receiving  $\beta$ . In this study, we used the existing Azure Cognitive services functionality to calculate the correction at each new iteration, considering the results of previous values for interactions in the system. This approach is since the calculation of such a coefficient is a separate area and there are several approaches to solving this problem. In our experiment, we use this service according to the black box principle: We do not know the internal structure of the service, but we introduce a correction value in the equation. Options for using the service are presented in Figs. 3 and 4 [19].

### 3 Empirical Results

We simulated the work of the enterprise within 30 days, with the participation of real users (who talked in the mail, worked with internal messengers, used the workflow system on SharePoint, reacted to systems simulating industrial capacities), and taking into account randomly generated events (breakdowns, failures, external contact by

**Table 2** Simulation results for each individual node

System node	Description of Interactions	Interactions with other system nodes	The presence of interaction from an external source	The difference between the predicted value and the real
Exchange e-mails	E-mail Correspondence	Chats, SharePoint DMS	True	0.245
Chats	Dialog daily conversations	E-mails	False	0.113
SharePoint DMS	Document workflow	Exchange	False	0.117
Imitation conveyor belt on Raspberries	The work of the product creation system	SAP	False	0.430
SAP modules	Customer workflow	Conveyor	True	0.175

mail and more). According to the idea of the experiment, for each interaction node, the target variable will be the log-return value for the time period after the incident. The results of the experiment are displayed in Table 2.

The closer the value in the last column to zero, the more accurate the forecast gives the system about the results of future interaction. Note some interesting features in the experimental results:

1. The best value for individual nodes showed a study of the chat system. This node is not critical for maintaining the efficiency of the enterprise model, as E-mail can be duplicated. This fact and the fact that in the calculation we used a cognitive analysis of chat text messages that can be deeply analyzed by the system at each iteration can explain the leadership in using this approach for text messages.
2. The relatively high (second) place of SharePoint DMS is due to the formalization of processes, where in the workflow for processing documents it is not possible to introduce serious disturbances (comments). At the same time, the usefulness of using cognitive services in such formalized systems can be questioned.
3. The relative success of modeling in SAP modules is also explained by the high formalization and standardization of processes. The result obtained (the value is worse than that of the previous paragraphs) may indicate the imperfection of the method when working with external disturbances and the difficulty of predicting emergency situations for the conveyor.
4. The simulation results for the conveyor may indicate the imperfection of the applied approach for a particular node.

## 4 Conclusion

During the research, an enterprise model was developed that uses various collaboration nodes to organize work. To predict the results of individual components based on the previous values, an extended implementation of the BART algorithm was used, which showed effectiveness in studying the results of individual nodes. In the future, this approach can be laid in the idea of applying emergency forecasting, possible safety problems [20] and other critical components in the enterprise. In future studies, the authors will strive to complicate and approximate the enterprise model to the existing ones; at the same time, it is planned to use various methods for working with a large amount of data on time series.

## References

1. Jo, B., Khan, R.M.A.: An internet of things system for underground mine air quality pollutant prediction based on azure machine learning. *Sensors (Basel)* **18**(4), 930 (2018). <https://doi.org/10.3390/s18040930>
2. Buchal, R., Songsore, E.: Collaborative knowledge building using Microsoft SharePoint. In: Proceedings of the 2018 Canadian Engineering Education Association (CEEA-ACEG18) Conf (2018)
3. Wang, Y., Lahiri, S.K., Chen, S., Pan, R., Dillig, I., Born, C., Naseer, I.: Formal Specification and Verification of Smart Contracts for Azure Blockchain, pp. 21–42 (2018). <https://doi.org/10.1023/j.compind.2018.08.257>
4. Kurgan, L., Musilek, P.: A survey of knowledge discovery and data mining process models. *Knowl. Eng. Rev.* **21**(1), 1–24 (2006)
5. Briggs, R.O., Kolfshoten, G.L., de Vreede, G.J., Dean, D.L.: Defining key concepts for collaboration engineering. In: Americas Conference on Information Systems, Acapulco, Mexico (2006)
6. Frost, S.: Meetings around the World: The Impact of Collaboration on Business Performance (2007)
7. Dennis, A.R., Nunamaker, J.F.J., Vogel, D.R.: A comparison of laboratory and field research in the study of electronic meeting systems. *J. Manage. Inf. Syst.* **3**, 107–135 (1991)
8. Soller, A., Martínez, A., Jermann, P.: From mirroring to guiding: a review of state of the art technology for supporting collaborative learning. *Int. J. Artif. Intell. Educ.* **15**, 261–290 (2005)
9. Webster, J., Staples, D.S.: Comparing virtual teams to traditional teams: An identification of new research opportunities. In: Martocchio, J.J. (ed.) *Research in Personnel and Human Resources Management*, vol. 25, pp. 181–215. JAI Press, San Diego, CA (2006)
10. Fjermestad, J., Hiltz, S.R.: An assessment of group support systems experimental research: methodology and results. *J. Manage. Inf. Syst.* **15**, 7–149 (1999)
11. Fjermestad, J., Hiltz, S.R.: A descriptive evaluation of group support systems case and field studies. *J. Manage. Inf. Syst.* **17**, 115–159 (2001)
12. Bostrom, R., Anson, R., Clawson, V.K.: Group facilitation and group support systems. In: Jessup, L.M., Valacich, J.S. (eds.) *Group Support Systems: New Perspectives*. Macmillan (1993)
13. Witte, E.H.: Toward a group facilitation technique for project teams. *Group Processes & Intergroup Relations* **10**(3), 299–309 (2007)
14. Alter, A.: Collaboration: Unlocking the Power of Teams. *CIO Insight* (2009). Retrieved April 15, from <https://www.cioinsight.com/c/a/Resea>

15. Breiman, L., Friedman, J., Stone, C., Olshen, R.: Classification and Regression Trees. Chapman and Hall/CRC, Boca Ration (1984)
16. Breiman, L.: Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
17. Tong, H.: Threshold models in nonlinear time series analysis. Springer-Verlag, New York (1983)
18. Chen, J., Chen, Z.: Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**(3), 759–771 (2008)
19. Azure Cognitive Services: <https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/>. Last accessed 1 Oct 2019
20. Kim, Y., Kim, J., Kim, W., Im, J.: Predicting fluctuations in cryptocurrency transactions based on user comments and replies. *PLoS ONE* **17**, 1–17 (2016). <https://doi.org/10.1371/journal.pone.0161197>

# A Computer-Aided Diagnosis System in the Diagnosis of Multiple Sclerosis



Polina Andropova , Dmitriy Cheremisin , and Anna Meldo 

**Abstract** This chapter reviews the basics and recent researches of computer-aided diagnosis (CAD) systems for assisting neuroradiologists in the detection, monitoring and prediction of multiple sclerosis (MS) in magnetic resonance (MR) images. The CAD systems consist of image feature extraction based on image processing techniques and machine learning classifiers such as linear discriminant analysis, artificial neural networks, and support vector machines. We introduce useful examples of the CAD systems in the neuroradiology and conclude with possibilities in the future of the CAD systems for MS in MR images.

**Keywords** Machine learning · Computer-aided diagnosis (CAD) systems · Artificial intelligence · Convolutional neural network · Multiple sclerosis

## 1 Introduction

Multiple sclerosis (MS) is a neurodegenerative disease characterized by chronic demyelination of the central nervous system (CNS) and which, as it develops, severely compromises patient quality of life. Although the cause of the disease remains unknown, it is assumed to be due to complex interactions between genetic and environmental factors. MS is not currently curable. The aim, therefore, is to diagnose it early, make competent monitoring and to provide treatment that reduces the risk of relapse and the progression of disability [1].

The current criteria used to diagnose forms of MS were originally formulated by [2] and revised by [3] and [4]. Diagnosis should consider evidence of damage to the CNS disseminated in time (on different dates) and in space (damage to at least two

---

P. Andropova (✉) · D. Cheremisin

N. P. Bechtereva Institute of the Human Brain of the Russian Academy of Sciences (IHB RAS),  
St. Petersburg, Russia

e-mail: [polin.and@icloud.com](mailto:polin.and@icloud.com)

A. Meldo

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

different parts of the CNS) and should exclude other conditions that, due to their clinical or laboratory profile, can mimic MS.

MS declares itself in a variety of clinical and imaging forms, differentiating both within patients' group and within individual patients in a time course. A diagnosis of MS could be revealed only by confluence of clinical, imaging, and laboratory findings, as there is no definite clinical or diagnostic feature of such entity. MRI features of MS mask themselves and mimic MRI abnormalities associated with other diseases and non-specific MRI findings, commonly found in a general population. Pursuing on timely diagnosis to lift a burden of uncertainty for patients and start modifying therapy could lead clinicians to increased risk of misdiagnosis [5].

While evaluating a patient, physicians rely on pattern recognition as human intuition dictates. Rational diagnosis making requires that clinical patterns be put in the context of disease prior to probability, yet physicians often exhibit flawed probabilistic reasoning. High rates of costly and even fatal diagnostic errors are the result of such issues in making a diagnosis. While being introduced more than half-century ago, computer-aided diagnosis systems and software remain not widely integrated into clinical routine. These systems cannot efficiently recognize patterns and are unable to consider the base rate of potential diagnoses [6].

Solomon et al. [7] conducted a study to assess the proper application of McDonald Criteria key elements among neurology residents and MS specialists. It was shown that neurology residents and their counterparts (with less mistakes though) inaccurately identified essential components of McDonald Criteria.

In MS, early treatment has been considered the best strategy [8], hence qualitative and faster clinical decisions are needed. In line with this, computer-aided diagnosis (CAD) has become one of the most important areas of research [9].

## 2 Evaluation of CAD Systems

In recent years, CAD software development and its implementation was rapidly expanding, which lead to promising results, improving accuracy and increased sensitivity in a variety of medical fields, such as proctology [10], ophthalmology [11], mammalogy [12], cardiology [13], urology [14, 15], radiology [16].

With that being said, many systems for diagnosing MS have been developed as well to ensure the successful detection of focal brain damage, monitoring, predicting the severity of neurological disorders and for a more informed decision about the start of treatment of the disease. Each method of reasoning has certain capabilities and limitations, with varying effectiveness and application possibilities in the diagnosis of this disease.

## 2.1 CAD Systems for Early Diagnosis of Multiple Sclerosis

The identification of MS was the predominant goal of researchers. Below we present some of the most successful CAD systems in this area.

In 2017, Wu et al. [17] presented an MS slice identification system, based on Haar wavelet transform, principal component analysis, and logistic regression. The accuracy of their software product was 87.65%.

Next, with the help of the Minkowski-Bouligand measurement and a neural network with one hidden layer, Zhang et al. [18] created a CAD with sensitivity, specificity, and accuracy—97.78%, 97.82%, and 97.77%, respectively. Data augmentation was used to increase the size of the training set. Improved convolutional neural network combined the parametric rectified linear unit (PReLU) and dropout techniques. Finally, a 10-layer deep convolutional neural network was established, with 7 convolution layer and 3 fully connected layers.

In recent studies, Wang et al. [19] proposed a new method for diagnosing MS using a CAD based on a 14-Layer convolutional neural network with batch normalization, dropout, and stochastic pooling which provided a sensitivity of  $98.77 \pm 0.35\%$ , specificity of  $98.76 \pm 0.58\%$ , and accuracy of  $98.77 \pm 0.39\%$ . Comparison showed that their method is superior to modern methods of artificial intelligence. During this comparison it was revealed that stochastic pooling algorithm gives better results than maximum and average pooling ones. Furthermore, Wang et al. compared proposed method with six state-of-the-art approaches, including five traditional artificial intelligence methods and one deep learning method. The comparison shows those method has a better outcome value than all other six state-of-the-art approaches.

Experimental results of previously mentioned segmentation methods of MS lesions on MR images reveal that representation deep learning-based methods are the most accurate and promising ones for the early detection of multiple sclerosis.

The unprecedented success of developers gives us clear data on the superiority of CAD systems over people in the detection of multiple sclerosis.

## 2.2 CAD Systems for Monitoring and Prognosis of Multiple Sclerosis

The growth of plaques and the development of new lesions in MRI are markers of new disease activity in MS patients. Treatment efficacy and understanding of worsening course of the disease heavily rely on our ability of successfully predicting future lesion activity. Doyle et al. [20] introduced the first, fully automatic, probabilistic framework for the prediction of future lesion activity in relapsing–remitting MS patients, based only on baseline multi-modal MRI, and use it to successfully identify responders to two different treatments. The authors developed a new Bag-of-Lesions (BoL) representation for patient images based on a battery of unique characteristics extracted from lesions. A probabilistic codebook of lesion types is



created by clustering features using Gaussian mixture models. Patients are represented as a probabilistic histogram of lesion types. A Random Forest classifier is trained to automatically predict future MS activity up to two years ahead based on the patient's baseline BoL representation. Automated identification of responders in two different treated groups of patients leads to sensitivity of 82% and 84% and specificity of 92% and 94% respectively, showing that this is a very promising approach toward personalized treatment for MS patients.

A crucial step in the management of patients with MS is an ability to steadily predict a worsening of the disease over the short-time period. A critical component in the management of patients with MS is correctly predicting which patients will experience worsening disease over the short term. This is particularly relevant given the expanding array of disease-modifying medications and the importance of identifying the patients who may benefit from more potent or aggressive treatment or closer monitoring. An identification of patients with a high risk of attacks within a given time frame gives clinicians an ability to treat such patients more actively thus improving course of their ailment. Yoo et al. [21] conducted a research to determine an ability of deep learning to extract, from segmented lesion masks, covert signals that able to predict spark in the disease's course within short timeframe in patients with early MS symptoms more accurately than lesion volume, generally used as MS imaging biomarker. They used convolutional neural networks to extract latent MS lesion patterns that are associated with early disease activity using lesion masks computed from baseline MR images. Main obstructions are that lesion masks are sparse resulting in a scarce number of samples relative to the dimensionality of the images. To cope with sparse voxel data, Yoo et al. [21] propose utilizing the Euclidean distance transform (EDT) for increasing information density by populating each voxel with a distance value. These prediction model can achieve an accuracy rate of 72.9% (SD = 10.3%) over 2 years using baseline MR images only, which is significantly higher than the 65.0% (SD = 14.6%) that is attained with the traditional MRI biomarker of lesion load.

A complex anatomical distribution is a main feature of pathological process in case of multiple sclerosis, and commonly used low-dimensional models could not firmly detect such changes. Thereby management of individual patients and interventional trials suffer from inadequate ability to detect treatment effects. Kanber et al. [22] conducted a comparison between conventional models and high-dimensional models incorporating a plethora of imaging factors in an ability to detect an imaging changes in response to provided treatment. They used fully automated image analysis to extract 144 regional, longitudinal trajectories of pre- and post-treatment changes in brain volume and disconnection in a cohort of 124 natalizumab-treated patients. Low- and high-dimensional models of the relationship between treatment and the trajectories of change were built and evaluated with machine learning, quantifying performance with receiver operating characteristic curves. Compared to existing methods, high-dimensional models were superior in treatment response detection (area under the receiver operating characteristic curve = 0.890 [95% CI = 0.885–0.895] versus 0.686 [95% CI = 0.679–0.693],  $P < 0.01$ ) and in statistical efficiency (achieved statistical power = 0.806 [95% CI = 0.698–0.872] versus 0.508 [95% CI =

0.403–0.593] with number of patients enrolled = 50, at  $\alpha = 0.01$ ). The study strongly suggested to incorporate high-dimensional models in the routine clinical imaging due to substantial enhancement in the detection of imaging changes in response to treatment, allowing clinicians enable more precise individual approach and prognosis, and greatly improving statistical data output for randomized controlled trials.

Tousignant et al. [23] presented the first automatic end-to-end deep learning framework for the prediction of future patient disability progression (one year from baseline) based on multi-modal brain Magnetic Resonance Images (MRI) of patients with MS. The model uses parallel convolutional pathways and is trained and tested on two large proprietary, multi-scanner, multi-center, clinical trial datasets of patients with Relapsing–Remitting MS (RRMS). Using only the multi-modal MRI provided at baseline, the model achieves an AUC of  $0.66 \pm 0.055$ . However, when supplemental lesion label masks are provided as inputs as well, the AUC increases to  $0.701 \pm 0.027$ . Clinicians provided with the predictions computed by the model can, therefore, use the associated uncertainty estimates to assess which scans require further examination.

Further development of systems for predicting MS is needed. This is important because it will allow us to control this disease, to prevent the disability of patients by changing therapy.

### 3 Cognitive Assessment in Multiple Sclerosis

Noteworthy is the study by Khaligh-Razavi et al. [24] by creating a self-administered, artificial intelligence (AI) platform for cognitive assessment in MS.

Cognitive impairment is common in patients with MS. Accurate and repeatable measures of cognition have the potential to be used as a marker of disease activity. Authors developed a 5-min computer test—named Integrated Cognitive Assessment (ICA)—that is self-administered and language-independent to evaluate a degree of cognitive impairment in patients with MS.

ICA demonstrated excellent test–retest reliability ( $r = 0.94$ ), with no learning bias. ICA was sensitive in discriminating against the MS patients from the healthy controls (HC) group and demonstrated a high accuracy (AUC = 95%) in discriminating cognitively normal from cognitively impaired participants.

ICA can be used as a digital biomarker for assessment and monitoring of cognitive performance in MS patients. Its advantages are shorter duration, not being learning biased, independent of language and utilizing performance capabilities of artificial intelligence in more definitive identification of the cognitive status of users. Since ICA is a digital test, it is possible for it to be integrated in electronic health record and further in research database with ease.

However, there are some limitations to this system.

As it was aforementioned, imaging changes could be detected far before clinical deterioration, so a pillar of treatment tactics, prediction and prognoses of relapses remains in a neuroradiological imaging. Clinical measures could not grasp

on all affected functional domains, does not have enough evidence to predict the progression of disability.

Such platform does not incorporate neuroradiological method, so it can go as only refining tool for diagnosis and monitoring of cognitive deficiency in patients with multiple sclerosis.

## 4 Conclusion

Intelligent computer systems are used in diagnosing MS and help physicians in the accurate and timely detection, monitoring, and prediction of the disease.

Rule-based, fuzzy-logic, and artificial neural network methods have had more applications in intelligent systems for the diagnosis of MS. The highest rate of sensitivity and accuracy indexes is associated with the neural network reasoning method at  $98.77 \pm 0.35\%$ , and  $98.76 \pm 0.58\%$ , respectively.

CAD systems for monitoring and prognosis of MS are still trailing human expertise on both detection and delineation criteria. In addition, we demonstrate that computing a statistically robust consensus of the algorithms performs closer to human expertise.

One of the authors of the CAD for detection of imaging response to treatment in multiple sclerosis, Dr. Nachev [25], said the algorithm combined “the finesse of a clinician with the objectivity of a machine,” adding: “Rather than copying what radiologists already do, this kind of complex modelling does what no human being can, drawing intelligence from a very rich data set to enable care that is individually tailored yet objective and reproducible.”

CAD systems require further improvement and greater accuracy.

At the moment, in the diagnosis of multiple sclerosis, in the future, it is possible to automate all stages of diagnosis. The opinions of different scientists are divided on the role of human in the diagnosis of multiple sclerosis: from the complete replacement of a person to fruitful cooperation and the performance of CAD functions inaccessible to humans. What will be more effective will be shown by the unprecedented ongoing progress in the field of artificial intelligence.

## References

1. Huang, W.J., Chen, W.W., Zhang, X.: Multiple sclerosis: pathology, diagnosis and treatments. *Exp. Ther. Med.* **13**(6), 3163–3166 (2017)
2. McDonald, W.I., Compston, A., Edan, G., Goodkin, D., Hartung, H.P., Lublin, F.D., et al.: Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis. *Ann. Neurol.* **50**, 121–127 (2001)
3. Polman, C.H., Reingold, S.C., Banwell, B., Clanet, M., Cohen, J.A., Filippi, M., et al.: Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann. Neurol.* **69**, 292–302 (2011)

4. Thompson, A.J., Banwell, B.L., Barkhof, F., Carroll, W.M., Coetzee, T., Comi, G., et al.: Diagnosis of multiple sclerosis: 2017 revisions of the McDonald criteria. *Lancet Neurol.* **17**, 162–173 (2018)
5. Solomon, A.J., Corboy, J.R.: The tension between early diagnosis and misdiagnosis in multiple sclerosis. *Nat. Rev. Neurol.* **13**, 567–572 (2017)
6. Cahan, A., Cimino, J.J.: A learning health care system using computer-aided diagnosis. *J. Med. Internet Res.* **19**(3), e54 (2017)
7. Solomon, A., Pettigrew, R., Naismith, R., Chahin, S., Krieger, S., Weinschenker, B.: Challenges in multiple sclerosis diagnosis: misapplication and misunderstanding of the McDonald criteria (S6.001). *Neurology*, 92 (2019)
8. Fernández, O., Fernández, V., Arbizu, T., Izquierdo, G., Bosca, I., Arroyo, R., et al.: Characteristics of multiple sclerosis at onset and delay of diagnosis and treatment in Spain (The Novo Study). *J. Neurol.* **257**, 1500–1507 (2010)
9. Nayak, D.R., Dash, R., Majhi, B., Prasad, V.: Automated pathological brain detection system: a fast discrete curvelet transform and probabilistic neural network-based approach. *Expert Syst. Appl.* **88**, 152–164 (2017)
10. Mori, Y., Kudo, S.E., Berzin, T.M., Misawa, M., Takeda, K.: Computer-aided diagnosis for colonoscopy. *Endoscopy* **49**, 813–819 (2017)
11. Hagiwara, Y., Koh, J.E.W., Tan, J.H., Bhandary, S.V., Laude, A., Ciaccio, E.J., Tong, L., Acharya, U.R.: Computer-aided diagnosis of glaucoma using fundus images: a review. *Comput. Methods Programs Biomed.* **165**, 1–12 (2018)
12. Saha, M., Mukherjee, R., Chakraborty, C.: Computer-aided diagnosis of breast cancer using cytological images: a systematic review. *Tissue Cell.* **48**, 461–474 (2016)
13. Faust, O., Acharya, U.R., Sudarshan, V.K., San, T.R., Yeong, C.H., Molinari, F., Ng, K.H.: Computer-aided diagnosis of coronary artery disease, myocardial infarction and carotid atherosclerosis using ultrasound images: a review. *Phys. Med.* **33**, 1–15 (2017)
14. Wang, S., Burt, K., Turkbey, B., et al.: Computer aided-diagnosis of prostate cancer on multi-parametric MRI: A technical review of current research. *Biomed. Res. Int.* **2014**, 789561 (2014)
15. Wang, Q., Li, H., Yan, X., et al.: Histogram analysis of diffusion kurtosis magnetic resonance imaging in differentiation of pathologic Gleason grade of prostate cancer. *Urol. Oncol. Semin. Orig. Invest.* **33**, 337 (2015)
16. Zaglam, N., Cheriet, F., Jouvret, P.: Computer-aided diagnosis for chest radiographs in intensive care. *J. Pediatr. Intensive Care* **5**(3), 113–121 (2016)
17. Wu, X., Lopez, M.: Multiple sclerosis slice identification by Haar wavelet transform and logistic regression. *Front. Neurosci.* **12**, 818 (2018)
18. Zhang, Y.-D., Pan, C., Sun, J., Tang, C.: Multiple sclerosis identification by convolutional neural network with dropout and parametric ReLU. *J. Comput. Sci.* **28**, 1–10 (2018). <https://doi.org/10.1016/j.jocs>
19. Wang, S.-H., Tang, C., Sun, J., Yang, J., Huang, C., Phillips, P., et al.: Multiple sclerosis identification by 14-layer convolutional neural network with batch normalization, dropout, and stochastic pooling. *Front. Neurosci.* **12**, 818 (2018). <https://doi.org/10.3389/fnins.2018.00818>(2018)
20. Doyle, A., Precup, D., Arnold, D.L., Arbel, T.: Predicting future disease activity and treatment responders for multiple sclerosis patients using a bag-of-lesions brain representation. In: *Medical Image Computing and Computer Assisted Intervention—MICCAI*, pp. 186–194 (2017)
21. Yoo, Y., Tang, L., Brosch, T., Li, D., Metz, L., Traboulsee, A., Tam, R.: Deep learning of brain lesion patterns for predicting future disease activity in patients with early symptoms of multiple sclerosis. *LABELS/DLMIA@MICCAI* (2016)
22. Kanber, B., Nachev, P., Barkhof, F., Calvi, A., Cardoso, J., Cortese, R., Prados, F., Carole, H., Sudre, C., Tur, C., Ourselin, S., Ciccarelli, O.: High-dimensional detection of imaging response to treatment in multiple sclerosis. *Digital Med.* **2**, article number: 49 (2019)

23. Tousignant, A., Lemaître, P., Precup, D., Arnold, D., Arbel, T.: Prediction of progression in multiple sclerosis patients. *Proc. Mach. Learn. Res.* **102**, 483–492 (2019)
24. Khaligh-Razavi, S.-M., Sadeghi, M., Khanbagi, M., Kalafatis, C., Nabavi, S.: A self-administered, artificial intelligence (AI) platform for cognitive assessment in multiple sclerosis (MS) (2019). bioRxiv 611335. <https://doi.org/10.1101/611335>
25. <https://www.ft.com/content/2ade7112-8927-11e9-a028-86cea8523dc2>

# Predicting Students' Performance on MOOC Using Data Mining Algorithms



Sergey Nesterov , Elena Smolina , and Tigran Egiazarov 

**Abstract** This paper describes the results of experiments in predicting students' performance on a massive open online course (MOOC). Grade reports from MOOC “Data management” on the Russian platform [openedu.ru](http://openedu.ru) were used for the analysis. It is well known that only a small percent of students who enrolled in MOOCs pass them through. Data mining methods could help to understand the causes of this problem. We tried to predict whether the student will finish an online course or not based on his results during the first weeks. Such prediction if it was performed early enough could help to keep students in the course.

**Keywords** E-learning · MOOC · Data mining · Machine learning · Classification

## 1 Introduction

Nowadays, e-learning platforms, especially massive open online courses (MOOC) platforms, accumulate a huge amount of data about activities of course participants and the results of their training. This data could be analyzed to make online teaching more effective and interesting for students. Along with other methods, data mining could be used for that purpose. This led to the appearance of a special area in data mining—data mining of the educational process [1, 2]. Some examples of popular tasks for that sphere are clustering [3], student's performance prediction [4–6], and exploring the reasons for dropping the online course by students [7].

In our work, we analyzed grade reports of the MOOC “Data Management” from the Russian platform [openedu.ru](http://openedu.ru) [8]. The course was first time launched in September 2016. Sessions of the course started once a semester, in September and February, and for our analysis, we used results of five sessions of that course.

---

S. Nesterov (✉) · E. Smolina  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [nesterov@spbstu.ru](mailto:nesterov@spbstu.ru)

T. Egiazarov  
Rotterdam School of Management, Erasmus University, Rotterdam, Netherlands

The duration of the “Data Management” course is 16 weeks. Each week is devoted to a new topic, and each topic of the course contains video lectures, practical tasks, and a short test marked in the grade report as homework. Also, there are midterm and final exams. The final grade of the course is calculated from the average result of the weekly tests and results of these exams. More information about that course could be found in [9, 10]. It is important to note that in contrast to the works [11, 12], where the logs of the MOOC platform are analyzed, in our analysis we used only the standard grade reports from `openedu.ru`. Each instructor at `openedu.ru` can download such a report for his course, while the logs can be accessed only through technical support with a rather big delay. That is why we did not use logs in our analysis.

## 2 Method and Data Mining Tools

We used the R environment as the main software tool for our analysis. It has many different packages for statistical data processing, visualization, and machine learning [13–16]. In our work for classification, we used “`rpart`” (for decision trees), “`class`” (for the *k*-nearest neighbor algorithm), and “`Naïve Bayes`” packages.

Grade reports in `*.csv` format were downloaded from the “Open Education” portal (`openedu.ru`). Five sessions of the course we worked with were fall 2016, spring and fall 2017, and spring and fall 2018. Each report was imported into the R as a data frame. Missed data was relabeled to be in the readable format by the R environment.

As was mentioned before, one of the main problems of MOOCs is that only a small percent of students who enrolled in courses pass them through. So the prediction of the result of the course at least in the form of classification of students in two classes those who pass the course and those who drop out could be a very useful task.

For such type of classification, we needed a target attribute. We could not use the result of the final exam in this capacity because this exam is available only for students who paid for online proctoring: During the exam, a candidate is monitored online with the help of a webcam and special software, which gives access to the screen of the candidate. But for us, the results of all participants were interesting—and those who were on the free track of the course and those who were on paid track. So we decided to make synthetic target attribute: If the result of the student in a weekly test of the 15th or 16th week of the course is more than zero, we consider that this student passed the course (at least was learning all over the course). Otherwise, the student will be dropped out.

After performing the binary classification and analyzing its results, we tried to analyze the classification accuracy for different groups of students. These groups were determined based on the results of cluster analysis.

### 3 Binary Classification

As was mentioned before in our analysis, we used grade reports from MOOC platform *openedu.ru*. This platform is based on Open edX software which is widely used all over the world [17], and the format of its reports could be named rather common. For our course for each participant, we analyzed 16 marks for weekly tests and the mark for the midterm exam, which is placed after the material of the 8th week.

Using this data, we tried to predict if a course will be completed by a student or not according to his results during the few first weeks. For our prediction model, we decided to use the results of the first four sessions of the course as a training set and the data from the result report for the fall 2018 session—as a testing set. The next question was how many weeks of the course we must take into account. If this number will be too little, it will not give enough data for the relevant prediction. But if we will take all of the weeks, our prediction will be trivial.

First, we visualized data in our datasets. Figure 1 shows the number of students who performed tasks during the first session of the course (fall 2016). The picture was nearly the same for other sessions of the MOOC. A rather large number of students enroll for the course (from 1.5 to 3 thousands student) and only between 1/3 and 1/5 of them perform the first assignment. Then, the number of active students decreases weekly, and after the midterm exam, the number of such students becomes practically stable. That is why we decided that for non-trivial prediction will student drop out or not we must use the results not more than 8 weeks of the course. In more detail, these results are described in [9, 10].

For the training set which was combined from the grade reports from the first four sessions of the course, we calculated the number of students who completed the tasks of the first 8 weeks and the percent of students who passed the whole course among them. The results are summarized in Table 1.

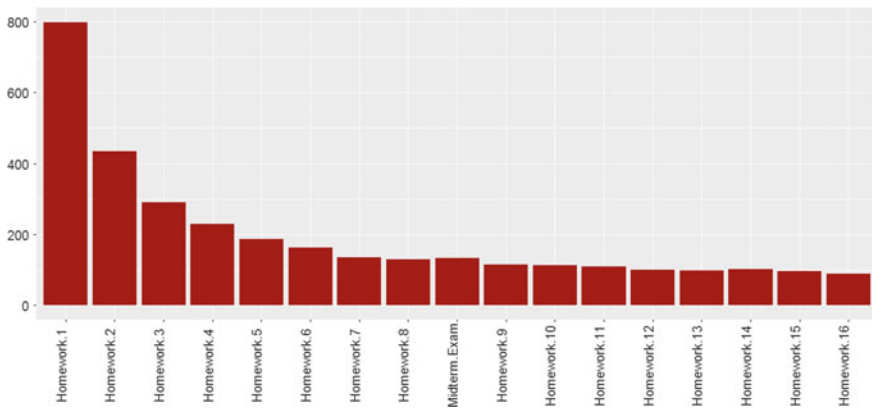


Fig. 1 Number of students who performed the task during fall 2016 session of the course



**Table 1** Number of students from training set who completed the tasks

Number of the task	Number of the students who passed	Percent of students who passed the whole course (%)
Homework 1	1869	22
Homework 2	1199	35
Homework 3	901	46
Homework 4	719	58
Homework 5	638	66
Homework 6	585	72
Homework 7	544	77
Homework 8	528	80
Midterm Exam	543	77
Passed the whole course	423	100

According to the results in Table 1, it could be mentioned that we could perform forecasting earlier than in the middle of the course. So we used data from first 4, 5, 6, 7, and 8 weeks to train classification models and used different data mining algorithms for them. The results are summarized in Table 2.

For binary classification after testing the classification model on data from the testing set, we can get four values:

**Table 2** Classification models

Algorithm	Number of weeks	accuracy	precision	recall	f1
k-nearest neighbor	4	0.79	0.75	0.73	0.74
	5	0.8	0.73	0.77	0.75
	6	0.81	0.75	0.77	0.76
	7	0.81	0.75	0.83	0.79
	8	0.82	0.76	0.81	0.79
Naïve Bayes	4	0.79	0.68	0.9	0.78
	5	0.82	0.71	0.94	0.81
	6	0.82	0.71	0.94	0.81
	7	0.84	0.73	0.95	0.82
	8	0.85	0.75	0.95	0.84
Decision Tree	4	0.79	0.74	0.73	0.73
	5	0.79	0.7	0.82	0.75
	6	0.83	0.78	0.79	0.78
	7	0.81	0.74	0.86	0.8
	8	0.84	0.77	0.87	0.82

- TP—number of true positive predictions;
- TN—number of true negative predictions;
- FP—number of false positive predictions;
- FN—number of false negative predictions.

Quality metrics that are mentioned in the table above are defined in the following way [14, 18]. Accuracy is the fraction of correct predictions:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Precision shows how accurate positive predictions were:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Recall is defined as the proportion of true positive predictions on the total number of positive instances:

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

*F1* score is the combination of precision and recall, which is defined as:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

After the analysis of accuracy and other quality metrics of our models, we tried to improve our classification models by training on the balanced training set. The problem was that most of the students in the original dataset have dropped out of the course. This is a common situation for MOOCs, but for our purposes, it was not particularly convenient.

So we used random undersampling [19] to form balanced training set where 40% of variants had target attribute with value 1 (passed), and 60%—0 (dropped out). But contrary to our expectations, training on such set did not help to improve the quality of our models.

## 4 Classification Accuracy for Different Groups of Students

In papers [9, 10], the results of cluster analysis of the same dataset are described. In current work, we will try to combine these results with classification. For clustering, we used the *k*-means algorithm and we chose the number of clusters based on the analysis of the sum of squared errors (distance between each point and the mean of its cluster) [14, 16]. For all course sessions, according to this indicator, four clusters

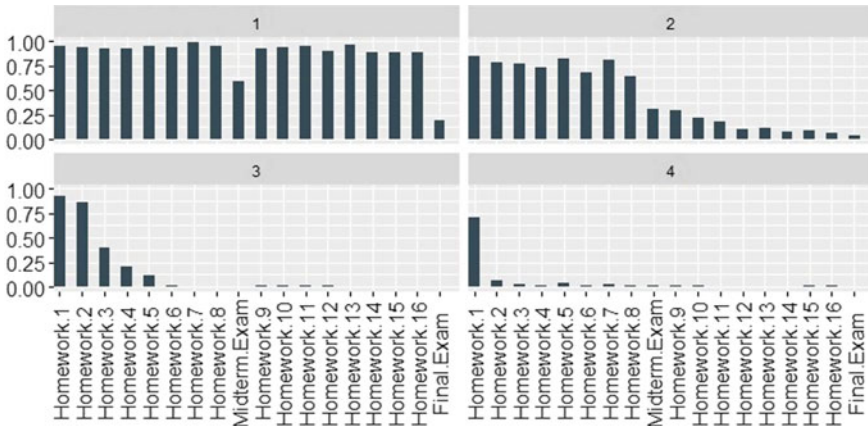


Fig. 2 Average results of weekly tests and the midterm exam for each cluster (spring 2018)

were chosen. Figure 2 shows the average results of weekly tests and the midterm exam for each cluster (spring 2018). For other sessions, the result looks like nearly the same.

As was mentioned before, in our analysis we used grade reports from MOOC platform openedu.ru.

- Students with good results over the course;
- Students who actively studied the first half of the course, and then lowered their academic performance, but completed the course;
- Students who were actively studying in the first weeks, then continued to study less actively, and dropped out after the middle of the course;
- Students who studied only about the first two weeks.

Algorithm	Class															
	1			2			3			4						
k-nearest neighbor	pred	fact			pred	fact			pred	fact			pred	fact		
		0	0	1		0	29	7		0	98	0		0	141	2
		1	11	158		1	40	8		1	0	0		1	2	1
Naïve Bayes	pred	fact			pred	fact			pred	fact			pred	fact		
		0	0	1		0	14	7		0	98	0		0	143	3
		1	11	189		1	55	8		1	0	0		1	0	0
Decision Tree	pred	fact			pred	fact			pred	fact			pred	fact		
		0	0	16		0	29	9		0	98	0		0	139	2
		1	11	174		1	40	6		1	0	0		1	4	1

Fig. 3 Confusion matrices for different groups (classes) of students

And for these groups of students, our prediction models show different accuracy. Figure 3 shows confusion matrices for prediction who will pass the course (target attribute value is 1) and who will drop out (target attribute value is 0). For different groups of students, our classifiers have different accuracy. In future work, in fact, it could be used to make predictions better.

## 5 Conclusion

E-learning and massive open online courses now become very popular. But the question is—are they effective? That is why different analysis methods are used to measure course effectiveness and help to rise it up [20, 21].

Now, the percent of participants who finish MOOCs is rather low. Most of the participants are dropped out. The analysis of the results of the MOOC studying helps to understand students and their reasons for leaving the course and maybe can help to keep them at the course.

An analysis similar to ours can be made for other courses of openedu.ru or similar platforms, which are based on Open edX software. Instructors of all courses have access to grade reports, and they can make an analysis themselves. For example, it could be done using some web-based instruments which could be developed for the exact MOOC platform and provided for all instructors who work with it.

During this work, we had access to the grade reports from different sessions of only one course—“Data management” [8]. Thus, it cannot be said that the results of the analysis of other courses will be similar. For example, we cannot say that cluster analysis will give the same four main groups of learners for other courses as it was on our course during all five sessions. This could be a subject of further research. And we believe that some patterns of learner's behavior may be the same for different courses. In [12], it was shown for activity log analysis of four different courses on the Coursera platform.

Also, we plan to combine the analysis of the grade reports of the course with the analysis of activity logs. Such logs can show how much time the student spent learning materials, how he worked with video and text materials, and so on. And it could be very useful for predicting a student's performance.

## References

1. Villanueva, A., Moreno, L.G., Salinas, M.J.: Data mining techniques applied in educational environments: literature review. *Digital Educ. Rev.* **33**, 235–266 (2018)
2. Algarni, A.: Data mining in education. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **7**(6) 2016. <https://doi.org/10.14569/IJACSA.2016.070659>
3. Crues, R.W., Henricks, G.M., Perry, M., Bhat, S., Anderson, C.J., Shaik, N., Angrave, L.: How do gender, learning goals, and forum participation predict persistence in a computer science

- MOOC? *ACM Trans. Comput. Educ.* **18**, 4, Article 18 (September 2018), 14 p. <https://doi.org/10.1145/3152892>
4. Sweeney, M., Lester, J., Rangwala, H., Johri, A.: Next-term student performance prediction: a recommender systems approach. *JEDM* **8**(1), 22–51 (2016)
  5. Liao, S.N., Zingaro, D., Thai, K., Alvarado, C., Griswold, W.G., Porter, L.: A robust machine learning technique to predict low-performing students. *ACM Trans. Comput. Educ.* **19**, 3, Article 18 (January 2019), 19 p. <https://doi.org/10.1145/3277569>
  6. Salal, Y.K., Abdullaev, S.M.: Using of data mining techniques to predict of student's performance in Industrial Institute of Al-Diwaniyah, Iraq. *Bull. South Ural State Univ. Ser. Comput. Technol. Autom. Control & Radioelectron.* **19**, 121–130 (2019). <https://doi.org/10.14529/ctcr190111>
  7. Yang, D., Kraut, R., Rose, C.: Exploring the effect of student confusion in massive open online courses. *JEDM* **8**(1), 52–83 (2016)
  8. Nesterov, S.A., Andreeva, N.V.: Data management (Massive open online course) <https://openedu.ru/course/spbstu/DATAM/>. Last accessed 19 Oct 2019 (in Russian)
  9. Nesterov, S.A., Smolina E.M.: Metody intellektual'nogo analiza dannyh v zadachah ocenki rezul'tatov distancionnogo obucheniya. [Methods of data mining in the analysis of the results of distance learning]. In: Proceedings of the XXIII International Scientific and Practical Conference "System analysis and control", pp. 406–412. Peter the Great St. Petersburg Polytechnic University. (2019). [https://elibrary.ru/download/elibrary\\_38582562\\_94678170.pdf](https://elibrary.ru/download/elibrary_38582562_94678170.pdf). Last accessed 19 Oct 2019 (in Russian)
  10. Nesterov, S.A., Smolina, E.M.: Some results of the analysis of 3 years of teaching of a massive open online course. In: *Cyber-Physical Systems and Control. Lecture Notes in Networks and Systems*, vol. 95. Springer International Publishing (2019). ISBN 978-3-030-34983-7
  11. Tabaa, Y., Medouri, A.: LASyM: a learning analytics system for MOOCs. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* **4**(5) (2013). <https://doi.org/10.14569/IJACSA.2013.040516>
  12. Gelman, B., Revelle, M., Domeniconi, C., Johri, A., Veeramachaneni, K.: Acting the same differently: a cross-course comparison of user behavior in MOOCs. In: Proceedings of the 9th international conference on educational data mining, EDM 2016, pp. 376–381 (2016). [https://www.educationaldatamining.org/EDM2016/proceedings/paper\\_136.pdf](https://www.educationaldatamining.org/EDM2016/proceedings/paper_136.pdf). Last accessed 1 Nov 2019
  13. Lantz, B.: *Machine Learning with R*, 2nd edn. Packt Publishing (2015). ISBN: 9781784393908
  14. Bruce, A., Bruce, P.: *Practical Statistics for Data Scientists*. O'Reilly Media (2017). ISBN: 9781491952955
  15. Grolemund, G., Wickham, H.: *R for Data Science*. <https://r4ds.had.co.nz>. Last accessed 19 Oct 2019
  16. Adler, J.: *R in a Nutshell: A Desktop Quick Reference*, 2nd edn. O'Reilly Media (2012)
  17. Open edX Homepage: <https://open.edx.org/>. Last accessed 19 Oct 2019
  18. Grus, J.: *Data Science from Scratch: First Principles with Python*. O'Reilly Media (2015). ISBN: 9781491901427
  19. Towards Data Science. Using Under-Sampling Techniques for Extremely Imbalanced Data. <https://towardsdatascience.com/sampling-techniques-for-extremely-imbalanced-data-part-i-under-sampling-a8dbc3d8d6d8>. Last accessed 19 Oct 2019
  20. Kalmykova, S.V., Pustyl'nik, P.N., Razinkina, E.M.: Role scientometric researches' results in management of forming the educational trajectories in the electronic educational environment. *Adv. Intell. Syst. Comput.* **545**, 427–432 (2017). [https://doi.org/10.1007/978-3-319-50340-0\\_37](https://doi.org/10.1007/978-3-319-50340-0_37)
  21. Surygin, A.I., Kalmykova, S.V., Alexankov, A.M.: Models of international virtual learning environment for international educational projects. In: 15th International Conference on Interactive Collaborative Learning, ICL 2012, 6402221 (2012). <https://doi.org/10.1109/ICL.2012.6402221>

# On the Implementation of the Planar3D Model Using the Explicit Time Integration Scheme and the Statistical Front Tracking Method



Egor Starobinskii , Nikita Mushchak , Svetlana Kraeva , Sergei Khlopin , and Egor Shel 

**Abstract** Nowadays, the relevance of problems associated with mathematical modelling of the propagation of hydraulic fractures is growing. In this paper, we consider a Planar3D plane crack model using the explicit numerical integration scheme over time to calculate the dynamics of crack propagation in a multilayer medium. When conducting numerical modelling, the key task is to ensure the speed of calculations while maintaining the stability of the calculation scheme. The present work is devoted to modelling the hydraulic fracturing of a layered formation under the influence of nonuniform injection of non-Newtonian fluid. We present the medium in the form of a set of horizontal layers, each of which is characterized by its own values of minimum compressive stress, toughness and leakage coefficient. To determine the geometry of the crack, we solve the front velocity equation and use universal asymptotes and the explicit time integration scheme.

**Keywords** Hydraulic fracturing · Planar3D model · Non-Newtonian fluid · Front tracking · Explicit time integration scheme · IMEX time integration scheme · Universal asymptotes

## 1 Introduction

Hydraulic fracturing of an oil-containing formation (HF) can significantly increase the efficiency of the oil production process by increasing the oil flow to the well. The complexity of conducting full-scale experiments determines the demand for computer models that predict the growth of a crack formed during hydraulic fracturing. The need for the development of modern approaches to hydraulic fracturing simulating is shown in [1–8]. The Planar3D model [9] considered in this paper

---

E. Starobinskii (✉) · N. Mushchak · S. Kraeva · S. Khlopin · E. Shel  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [st.eb@ailurus.ru](mailto:st.eb@ailurus.ru)

E. Shel  
Gazpromneft Science & Technology Centre, St. Petersburg, Russia

makes it possible to evaluate the geometry and dynamics of crack propagation in layered rock. To solve the system of differential equations of this model, implicit time integration schemes are traditionally used, which is primarily due to the speed of calculations. At the same time, the advantages of using an explicit scheme include the simplicity of taking into account a number of physical effects, the absence of the need for matrix inversion and the efficiency of parallel computing. This paper describes an algorithm for calculating the fracture geometry using the explicit or implicit time integration scheme.

## 2 Formulation of the Problem

Let us formulate a system of equations describing fluid transfer and medium deformation under the fluid injection from a point source. Let the formation consist of horizontal layers, isotropic and homogeneous in their mechanical properties. The layers are characterized by different values of minimum compressive stress, toughness and leak-off coefficient. A crack grows in a plane without lag (see Fig. 1), perpendicular to the interface of the layers, under the pressure of the injected non-Newtonian incompressible fluid (the rheology of the fluid is described by a power law).

The relationship between hydraulic fracture pressure and crack opening is determined by the hypersingular integral [11, 12]:

$$p(x, y, t) = \sigma(y) + \frac{E'}{8\pi} \iint \frac{w(x, y, t)}{\left( (x - \hat{x})^2 + (y - \hat{y})^2 \right)^{3/2}} d\hat{x}d\hat{y}, \quad (1)$$

where  $p(x, y, t)$  is the pressure at the point with coordinates  $(x, y)$  at time  $t$ ,  $\sigma$  are compressive reservoir stresses in the direction perpendicular to the plane of the crack,

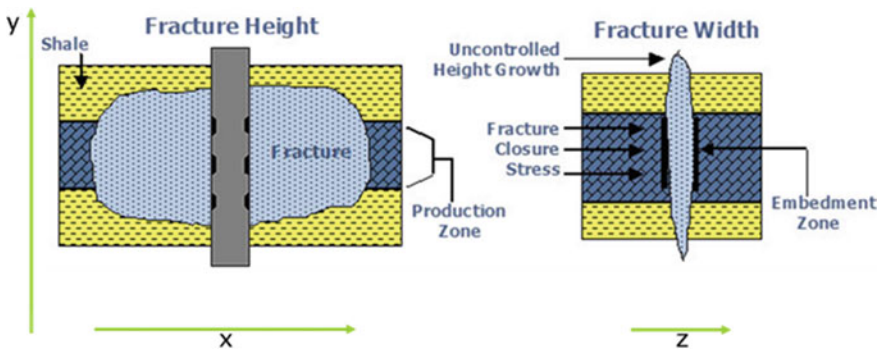


Fig. 1 Hydraulic fracture propagation schematic [10]

$E' = \frac{E}{1-\nu^2}$  is plane strain modulus,  $E$  is Young's modulus,  $\nu$  is Poisson's ratio,  $w$  is the crack opening.

The mass balance for the liquid, taking into account the injection of fluid and leaks according to the empirical law of Carter, we write in the following form:

$$\frac{\partial w(x, y, t)}{\partial t} = \nabla \cdot \left( \left( \frac{w^{2n+1}(x, y, t)}{\mu'} \right)^{\frac{1}{n}} |\nabla p(x, y, t)|^{\frac{1-n}{n}} \nabla p(x, y, t) \right) + Q(0, 0, t) - \frac{2C_l(y)}{\sqrt{t - t_0(x, y)}}, \tag{2}$$

where  $\frac{\partial w}{\partial t}$  is the partial time derivative,  $\nabla$  is del operator,  $n$  is the behaviour index,  $\mu' = 2\left(4 + \frac{2}{n}\right)^n \mu$  is the generalized consistency coefficient,  $\mu$  is the coefficient dynamic viscosity,  $Q(x, y, t)$  is the term describing the injection of fluid into the crack,  $C_l(y)$  is the leakage factor according to the Carter formula,  $t_0(x, y)$  is the activation time corresponding to the moment the crack front passes through the point with coordinates  $(x, y)$ .

Using the calculated magnitude of the increase in the fracture opening from the system of Eqs. (1)–(2), we can calculate the value of the opening at the next time step:

$$w(x, y, t + \Delta t) = w(x, y, t) + \int_t^{t+\Delta t} \frac{\partial w}{\partial t} dt, \tag{3}$$

where  $\Delta t$  is the time step.

To account for proppant transfer in the formulation given in [2], we will use the following form of the mass balance:

$$\frac{\partial(cw)}{\partial t} + \nabla \cdot (cw\mathbf{v}_p) = 0, \tag{4}$$

where  $c$  is the proppant concentration,  $\mathbf{v}_p$  is the proppant velocity.

We associate the effective viscosity of the liquid with the proppant concentration [6]:

$$\mu'_s = \mu' \left( 1 - \frac{c}{c_{\max}} \right)^{\frac{5n}{2}}, \tag{5}$$

where  $c_{\max}$  is the given maximum concentration for particular proppant used.

Thus, we get  $c$  from Eq. (4), substitute its value to formula (5) and then replace  $\mu'$  in Eq. (2) with  $\mu'_s$ . However, the discretization of Eq. (4) and the proppant velocity (namely, its settlement component) can be written in different forms. The subsequent work will be devoted to the study of this issue. In the framework of this article, we will



then focus on the simplified formulation of the problem in the absence of proppant injection into the fracture.

### 3 Numerical Integration Scheme and Discretization

The considered system of equations can be solved by both explicit and implicit methods. The advantages of the explicit scheme include the simplicity of the algorithm, the convenience of parallel computing and the absence of the need for matrix inversion. A significant drawback is the need to choose a small (compared to implicit schemes) time step to ensure the stability of the scheme.

The possibility of using an explicit calculation scheme was shown in [13]. Moreover, the use of the explicit scheme is convenient for several reasons. Since the use of the initial approximation is not required, the resulting solution can be considered more accurate. In addition, the use of a small time step allows us to consider the problem of finding a crack opening by a fluid pressure one-way connected; that is, according to the calculated opening, it is not necessary to adjust the pressure values.

Hereinafter, the numerical approximation is achieved by the finite difference method. Thus, a new opening value at each time step can be found as follows:

$$w_{i,j}^{k+1} = w_{i,j}^k + \frac{\partial w_{i,j}^k}{\partial t} \Delta t, \quad (6)$$

where  $w_{i,j}^k$  is the opening in the cell  $\{i, j\}$  at the  $k$ -th time step,  $\frac{\partial w_{i,j}^k}{\partial t} = \left(\frac{\partial w_{i,j}}{\partial t}\right)^k$  is obtained from Eq. (2).

In addition to the explicit integration scheme by the Euler method, we will use spatial discretization with a fixed cell size ( $\Delta x, \Delta y = \Delta x$ ). Odd numbers horizontally and vertically give the number of cells of the computational grid. In this case, it is convenient to represent the pressure and opening in the form of column vectors using the procedure of drawing the corresponding matrices [9]. The hypersingular integral from Eq. (1) can be replaced by the product of the influence matrix  $A_{i,j}$  and the opening column  $w_{i,j}$ :

$$p_{i,j}(t) = \sigma_j + \frac{E'}{8\pi} A_{i,j} \cdot w_{i,j}(t), \quad (7)$$

where  $p_{i,j} = p(\Delta x(i - i_0), \Delta y(j - j_0))$  and  $w_{i,j} = w(\Delta x(i - i_0), \Delta y(j - j_0))$ ,  $\sigma_j = \sigma(\Delta y(j - j_0))$ ,  $i_0$  is the index of the central column of the computational mesh,  $j_0$  is the index of the central row,  $\{i_0, j_0\}$  are the discrete coordinates of the central cell.

The values of the influence matrix can be found as a solution of the double integral:

$$A_{i,j} = \int_{y_{i,j}-\frac{\Delta y}{2}}^{y_{i,j}+\frac{\Delta y}{2}} \int_{x_{i,j}-\frac{\Delta x}{2}}^{x_{i,j}+\frac{\Delta x}{2}} \left( (x_{i,j} - \hat{x})^2 + (y_{i,j} - \hat{y})^2 \right)^{-3/2} d\hat{x}d\hat{y}, \quad (8)$$

where  $(x_{i,j}, y_{i,j})$  are the coordinates of the centre of the cell  $\{i, j\}$ .

Let us write the discrete form of Eq. (2):

$$\frac{\partial w_{i,j}}{\partial t} = \frac{1}{\Delta x} (s_{i,j}^{i+1} + s_{i,j}^{i-1}) + \frac{1}{\Delta y} (s_{i,j}^{j+1} + s_{i,j}^{j-1}) - \frac{2(C_l)_j}{\sqrt{t - (t_0)_{i,j}}}, \quad (9)$$

$$s_{i,j}^{\pm 1} = \text{sgn}(p_{i\pm 1,j} - p_{i,j}) \left( \frac{w_{i\pm 1,j} + w_{i,j}}{2} \right)^{\frac{2n-1}{n}} \left| \frac{p_{i\pm 1,j} - p_{i,j}}{\mu' \cdot x} \right|^{\frac{1}{n}}, \quad (10)$$

$$s_{i,j}^{j\pm 1} = \text{sgn}(p_{i,j\pm 1} - p_{i,j}) \left( \frac{w_{i,j\pm 1} + w_{i,j}}{2} \right)^{\frac{2n-1}{n}} \left| \frac{p_{i,j\pm 1} - p_{i,j}}{\mu' \cdot y} \right|^{\frac{1}{n}}, \quad (11)$$

where  $\text{sgn}(x)$  is the sign function. If  $i = i_0$  and  $j = j_0$  we also add source pump value to the time derivative of the opening in this cell.

A different approach to integrating is to use a hybrid implicit-explicit scheme (IMEX). Equation (3) in this case can be rewritten by adding the resida, in the form of the Laplacian of the opening to the time integral:

$$w(t + dt) = w(t) + \int_t^{t+dt} \left( \frac{\partial w}{\partial t} \pm (\nabla \cdot \nabla)(Dw) \right) dt, \quad (12)$$

where  $D$  is a small non-negative coefficient that determines the effect of the residual. For  $D = 0$ , the scheme becomes explicit.

The hybrid scheme is more stable than the explicit scheme, which allows us to increase the time step by about 400%. In this case, the time step remains small in comparison with a purely implicit scheme. Let us write the calculation scheme for Eq. (12), introducing the residual with a positive sign at  $k + 1$  time step and with a negative one at step  $k$ :

$$\begin{aligned} & w_{i,j}^{k+1} \left( \frac{\Delta x \Delta y}{D \Delta t} + 4 \right) - \left( w_{i+1,j}^{k+1} + w_{i-1,j}^{k+1} + w_{i,j+1}^{k+1} + w_{i,j-1}^{k+1} \right) \\ &= w_{i,j}^k \left( \frac{\Delta x \Delta y}{D \Delta t} + 4 \right) + \frac{\Delta x \Delta y}{D} \frac{\partial w_{i,j}^k}{\partial t} - \left( w_{i+1,j}^k + w_{i-1,j}^k + w_{i,j+1}^k + w_{i,j-1}^k \right). \end{aligned} \quad (13)$$

Here we used the assumption that  $(\nabla \cdot \nabla)w_{i,j}^k = \frac{w_{i+1,j}^k - 2w_{i,j}^k + w_{i-1,j}^k}{(\Delta x)^2} + \frac{w_{i,j+1}^k - 2w_{i,j}^k + w_{i,j-1}^k}{(\Delta y)^2}$  and  $\Delta x = \Delta y$ .

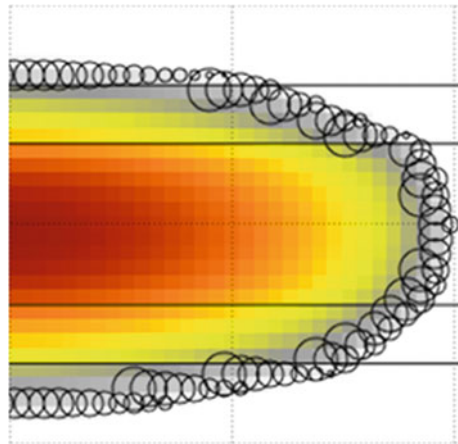
The resulting system of linear algebraic Eqs. (13) can be solved by the conjugate gradient method, since the matrix of SLAE coefficients turns out to be five-diagonal. The procedure for finding  $\frac{\partial w}{\partial t}$  remains unchanged.

Thus, Eqs. (7)–(11) in combination with Eq. (6) or (13) form the dynamic system that describes the change in the geometry of the crack. To speed up the calculations and to clarify the position of the crack front, we will use universal asymptote, more on this in Parts 4 and 5. For each layer in the reservoir, we consider the known values of minimum compressive stress  $\sigma$ , plane strain modulus  $E'$ , leakage coefficient  $C_l$  and toughness  $K_{Ic}$  (will be used to write down the asymptotic formula). At the initial moment, it is necessary to know the opening at each point and the position of the crack front. The boundaries of the computational domain are assumed to be free. A more detailed discussion of the initial and boundary conditions is given in Part 7.

## 4 Front Tracking

Those cells of the computational grid, to which the crack front has reached, are called active elements. Additionally, we divide the active elements into three subtypes: tip, ribbon and channel. The tip elements are those cells that contain the front of the crack. If the cell has a common vertex with the tip element, but itself is located behind the front of the crack, then such a cell will be called the ribbon cell. All other activated cells will be considered internal (channel). To track the crack front, we use the statistical method that does not require finding the norm to the front line. The basis of the method is to build tracking circles around the ribbon elements (see Fig. 2). The radius of the circle built around the element is equal to the distance from the centre of the element to the front of the crack. When the tip element is completely inside the circle built around one of the ribbon elements, it itself becomes a ribbon element. We also monitor the moment of changing the type of an element

**Fig. 2** Crack opening (right wing). Tracking circles are indicated in black. Horizontal lines indicate layer surfaces



from a ribbon to a channel by the radius of the circle: when the circle constructed from the centre of the ribbon element completely includes the diagonally closest elements, this ribbon element becomes channel. The advantage of this approach is the computational speed, since the determination of new ribbon elements occurs according to the simple geometric criterion: the distance to the front in the ribbon element must belong to  $\left[ \frac{\sqrt{2}}{2} \Delta x, \frac{3\sqrt{2}}{2} \Delta x \right]$ . Moreover, if the distance is greater than  $\frac{\sqrt{10}}{2} \Delta x$ , then adjacent cells vertically and horizontally can also become ribbons. To calculate the distance to the front at the next time steps after the first step, it is convenient to use universal asymptotes [14–18] (at the first time step, we consider the position of the front known).

## 5 Universal Asymptotic Umbrella

To determine the new distance to the front in each boundary element from the known value of the opening in the cell, we use the universal asymptotic umbrella (UAU) described in [17, 18]. The general form of the asymptote connects the distance to the crack front with its opening:

$$w = A_w(v)r^\alpha, \tag{14}$$

where  $v$  is the front velocity,  $A_w(v)$  is the known function of the front velocity,  $r$  is the distance to the front, and  $\alpha$  is the quantity depending on the rheology of the fluid.

The value of  $\alpha$  and the form of the function  $A_w$  are determined by the degree in the rheological law of the fluid and the mode of crack propagation. UAU is described for three propagation modes: the dominant leak-off regime, the dominant viscosity regime and the dominant toughness regime. Let us introduce the auxiliary function  $B(\alpha) = \frac{\alpha}{4} \cot(\pi - \pi\alpha)$  and write the UAU for each of the three regimes (we assume that Carter’s law always holds for leaks):

- dominant leak-off regime:

$$\alpha = \frac{n + 4}{4n + 4}, A_w(v) = A_l v^{\frac{1-\alpha}{3}}, A_l = \left( 2^n C_l^2 \frac{\mu'}{E'} \right)^{\frac{1}{2n+2}} \left[ \frac{3n}{4n + 4} B\left( \frac{n + 4}{4n + 4} \right) \right]^{-\frac{1}{2n+2}}, \tag{15}$$

where  $C_l$  is the leakage coefficient for the considered layer;

- dominant viscosity regime:

$$\alpha = \frac{2}{n+2}, A_w(v) = A_v v^{1-\alpha}, A_v = \left(\frac{\mu'}{E'}\right)^{\frac{1}{n+2}} \left[ \frac{n}{n+2} \mathbf{B}\left(\frac{2}{n+2}\right) \right]^{-\frac{1}{2n+2}}, \quad (16)$$

– dominant toughness regime:

$$\alpha = \frac{1}{2}, A_w(v) = A_t = \sqrt{\frac{32}{\pi}} \frac{K_{Ic}}{E'}, \quad (17)$$

where  $K_{Ic}$  is the critical value of normal stress intensity factor.

Let crack opening  $w$  and normalized opening  $\tilde{w}$  be connected by the following relation:

$$w = \left(\frac{\mu'}{E' t_s^n}\right)^{\frac{1}{n+2}} \tilde{w}, \quad (18)$$

where  $t_s$  is the scale factor equal to the number of seconds per unit time.

Then, UAU for the normalized opening  $\tilde{w}$  can be written in the following form:

– dominant leak-off regime:

$$\tilde{w} = \left[ \frac{3n}{4n+4} \mathbf{B}\left(\frac{n+4}{4n+4}\right) (4C_l \sqrt{t_s})^{-n} \left(\frac{\mu'}{E' t_s^n}\right)^{\frac{n}{n+2}} \right]^{-\frac{1}{2n+2}} (t_s v)^{\frac{n}{4n+4}} r^{\frac{n+4}{4n+4}}, \quad (19)$$

– dominant viscosity regime:

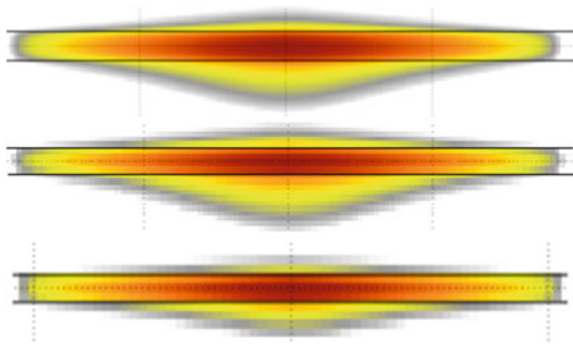
$$\tilde{w} = \left[ \frac{n}{n+2} \mathbf{B}\left(\frac{2}{n+2}\right) \right]^{-\frac{1}{n+2}} (t_s v)^{\frac{n}{n+2}} r^{\frac{2}{n+2}}, \quad (20)$$

– dominant toughness regime:

$$\tilde{w} = \sqrt{\frac{32}{\pi}} K_{Ic} \left[ \frac{\mu'}{t_s^n} (E')^{n+1} \right]^{-\frac{1}{n+2}} r^{\frac{1}{2}}. \quad (21)$$

Note that the described approach allows us to specify in each cell any of the three distribution regimes. It also seems possible to generalize the method for simultaneously taking into account two propagation modes in one cell: in the horizontal and in the vertical directions. In this case, instead of tracking circles, it is worth using tracking ellipses, the half-axes of which will be set by universal umbrellas.

**Fig. 3** Crack opening during calculation with a fixed cell size (top), with a single scaling (middle) and double scaling (bottom). Horizontal lines indicate layer surfaces



## 6 Mesh Scaling

To increase the speed of calculations and maintain the stability of the calculation scheme, the use of automatic mesh scaling is proposed. For example, if the crack size is doubled relative to the original (or when the crack front reaches the boundary of the computational domain), the size of each cell can also be doubled. This leads to a multiple acceleration of calculations, since the number of active elements nonlinearly grows relative to the growth of the crack itself. Figure 3 shows the example of calculating a hydraulic fracture in a three-layer medium for three cases: when the cell size was kept constant, when the cell size was increased once, and when the cell size was increased two times. The differences in the opening value at the injection point were less than 0.5%, and the differences in determining the ratio of the crack length to its height were less than 3% and less than 5%, respectively, for single and double scaling of meshes. The compressive stresses in the upper layer were 4 MPa more than in the central one, and in the lower layer 3 MPa more than in the central layer. Newtonian fluid ( $n = 1$ ) was injected with the dynamic viscosity of  $\mu = 0.4$  Pa s for 1 h with the variable flow rate of 3.2–3.6 m<sup>3</sup>/min. The initial cell size was 2.7 × 2.7 m, after double scaling becoming 10.8 × 10.8 m. A single scaling led to the acceleration of calculations by 15,000%, and a double scaling led to the additional acceleration by another 1200%.

## 7 Boundary and Initial Conditions

To specify the initial crack configuration, one needs to know the opening in each cell of the computational mesh. One will also need to determine which elements will be ribbon, as well as the distance to the front in these elements. The crack initiation process is not considered in this work, although such a problem can be simplified as the growth of a crack consisting of only one element with the radius of the tracking circle defined in it at the initial moment of time. There are special models that can more accurately calculate the process of crack initiation in a layered medium. In the

framework of this work, we will use a simplified approach; in particular, we will not take into account the effect of perforation in the injection pipe on the shape of the formed crack, since the influence of this factor on the final shape of the crack is small [17].

We determine the initial state of the crack from the self-similar solution [17] for the radial crack formed by the injection of a viscous fluid. In [17], the parameters of a self-similar solution for hydraulic fracturing are presented for fluids with various  $n$ , namely the radius of the crack  $r_A$  and the crack opening  $w_A$  as a function of distance from the source  $d_n (0 \leq d_n \leq r_A)$ . We write the initial crack opening at time  $t_0$  with constant injection with a flow rate of  $Q_0$ :

$$w(d) = w_A \left( \frac{r_A}{r_0} d \right) \left( \left( \frac{\mu'}{E'} \right)^2 t_0 \right)^{1/9} Q_0^{1/3}, \quad (22)$$

where  $0 \leq d \leq r_0$ ,  $r_0$  is the distance from the injection source to the front of the radial crack at the initial moment of time.

Since a self-similar solution implies orthotropic propagation of a crack from an injection source, a natural time limit appears in a layered medium before which such a solution can be used. The time  $t_0$  is associated with the initial crack radius  $r_0$  by the relation:

$$t_0 = \left( \left( \frac{r_0}{r_A} \right)^3 \frac{\mu'}{Q_0 E'} \right)^{3/4}. \quad (23)$$

The crack radius  $r_0$  is limited by the dimensions of the central layer (the layer containing the injection source). On the other hand, during time  $t_0$ , the fluid flow rate must remain constant (or close to constant). To satisfy both conditions, the time  $t_0$  will be chosen as the minimum value between the time to reach the surface of the central layer and the moment of change in the flow rate of the fracturing fluid. It is worth noting that in the practical application of the described model in hydraulic fracturing simulators, the initial crack state for Planar3D can be obtained from the corresponding calculation modules.

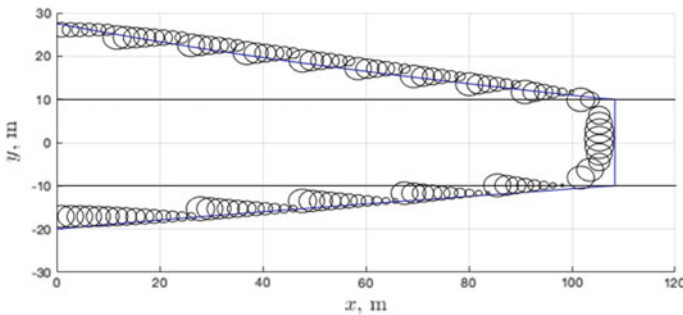
The boundary condition may be the requirement of symmetry of the left and right wings of the crack. This assumption is true if multistage hydraulic fracturing is not considered with the simultaneous development of nearby fractures and their mutual influence [19]. We consider two options for applying the symmetry condition to discretized equations. The first way is to save the central column of the mesh and modify the calculation of the elements of the influence matrix (8) to take into account the reflection  $f_{i,j} = f_{2i_0-i,j}$ . The source is then shifted by  $\frac{\Delta x}{2}$  relative to the axis of symmetry, and the fluid flows to the right from the cells of the column containing the source will be equal to the flows to the left, taken with the opposite sign. Another option is to rebuild the mesh so that the source is located strictly on the axis of symmetry and has half the initial power, and the flows through this boundary with

the condition of symmetry are zero. Both methods have proven themselves in practice and have led to a significant increase in the speed of calculations.

## 8 Research Results and Discussion

The described algorithm was implemented as the calculation module in C++. The program allows you to track over time the progress of hydraulic fracturing in a layered medium, the opening of the fracture and the pressure of the fracturing fluid, as well as the concentration of proppant. The input parameters are lithology data, as well as fluid and proppant injection plans.

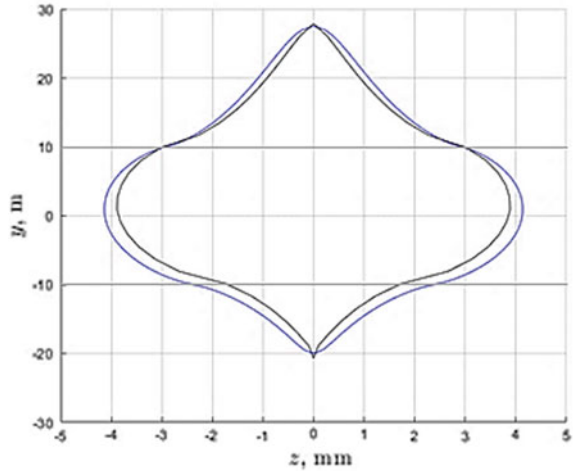
As an example of calculations, we consider a three-layer medium with a thickness of the central layer of 20 m and a contrast of stresses relative to the central layer of + 4 MPa for the upper half-space and +5 MPa for the lower half-space. We will pump into the centre of the layer a Newtonian fluid with characteristics  $n = 1$ ,  $\mu = 0.2 \text{ Pa} \cdot \text{s}$  at a constant flow rate  $Q = 3.6 \text{ m}^3/\text{min}$ . Medium characteristics:  $E = 2.5 \text{ GPa}$ ,  $\nu = 0.23$ ,  $C_l = 0$ . Simulation time is 439 s. Figures 4 and 5 show the crack profiles obtained in the Planar3D calculation module and the module that implements the Pseudo3D approach from [20]. The difference in determining the crack length was less than 1% and in determining the height is less than 0.5%. Additionally, in [9], simulation results in the developed software are compared with published calculations for EP3D and Planar3D ILSA [21], and the good agreement between the results was also obtained.



**Fig. 4** Front of the right wing of the hydraulic fracture. The black circles show the tracking circles for the Planar3D model; the blue colour shows the front line according to the Pseudo3D model. Solid horizontal lines mark layer boundaries



**Fig. 5** Hydraulic fracture opening profiles calculated using the Planar3D (black line) and Pseudo3D (blue line) models. Solid horizontal lines mark layer boundaries



## 9 Conclusions

This paper describes a planar crack propagation model in a layered medium based on the Planar3D method. Minimum compressive stress, toughness and leak-off coefficient are considered different in different layers. The key features of the proposed model are reducing the system of partial differential equations to a dynamical system, using universal asymptotes to describe various regimes of crack propagation under the influence of a non-Newtonian fluid, and determining the position of the crack front without calculating the norm. We also propose the use of the hybrid implicit–explicit integration scheme to speed up calculations.

**Acknowledgements** This work was supported by Ministry of Science and Higher Education of the Russian Federation within the framework of the Federal Program “Research and development in priority areas for the development of the scientific and technological complex of Russia for 2014–2020” (activity 1.2), Grant No. 14.575.21.0146 of September 26, 2017, unique identifier: RFMEFI57517X0146. The industrial partner of the grant is LLC “Gazpromneft Science & Technology Centre”.

The authors are deeply grateful to A. M. Linkov, V. A. Kuzkin, A. D. Stepanov for useful discussions.

## References

1. Pitakbunkate, T., et al.: Hydraulic fracture optimization with a P-3D model. In: SPE Production and Operations Symposium/Society of Petroleum Engineers
2. Adachi, J., et al.: Computer simulation of hydraulic fractures. *Int. J. Rock Mech. Mining Sci.* **44**, 739–757 (2007)

3. Peirce, A.: Implicit level set algorithms for modelling hydraulic fracture propagation. *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **374**, 20150423
4. Mack, M.G., Warpinski, N.R.: Mechanics of hydraulic fracturing. *Reservoir Stimulation*, 6–1
5. Linkov, A.M., et al.: Modified formulation,  $\varepsilon$ -regularization and the efficient solution of hydraulic fracture problems. In: ISRM International Conf. for Effective and Sustainable Hydraulic Fracturing/International Society for Rock Mechanics and Rock Engineering (2017)
6. Osipov, A.A.: Fluid mechanics of hydraulic fracturing: a review. *J. Petrol. Sci. Eng.* **156**, 513–535 (2017)
7. Khasanov, M.M., et al.: Scientific engineering as the basis of modeling processes in field development. *Georesources* **20**, 142–148 (2018)
8. Zia, H., Lecampion, B.: PyFrac: A planar 3D hydraulic fracture simulator (2019). arXiv preprint 1908.10788
9. Starobinskii, E.B., Stepanov, A.D.: Adapting the explicit time integration scheme for modeling of the hydraulic fracturing within the planar3D approach. *J. Phys. Conf. Ser.* **1236**(1), 012052
10. Core Lab: Hydraulic fracture design. <https://corelab.com/ps/hydraulic-fracture-design>
11. Peirce, A.: Modeling multi-scale processes in hydraulic fracture propagation using the implicit level set algorithm. *Comput. Methods Appl. Mech. Eng.* **283**, 881–908 (2015)
12. Hills, D.A., et al.: *Solution of Crack Problems: The Distributed Dislocation Technique*, vol. 44. Springer Science & Business Media
13. Stepanov, A.D., Linkov, A.M.: On increasing efficiency of hydraulic fracture simulation by using dynamic approach of modified theory. In: *Proceedings of Summer School-Conference on Advanced Problems in Mechanics*, pp. 393–403 (2016)
14. Dontsov, E.V., et al.: Implementing a universal tip asymptotic solution into an implicit level set algorithm (ILSA) for multiple parallel hydraulic fractures. In: *50th US Rock Mechanics/Geomechanics Symposium/American Rock Mechanics Association* (2016)
15. Peirce, A., Detournay, E.: An implicit level set method for modeling hydraulically driven fractures. *Comput. Methods Appl. Mech. Eng.* **197**, 2858–2885 (2008)
16. Garagash, D.I., Detournay, E., Adachi, J.I.: Multiscale tip asymptotics in hydraulic fracture with leak-off. *J. Fluid Mech.* **669**, 260–297 (2011)
17. Linkov, A.M.: The particle velocity, speed equation and universal asymptotics for the efficient modelling of hydraulic fractures. *J. Appl. Math. Mech.* **79**, 54–63 (2015)
18. Linkov, A.M.: Universal asymptotic umbrella for hydraulic fracture modeling (2014). arXiv preprint) 1404.4165
19. Cherevko, M.A., et al.: *Development of the oil fields in Western Siberia using the horizontal wells with a multistage hydraulic fracturing (in Russian)*. Tyumen'–Kurgan: Zaural'e, p. 265 (2015)
20. Markov, N.S., Linkov, A.M.: correspondence principle for simulation hydraulic fractures by using pseudo 3D model. *Mater. Phys. Mech.* **40**, 181–186 (2018)
21. Dontsov, E.V., Peirce, A.P.: An enhanced pseudo-3D model for hydraulic fracturing accounting for viscous height growth, non-local elasticity, and lateral toughness. *Eng. Fracture Mech.* **142**, 116–139 (2015)

# Fast Fourier Transform in Planar3D Model Using an Explicit Numerical Integration Scheme



Nikita Mushchak , Egor Starobinskii , Sergei Hlopin , and Egor Shel 

**Abstract** In this paper, the authors propose methods to speed up calculations of a fracture propagation model using fast Fourier transform (FFT). We consider the Planar3D model with the explicit numerical integration scheme. Current research decided to implement the radix-2 Cooley–Tukey FFT algorithm in C++ code using STL containers, which provided fast calculations and gave advances to work with memory and cache. We compare the speed of FFT computation with other libraries (FFTW3, GSL, Eigen3). Analysis of results has been shown as a comparable time of calculations. We consider a method for accelerating the calculations of the Planar3D module in the framework of matrix–vector multiplication and processing of input data using a low-pass filter. The considerate model uses the product of matrix–vector multiplication. This procedure engages from time to time throughout the program. In this paper, we implement a modified method for calculating the matrix–vector multiplication product using FFT, which allows us to speed up the calculations. Another technique is used to process input data by the example of averaging lithology layers. In the presence of thin layers with high contrasts of mechanical properties, one can apply a low-pass filter. Such processed layers make it possible to obtain an increase in the computational speed when simulating the evolution of a hydraulic fracture model.

**Keywords** Fast Fourier Transform · Planar3D model · Low-pass filter · Explicit numerical integration scheme

---

N. Mushchak (✉) · E. Starobinskii · S. Hlopin · E. Shel  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [mutchak\\_nd@spbstu.ru](mailto:mutchak_nd@spbstu.ru)

E. Shel  
LLC Gazpromneft Science and Technology Centre, St. Petersburg, Russia

## 1 Introduction

The first acceleration method of discrete Fourier transform (DFT) was proposed by Gauss in the nineteenth century [1]. Fourier transform is widely used in engineering, sound recording and sound transmission, and signal processing. One of the easiest FFT implementations was proposed by Cooley–Tukey [2]. This method requires the length of the input array to be a power of two, which represents the divide and conquer algorithm. It is also worth noting the prime factor algorithm [3], which generalizes a one-dimensional DFT to a two-dimensional DFT. Bruun’s FFT algorithm [4] uses the recursive polynomial-factorization approach. The Rader’s FFT algorithm [5] calculates the discrete Fourier transform for the lengths of arrays expressed by primes by repeatedly expressing the DFT as a cyclic convolution.

For frequency-domain signals, FFT can also be applied in real time [6, 7], if the signal delay exceeds the time required to calculate the Fourier transform. The asymptotic complexity of FFT is  $O(N \log(N))$ , where  $N$  is the number of elements in the signal sample. FFT is used for low-pass filtering [8], high-pass filtering [9], and band-pass filtering [10].

Fast Fourier transform allows us to speed up bottleneck fragments of the Planar3D calculation module. Within the framework of this study, we realize the single-header library of FFT [11] available in the open Git repository. We describe the matrix–vector product acceleration using FFT, as well as a technic of processing lithology data via low-pass filter also based on FFT.

There are some well-known libraries that have FFT implementations. Nowadays, acceleration of computation speed is achieved through the use of specific architectures [12, 13] and the use of parallel computing [14, 15]. In this work, we compare the speed of calculation of different libraries with C++ implementations that can be used with popular compilers like GCC, g++, MVC, clang, and clang++. Besides our FFT library, we also consider FFTW3 [16], GSL [17], and Eigen3 [18]. Libraries have proven themselves for their use in scientific research. FFTW3 uses several variants of the Cooley–Tukey FFT algorithm [2], as well as the Bluestein’s FFT algorithm [19]. The GSL library uses radix-2 and mixed-radix algorithms. Eigen takes advantage of processor architecture features.

## 2 Fast Fourier Transform

There are various FFT algorithms; one of the easiest to implement is the radix-2 Cooley–Tukey algorithm [2]. This algorithm recursively divides the sequence into two equal parts and then applies the discrete Fourier transform:

$$F_k = \sum_{m=0}^{N/2-1} f_{2m} e^{-2i\pi k(2m)/N} + \sum_{m=0}^{N/2-1} f_{2m+1} e^{-2i\pi k(2m+1)/N}, \quad (1)$$

which is the separate DFT for even indices  $2m$  and for odd indices  $2m + 1$ .

Inverse Fourier transform consists of the following steps: complex conjugation of the input vector, direct Fourier transform, complex conjugation of the result, and normalization to satisfy

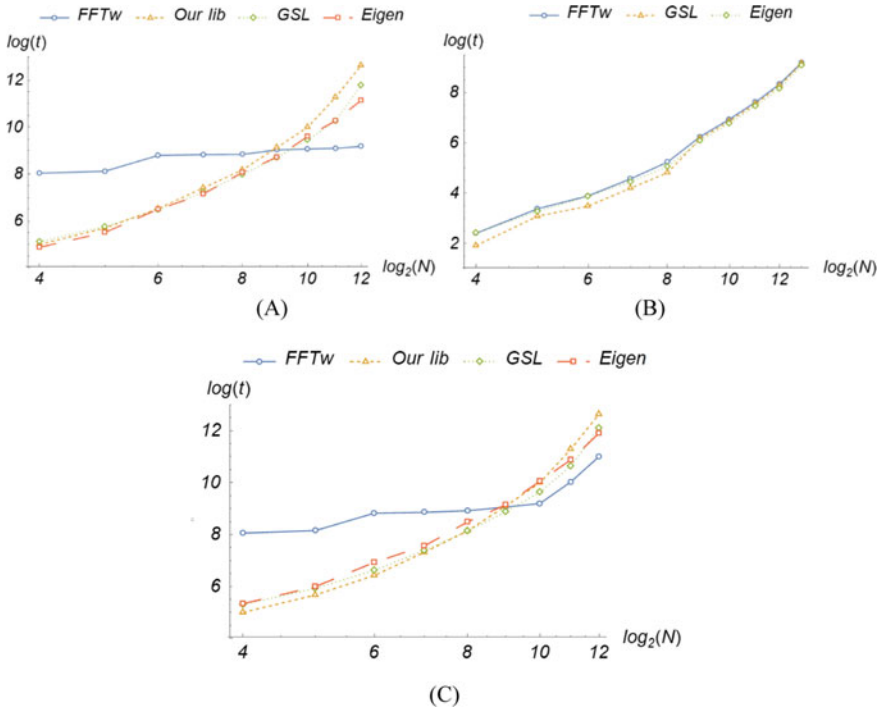
$$y = \text{IFFT}(\text{FFT}(f)) \quad (2)$$

While implementing the FFT in the library, we took into account state-of-the-art programming methods. The library uses STL containers, which are a convenient and efficient tool for programming in C++. The main advantages of containers are automatic allocation, memory clearing, and cache monitoring. It is most convenient to work with the STL container `std::vector()`, which implements quick operations of adding an element to the vector end and accessing its element [20]. To store complex numbers, a vector of vectors is used, with each element of the original vector containing a two-element vector to store the real and imaginary parts. That said, the input argument to the FFT function is a complex vector, the values of which during the calculation are replaced by the values of its Fourier image.

When testing the library, the calculation speed was compared with FFTW3, GSL, and Eigen3. The same randomly generated vector was sent as an input to all tested functions. The double data type was used to achieve acceptable accuracy for scientific calculations, and the execution times were obtained using the standard method `std::chrono()`. Fourier transform was calculated 1 million times, after which the average time of calculation was found (Fig. 1a).

Libraries of other developers use different types of data, so conversion is required. Conversion was implemented as a sequential value copying from the generated vector into the data type that each tested library works with; this operation was also performed 1 million times (Fig. 1b). Additionally, collected calculation times with data conversion and subsequent FFT calculations of the tested libraries were checked (Fig. 1c).

Due to the data that we received, it can be noted that our library has the same time with the other libraries at the length of vector less than 256. With the length of the vector, more than 512, our library has up to 5 times slower calculations (Fig. 1a). Conversion from `std::vector()` to data that FFTW library uses takes more time than conversion to other library types of data (up to 15% with length of vector 4096) (Fig. 1b). Our library has an advance in speed of calculation of FFT with data conversion with the vector length less than 512 and 6 times slower with vector length 4096. FFTw, GSL, and EIGEN libraries have more complicated algorithms than our library. It is possible to accelerate the algorithm adding the implementation of radix-4 or more complicated computational schemes.



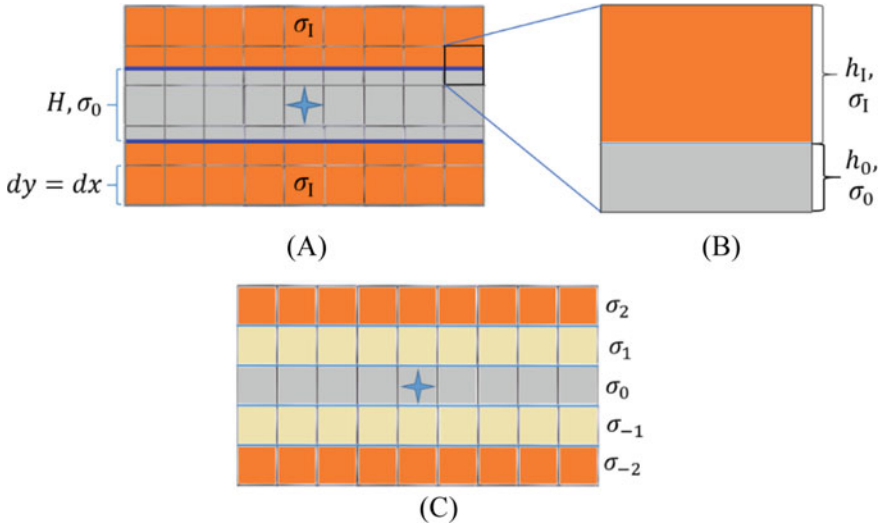
**Fig. 1** Times of calculation of FFT using different libraries (1 million times averaging) (a). Time of conversion to libraries' data (1 million times averaging) (b). Time of conversion and FFT (1 million times averaging) (c)

### 2.1 Layers Approximation Using FFT

The Planar3D [21–25] is a model of a plane fracture propagating in a multi-layered medium. The assumption that the rock formation consists of homogeneous isotropic horizontal layers is used. The fracture propagates in a plane perpendicular to the minimum compressive stresses.

Analysis of geological data means obtaining necessary information about the mechanical properties of the layers. One of the problems is to take into account the Planar3D [23] model layers with thickness of less than one mesh element and layers for which its boundaries do not coincide with the boundaries of the cells. In this case, it is necessary to interpolate the mechanical properties of the current layer on the computational mesh. As an illustration, we consider a three-layer medium ( $H = 10\text{m}$ ,  $\sigma_0 = 0\text{MPa}$ ,  $\sigma_1 = 6\text{MPa}$ ,  $dy = dx = 6\text{m}$ ) (Fig. 2a). The stresses are found as the weighted arithmetic mean (Fig. 2b):

$$\sigma_k = \frac{\sigma_l \cdot h_l + \sigma_{l+1} \cdot h_{l+1}}{h_l + h_{l+1}} \tag{3}$$



**Fig. 2** Three-layered medium (a), stresses in the cell of the computational mesh (b), interpolation of stresses on the computational grid (c). Star shows the injection source

where  $k$  is the current number of interpolated layer,  $l$  is the current number of the original layer,  $\sigma_l$  is the stress of the original layer with height  $h_l$ , calculated as the distance from the surface of layer to the boundary of the current cell. After applying interpolation, we obtain a five-layer medium (Fig. 2c) with stress  $\sigma_{-2} = 6$  MPa,  $\sigma_{-1} = 4$  MPa,  $\sigma_0 = 0$  MPa,  $\sigma_1 = 4$  MPa,  $\sigma_2 = 6$  MPa.

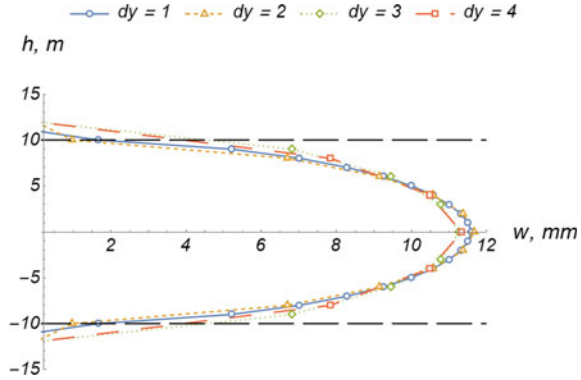
We can interpolate the characteristics of layers in the different computational mesh. Let us consider a three-layered symmetric medium with stress contrasts  $\Delta\sigma = 6$  MPa, the central layer thickness is  $H = 20$  m, plane strain modulus is  $E' = 20$  GPa. Newtonian fluid with the dynamic viscosity coefficient  $\mu = 0.4$  is injected for 3 min with the pumping rate of  $Q = 4$  m<sup>3</sup>/min. The dimensions of the fracture were calculated depending on the computational cell size  $dx = dy = 1, 2, 3, 4$  m; see Table 1.

When we increase the mesh step, the height and opening of the fracture in the source of injection do not change significantly (<4%) and the difference in length is less than 10%, which is acceptable for engineering calculations. When we increase

**Table 1** Time of calculations and fracture geometry by different cell size  $dx$

Cell size, m	Length, m	Height, m	Opening, mm	Calc. time, s
1	88.23	22.13	11.59	217.24
2	91.12	22.66	11.67	5.91
3	81.56	23.57	11.26	1.34
4	83.59	23.14	11.33	0.10

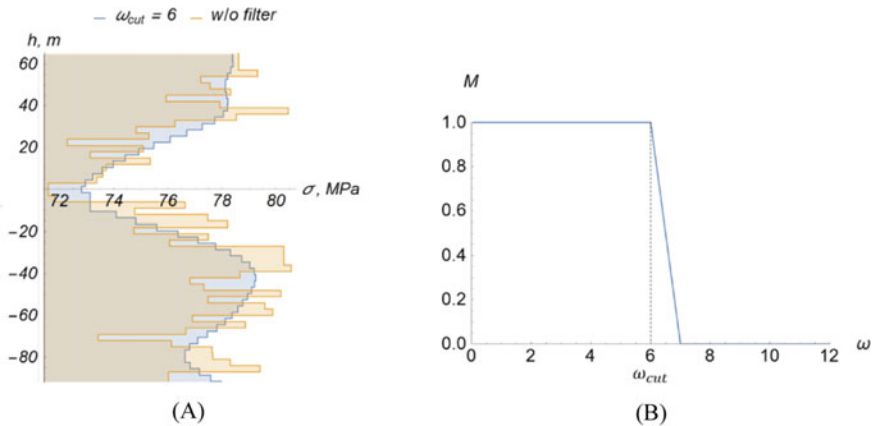
**Fig. 3** Transverse profile of the crack opening in the three-layer medium obtained with different mesh steps  $dy$ , black dashed horizontal lines show the surfaces of the layers



the mesh step, the speed of the calculation significantly accelerates. The transverse profiles of the openings are constructed depending on the various  $dy$  (see Fig. 3).

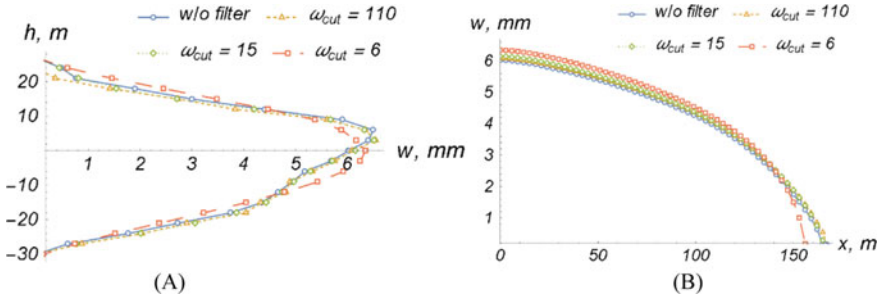
For thin layers with stress contrasts greater than 3 MPa, a low-pass filter is used. To obtain a solution we make:

1. Characteristics interpolation onto a mesh with thickness  $dy = \frac{H_{\text{bottom}} - H_{\text{top}}}{N}$ , where  $H_{\text{bottom}}$  is a coordinate of the lower surface of the lower layer,  $H_{\text{top}}$  is a coordinate of the upper surface of the upper layer and the number of elements  $N = 2^n$ ,  $n$  is an integer number. For example, let us consider multi-layered medium with stress contrasts (Fig. 5a). We can interpolate stress contrasts and receive a vector  $\sigma_l = (\sigma_0, \sigma_1, \dots, \sigma_{N-1})$ .
2. Find the Fourier image of interpolated stress contrasts  $\bar{\sigma}_l = \text{FFT}(\sigma_l)$ .
3. Calculate transfer function  $M_l = (M_0, M_1, \dots, M_{N-1})$  of low-pass filter (Fig. 4b), as:



**Fig. 4** Layers' interpolation without filter (yellow) and with the low-pass filter with  $\omega_{\text{cut}} = 6$  (a). Transfer function with cut-off frequency  $\omega_{\text{cut}} = 6$  (b)





**Fig. 5** Comparison of opening  $w$  profiles (**a** transverse, **b** longitudinal) without any filter and with different  $\omega_{cut}$

$$M_l = \begin{cases} 1 & \text{if } |l - N/2 + 1| \leq N/2 - \omega_{cut} \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where  $\omega_{cut}$  is a cut-off frequency.

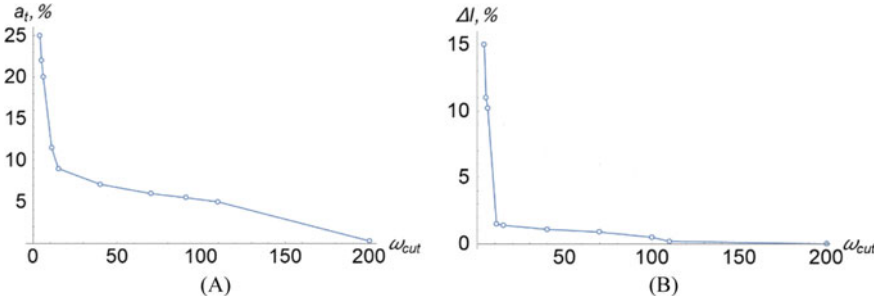
4. Obtain  $\xi_l = M_l * \bar{\sigma}_l$  using wise-element multiplication.
5. Calculate  $\sigma_l = \text{IFFT}(\xi_l)$  and interpolate onto computational mesh (Fig. 4a) with mesh step  $1 \leq dy \leq 10$  for real cases.

Let us consider a case with stress contrasts (see Fig. 5a), we apply to the original stress contrasts the algorithm, which described below, obtain filtered and interpolated in computational mesh layers. We apply different cut-off frequencies  $\omega_{cut}$  to the original stress contrasts. We use non-Newtonian fluid with the dynamic viscosity coefficient  $\mu = 0.05$  which is injected for 20 min with the pumping rate of  $Q = 5 \text{ m}^3/\text{min}$ , plane strain modulus is  $E' = 30 \text{ GPa}$ , Carter’s leak-off coefficient is  $C_L = 10 \text{ }\mu\text{m/s}^{1/2}$ , toughness is  $K_{IC} = 1.01325 \text{ MPa m}^{1/2}$ . In this study, the length of the vector is  $N = 1024$  elements.

The algorithm is simple and straightforward. Figure 5 shows the transverse and longitudinal profiles of fracture opening  $w$  with the different cut-off frequencies  $\omega_{cut}$ . We consider the transverse profile of the opening  $w$  along the vertical line passing through the injection source (Fig. 5a) and the longitudinal one along the horizontal line (Fig. 5b).

As  $\omega_{cut}$  decreases, layers’ characteristics vary less, and at  $\omega_{cut} = 0$ , the medium becomes homogeneous. When we take  $\omega_{cut} = 6$ , the dimensions of fracture (length, height, opening) differ by less than 10% in comparison with the case when no filters are used, which is an acceptable error for engineering calculations. Also, when the cut-off frequency  $\omega_{cut}$  is reduced, the calculations of the Planar3D model accelerate (Fig. 6). Let us introduce the acceleration of calculations:

$$a_t = \frac{|t - t_{w/o}|}{\max(t, t_{w/o})} \cdot 100 \quad (5)$$



**Fig. 6** Acceleration of calculations  $a_t$  (a) and difference of fracture's length  $\Delta l$  (b) of the Planar3D model for various  $\omega_{cut}$

where  $t$  is the calculation time at  $\omega_{cut}$ ,  $t_{w/o}$  is the time of calculations without using any filter, which is taken as the basis. Difference in length of fracture is:

$$\Delta l = \frac{|l - l_{w/o}|}{\max(l, l_{w/o})} \cdot 100, \quad (6)$$

where  $l$  is the fracture's length at  $\omega_{cut}$ ,  $l_{w/o}$  is the fracture's length without using any filter.

When we decrease cut-off (Fig. 6) frequency to value 7, the acceleration  $a_t$  and differences in length  $\Delta l$  of fracture slowly increase. When  $\omega_{cut} < 6$ , the acceleration of the calculations reaches 25% with a difference in the fracture size of 15%.

## 2.2 Matrix–Vector Product Computation Using FFT

Let us consider the technique of accelerating matrix–vector multiplication based on FFT. We assume that all calculations are carried out on a uniform mesh consisting only of square cells of the same size. In the Planar3D model, the connection between the distributions of opening and fluid pressures is carried out through a matrix depending on the computational domain. For more details, see [23]. The matrix contains the influence coefficients, which are calculated on the basis of the Green's function [26] and will be referred to below as the influence matrix  $A$  (here and below all matrices will be noted in bolder weight). Coefficients depend on the distance from the source to the point. Each element of such a matrix is a submatrix showing the influence of the concentrated force in a given cell on other cells of the computational mesh. We use a square uniform mesh of size  $N \times N$ , and the influence matrix of the fourth rank is  $N \times N \times N \times N$ , where  $N$  is the number of mesh elements in its row. The pressure distribution  $p$  and the opening  $w$  of the fracture are  $N \times N$  matrices. It is convenient to rearrange each submatrix of the influence matrix into a vector and rebuild the resulting row matrix into a column. Similarly, the pressure matrix is rearranged into

a column vector, after which the fracture opening in vector form can be found as the product of the new influence matrix (size is  $N^2 \times N^2$ ) and pressure vector (size is  $N^2$ ). The pressure calculation based on the direct matrix–vector multiplication requires significant time expenditures, and asymptotic complexity of the algorithm is  $O(N^4)$  algorithm,  $N \sim 100$ .

The application of the fast Fourier transform will reduce the asymptotic complexity of the algorithm from  $O(N^4)$  for the direct method of multiplication to  $O(N^2 \log N)$ , while the boundary conditions (BC) in the problem should be replaced by periodic ones. When periodic BC is emplaced, the influence matrix has a symmetry relative to the point source and can be rebuilt as a one-dimensional vector of influence coefficients. The symmetry of this vector can be formulated in the form:

$$c_i = c_{N-i}, 1 \leq i \leq N - 1, \tag{7}$$

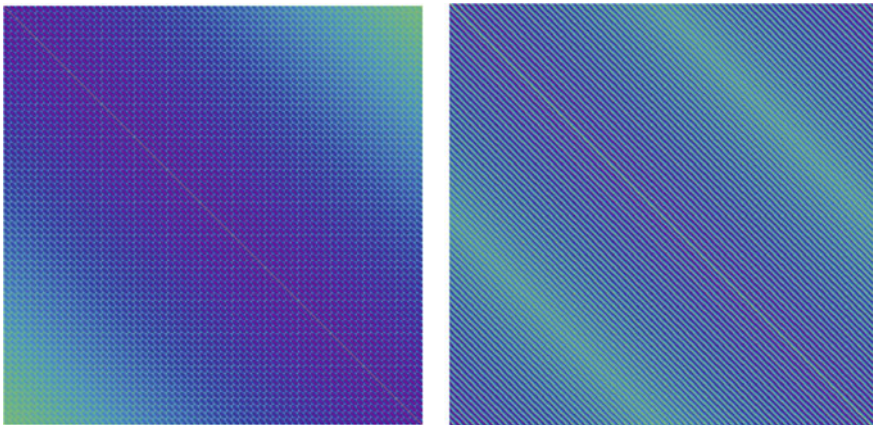
where  $c_i = (c_0, c_1, \dots, c_{N-1})$  is influence vector,  $N$  is number of elements.

It can be shown that when the size of the computational domain is of the order of two maximum dimensions of the propagating fracture introduced by modified BC, the error will be insignificant.

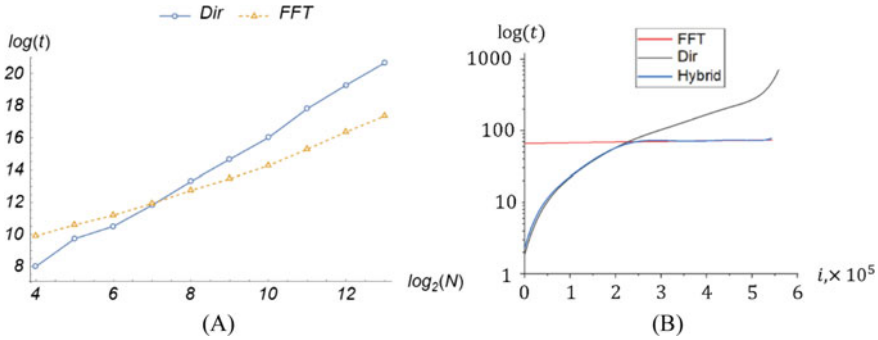
Hence, the influence matrix has the form of a symmetric matrix:

$$A = \begin{pmatrix} a_0 & \cdots & a_{n-1} \\ \vdots & \ddots & \vdots \\ a_{n-1} & \cdots & a_0 \end{pmatrix} \tag{8}$$

Due to the imposition of boundary periodic conditions on the matrix  $A$ , it takes the form (see Fig. 7):



**Fig. 7** Influence matrix  $A$  (left) and its circulant form  $C$  (right)



**Fig. 8** Time of calculation direct and modified algorithms (a), time of calculation 1-h hydraulic fracturing with deferent methods of matrix–vector multiplication (b)

$$C = \begin{pmatrix} a_0 & \cdots & a_{n-1} \\ \vdots & \ddots & \vdots \\ a_1 & \cdots & a_0 \end{pmatrix} \tag{9}$$

Computation of matrix–vector product consists of following steps [27]:

1.  $y = \text{FFT}(c)$ , where  $c$  is the first row of  $C$ .
2.  $x = \text{FFT}(w)$ , where  $w$  is the vector of opening.
3.  $h = x * y$ , wise-element multiplication.
4.  $p = \text{IFFT}(h)$ , obtain pressure vector using IFFT.

The advantage of this method is that the only one column vector  $c$  is stored in memory, while the direct method of matrix multiplication requires storing the full matrix  $A$  of size  $N^2 \times N^2$ .

Computer experiments were conducted to compare the calculation speed with different algorithms (using the direct method of multiplying a matrix by a vector and using the algorithm described above (see Fig. 8a). In the experiments, a random circulant matrix and a random vector were generated, after which the average execution time of the two algorithms was compared. Results have been calculated 1000 times and averaged.

In the case when the number of elements of the opening vector  $w$  is less than 128, speed of the modified method is inferior to the direct method of multiplication, but starting with 1024 elements manifests itself a multiple acceleration of the matrix–vector multiplication. With a vector size of 4096, the multiplication procedure is accelerated by about 20 times. It should be noted that at small vector lengths it turns out to be about two times slower than the direct method; that is, the speed gap is small and is associated with the quality of the used computer optimization of matrix multiplication.

The proposed algorithm is implemented in the Planar3D model with the explicit numerical integration scheme. As a result, we compared time of calculations with different methods of the matrix–vector multiplication on each numerical integration

step  $i$  (see Fig. 8b). Three versions of the product algorithm were compared: direct, modified, and hybrid methods of matrix–vector multiplication. The hybrid method implies that at first the direct method of matrix multiplication is used, and when a certain number of elements on the crack are reached, it is changed to a modified one. The calculations were carried out for the propagation of a fracture in a homogeneous medium. Three versions of the calculation program are written in C++ and use the maximum optimization level of the compiler. To average the results, five series of calculations were performed.

The size of the influence matrix and the opening vector in the crack depends on the number of mesh elements. In this way, faster calculations of the matrix–vector product on small crack sizes are achieved. The study revealed fracture size (1000 elements) at which the modified method gives the advantage of (Fig. 8b). The acceleration of the program execution by two times was obtained during the simulation of hourly hydraulic fracturing with a constant injection rate using the hybrid method of matrix–vector multiplication.

Thus, in this section, we apply the method of accelerating matrix–vector multiplication using FFT for matrices that are close to the circulant structure. With a matrix size of 4096, an increase in the computational speed by about 20 times is obtained. When this method was used in the implementation of the planar hydraulic fracturing model, the performance of the computational module has increased approximately twice.

### 3 Conclusions

In this paper, we propose methods for accelerating the Planar3D model of a hydraulic fracturing with the explicit time integration scheme using FFT. The C++ implementation uses the radix-2 Cooley–Tukey algorithm and the `std::vector()` STL container. The correctness and speed of calculations were compared with well-known libraries. During the analysis, it was found that the speed of calculations is comparable with other libraries. The library was used to accelerate the matrix–vector multiplication and process the input data of lithology. In the Planar3D model, the influence matrix can be reduced to a circular form by applying periodic boundary conditions, and the use of a simple FFT-based algorithm can increase the model computation speed. The acceleration was 200% when calculating the propagation of a crack in a homogeneous medium. FFT can also be used in processing the input data; an example of such an application with layer approximation is presented in 2.1. The use of a low-pass filter allows you to accelerate the calculation by 20% within the acceptable accuracy for engineering calculations.

**Acknowledgements** The authors are grateful to A. S. Linkov, V. A. Kuzkin, and N. S. Markov for useful and stimulating discussions.

This research was financed by Ministry of Education and Science of the Russian Federation within the framework of the Federal Program “Research and development in priority areas

for the development of the scientific and technological complex of Russia for 2014–2020,” Activity 1.2., Agreement on Grant No. 14.575.21.0146 of 26 September 2017, unique identifier: RFMEFI57517X0146.

## References

1. Heideman, M.T., Johnson, D.H., Burrus, C.S.: Gauss and the history of the fast Fourier transform. *Arch. Hist. Exact Sci.* **34**(3), 265–277 (1985)
2. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex Fourier series. *Math. Comput.* **19**(90), 297–301 (1965)
3. Good, I.J.: The interaction algorithm and practical Fourier analysis. *J. Roy. Stat. Soc.: Ser. B (Methodol.)* **20**(2), 361–372 (1958)
4. Bruun, G.: z-transform DFT filters and FFT’s. *IEEE Trans. Acoust. Speech Signal Process.* **26**(1), 56–63 (1978)
5. Rader, C.M.: Discrete Fourier transforms when the number of data samples is prime. *Proc. IEEE* **56**(6), 1107–1108 (1968)
6. Dentino, M., McCool, J., Widrow, B.: Adaptive filtering in the frequency domain. *Proc. IEEE* **66**(12), 1658–1659 (1978)
7. Van Nee, D.J.R., Coenen, A.J.R.M.: New fast GPS code-acquisition technique using FFT. *Electron. Lett.* **27**(2), 158–160 (1991)
8. Bellanger, M., Daguët, J.: TDM-FDM transmultiplexer: digital polyphase and FFT. *IEEE Trans. Commun.* **22**(9), 1199–1205 (1974)
9. Raja, J., Radhakrishnan, V.: Filtering of surface profiles using fast Fourier transform. *Int. J. Mach. Tool Des. Res.* **19**(3), 133–141 (1979)
10. White, S.: A simple FFT butterfly arithmetic unit. *IEEE Trans. Circ. Syst.* **28**(4), 352–355 (1981)
11. Lib Homepage: <https://github.com/NikitaMushchak/Fast-Fourier-Transform>
12. Wang, E., et al.: Intel Math Kernel Library. High-Performance Computing on the Intel® Xeon Phi™, pp.167–188. Springer, Cham (2014)
13. Li, Y., et al.: MPFFT: an auto-tuning FFT library for OpenCL GPUs. *J. Comput. Sci. Technol.* **28**(1), 90–105 (2013)
14. Gu, L., Li, X., Siegel, J.: An empirically tuned 2D and 3D FFT library on CUDA GPU. In: Proceedings of the 24th ACM International Conference on Supercomputing. ACM (2010)
15. Taboada, J.M., et al.: MLFMA-FFT parallel algorithm for the solution of large-scale problems in electromagnetics. *Progr. Electromagn. Res.* **105**, 15–30 (2010)
16. Frigo, M., Johnson, S.G.: FFTW: an adaptive software architecture for the FFT. In: Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP’98 (Cat. No. 98CH36181), vol. 3. IEEE (1998)
17. Galassi, M., et al: GNU scientific library. Network Theory Limited, 65–76(2002).
18. Eigen library homepage, [https://eigen.tuxfamily.org/index.php?title=Main\\_Page](https://eigen.tuxfamily.org/index.php?title=Main_Page)
19. Swartztrauber, Paul, N., et al.: Bluestein’s FFT for arbitrary n on the hypercube. *Parallel Comput.* **17**(6–7), 607–617 (1991)
20. Pieterse, V., et al.: Performance of C++ bit-vector implementations. In: Proceedings of the 2010 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists, pp. 123–145. ACM (2010)
21. Peirce, A.: Modelling multi-scale processes in hydraulic fracture propagation using the implicit level set algorithm. *Comput. Methods Appl. Mech. Eng.* **283**, 881–908 (2015)
22. Khasanov, M.M., et al.: Scientific engineering as the basis of modeling processes in field development. *Georesources* **20**, 142–148 (2018)

23. Starobinskii, E.B., Stepanov, A.D.: Adapting the explicit time integration scheme for modeling of the hydraulic fracturing within the Planar3D approach. *J. Phys. Conf. Ser.* **1236**(1), 43–47 (2019)
24. Linkov, A.M.: The particle velocity, speed equation and universal asymptotics for the efficient modelling of hydraulic fractures. *J. Appl. Math. Mech.* **79**(1), 54–63 (2015)
25. Linkov, A.M.: Universal asymptotic umbrella for hydraulic fracture modeling. arXiv preprint [arXiv:1404.4165](https://arxiv.org/abs/1404.4165) (2014)
26. Van Tiggelen, B.A.: Green function retrieval and time reversal in a disordered world. *Phys. Rev. Lett.* **91**(24), 243904 (2003)
27. Strang, G.: A proposal for Toeplitz matrix calculations. *Stud. Appl. Math.* **74**(2), 171–176 (1986)

# Employee Performance Analytics Approach Based on Anomaly Detection in User Activity



Aleksey Lukashin , Mikhail Popov , Dmitrii Timofeev ,  
and Igor Mikhalev 

**Abstract** In this paper, we highlight the distinctive features and critical areas of analytical tool application for the employee performance analytics of user activity. We describe problems of applying data analytics methods and technologies to ensure employee performance analytics. We also discuss the use of user activity time-series data analysis methods and techniques to provide employee performance analytics and describe approaches for processing unstructured data from different sources of user activity for further analytics using anomaly detection methods. Finally, we introduce a new strategy of building features from hybrid data streams from different sources and compare it with current practices.

**Keywords** Data analysis practices · Anomaly detection · Machine learning · Feature engineering · Time-series

## 1 Introduction

To assess the quality of the performed tasks, the effectiveness of the production targets, and increase the labour outcome in general, it is necessary to develop methods and technical solutions that will allow managing the process of work of employees in an automated mode to increase their efficiency. Two approaches are possible. One approach requires the formalisation of work quality criteria. Another one does not use formal criteria and relies upon the feedback from the managers, the task planning system, or the comparison with the industrial plan. The formalised criteria allow the precise description of the employee's activity and the representation of his or her efficiency in quantitative evaluation. However, this approach has the following disadvantages.

---

A. Lukashin · M. Popov (✉) · D. Timofeev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [popov\\_m@spbstu.ru](mailto:popov_m@spbstu.ru)

I. Mikhalev  
University of Amsterdam, Amsterdam, The Netherlands



1. Efficiency criteria will be quite different for employees of different professions. It is not correct to compare a sales manager, a researcher, or a production worker. Each profession requires an individual approach and different performance criteria, which makes it much more challenging to develop a unified solution.
2. For some professions, especially intellectual work, it is rather challenging to present adequate efficiency criteria. Besides, the positive effect of working activity can be postponed over time.

An alternative solution is the use of an informalised approach to evaluating efficiency using artificial intelligence technologies. For this purpose, we proposed to collect metrics about the activities of the company's employees and put them in common storage of poorly structured heterogeneous data (data lake class systems).

Similar approaches are actively used to solve cybersecurity problems in large corporate infrastructures (SIEM and UEBA class systems) [1]. Depending on the specifics of the work and profession of an employee, the following metrics, among the others, can be used:

1. Events about the time of arrival/departure from work;
2. Emails, including their content;
3. Number of calls made (e.g., for sales managers);
4. Number of solved tasks;
5. Domain-specific performance metrics, like the number of commits, software builds, or added lines of code in the software engineering field;
6. Number and duration of business meetings;
7. Employee's location;
8. Workplace activity, including the intensity of typing, mouse movements and so on;
9. Network activity, visited Internet resources, entry points.

Metrics should be collected by software agents installed in the corporate infrastructure and stored in a common repository in the form of messages.

To analyse the activity of the employee, we propose to apply methods for detecting anomalies in the data stream using machine learning and artificial intelligence technologies. To provide universality of the work of the developed software complex, we propose to create a design tool to allow the system operator to select the specific activity types and their attributes to analyse the effectiveness of a particular kind of employee. The data stream entering the storage system can be split into streams belonging to different entities (e.g., "Ivan Kornilov", "Accounting Department", manufacture) and apply an appropriate analyser to each stream.

However, it is not enough to identify anomalies in the behaviour of subjects. It is also necessary to classify the identified anomaly to allow the manager to take decisions. Additionally, we may apply well-known process mining methods [2] to extract the business process model from logs. An anomaly may correspond to the employee's health issues, the beginning of solving an atypical task, an extra cup of coffee, or a good mood, and, as a result, increased productivity.

In this paper, we make the following contributions:

1. A novel architecture of a software platform to analyse the activity of employees. This platform processes the data collected by software agents, including the data extracted from informational systems such as issue trackers and task management services, using the extensible set of analysers. The software platform orchestrates the execution of analysers, feeds the relevant events to their input, and sends the analysis results to the consumers.
2. An algorithm of anomaly detection based on the isolation forest method. We illustrate the algorithm using the sample sequence of operations performed by software developers in a task management system.

The proposed approach allows us to implement a system for assessing and managing the performance of employees without explicit criteria, which, in turn, will allow to significantly expand the range of applications of the developed program complex, including in fundamentally different fields of activity.

## 2 The Architecture of the Software Platform for Employee Performance Analytics

The system under development is designed to support the interaction of a complex of automation tools with the popular corporate applications, as well as to evaluate the effectiveness of the employees' work, his work in the group based on data collected by the adapters corporate services and data on physiological conditions and on computer actions (keystrokes, mouse actions, launching applications, operations in Visual Studio, and so on). Also other types of events may be added later by an event-processing software that analyses the entire event log using a personalised user activity model corresponding to the set of business processes that the user implements at the workplace (in particular, a hidden Markov model [3], a hierarchical hidden Markov model [4], or a hidden semi-Markov model [5] that detect a specific set of complex activities consisting of several simpler actions). For example, for software developers, the set of inferred activities includes operations on the source code, like the creation of classes and methods, or even more higher-level activities like introducing new functionality or fixing bugs [6].

Physiological conditions, which are directly interpreted from the indicators of a person's psychophysiological state by pulse, skin electroconductivity, temperature, EEG signals and other indicators of a person's condition, directly affect the efficiency and quality of the work performed and serve as an input to the methods for efficiency analysis [7]. Although signals like temperature and pulse may be directly interpreted, the EEG samples are too low level to use in the context of the overall user activity. To provide the necessary feedback to the user, we need to compute interpretable aggregate states based on the EEG data, such as levels of concentration, cognitive load, stress, or fatigue [8–11].

Corporate services are understood as business systems of the organisation, such as task management system (Jira, Redmine, YouTrack, and others), user support systems (Jira Service Desk, ServiceNow, and so on), sales and customer relationship management systems (CRM), knowledge management systems (Confluence and others).

The tasks of the corporate services adapter are:

1. Receiving and structuring information from the business systems of the organisation;
2. Managing of the organisation’s business systems in terms of creating new entities (tasks, meetings, documents, applications, and other).

The adapters should perform the transformation of an internal representation of requests to corporate services into external formats, and inverse transformation of responses from corporate services to an internal form to perform these tasks.

Figure 1 shows the prototype of the common architecture of the software complex.

For common use of hybrid data streams from different sources, the system should have an expandable architecture and provide connection of new business systems by developing extensions (plug-ins) that implement the functionality of interaction and integration with the business system with the conversion of data between external formats of corporate services and internal formats. These features complicate the direct application of machine learning methods, in particular, neural network models [12–14]. The data must be pre-processed to be used with advanced analytics methods.

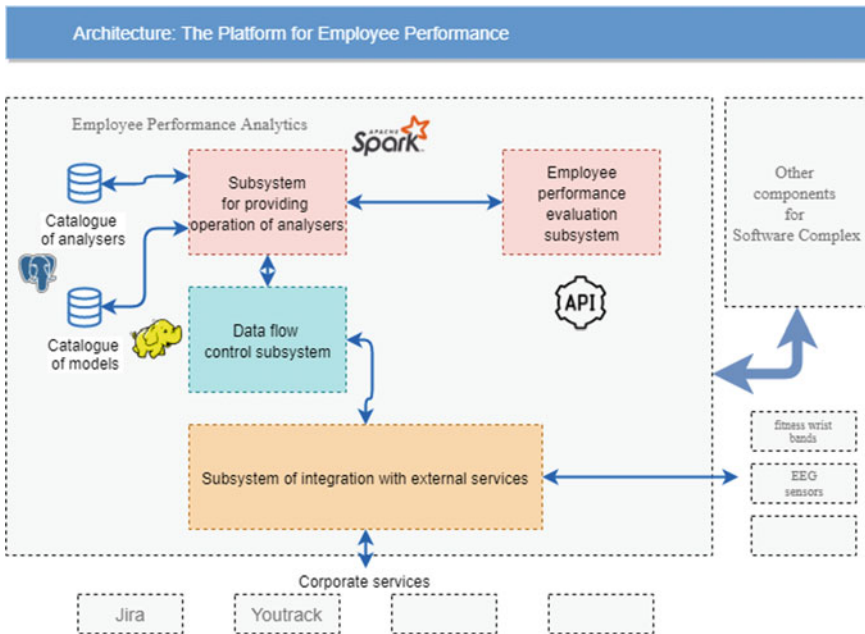


Fig. 1 Architecture of the platform for employee performance prototype

The development of appropriate pre-processing procedures (Extract, Transform, Load (ETL)) takes 80% of design time of a data analyst (engineer) [15].

The required functionality is implemented using the concept of “catalogue of intelligent analysers”. Within the context of this concept, the subsystem includes the following services:

*Catalogue of analysers.* It is a service providing cataloguing of heterogeneous analysers implementing functions of detecting anomalies and calculating the performance of an employee in data flows. Thus, users of the system or third-party services can quickly change the execution environment of models and data sources for its verification, which makes the process of building high-precision forecasts much easier. Each analyser in the catalogue has a version and type, which allows you to facilitate the development of analysers.

*Catalogue of models.* It is a service that provides cataloguing and control over the life cycle of intelligent and analytical models to assess the performance of employees. The use of machine learning allows for deep analysis of hidden relationships between controlled parameters. However, the use of this approach requires high costs for the preparation and configuration of model parameters, the formation of the necessary data set for training and verification of the model, as well as significant computational resources/time to complete the training process.

The use of the model catalogue allows iterative improvement of the quality of the models, controlling the learning and verification process at each step. The user of the system, using the model from the model library, knows precisely on which data set it was trained and verified.

*Data flow control service.* The service is intended for separation of the necessary data from the general stream (e.g., data about specific employees or data from particular services) and transfer of the data stream to the input of the necessary analyser.

*Service of the analysers operation support.* It is necessary to give the possibility of launching multiple analysers in a single execution environment, with shared access to the Apache Spark cluster, to provide flexibility and scaling of the developed software complex. This service should manage analysers life cycle (creation, deletion, scheduled, run).

The architecture of the proposed solution is a server application with access to the PostgreSQL-based DBMS. It allows storing the information about the analysers being launched and the schedule for their launch. Inside each analyser, an performance evaluation code is executed, which can be implemented using the implementation of machine learning methods in the Python language and the PySpark library for interaction with the Apache Spark cluster [16].

### 3 Anomaly Detection Approach for Semistructured Data

The performance of employees has always been a significant concern in organisations. There are many metrics for assessing employee performance; some of these metrics can be calculated using time-series [17]. Prediction of a specific indicator and its comparison with the real value allows you to detect for abnormal behaviour. Time series analysis is an approach to analyse time-series data to extract essential characteristics of data and generate other useful insights applied in a business situation. Generally, time-series data is a sequence of observations stored in time order. Time-series data often stands out when tracking business metrics, monitoring industrial processes, etc. In data forecasting, research is widely conducted, for example, employees of the Facebook company have released a library “Prophet” [18, 19], which allows you to study time series taking into account the nature of the data and predict the parameters taking into account seasonality, weekends and other.

The unsupervised method based on Isolation Forest [20] was chosen to identify anomalies. Feature engineering for analysis was developed drawing on statistics collection, which consists of collecting information on the frequency of occurring values for the observed keys of the analysed fields in events [21].

Collecting statistics is the process of obtaining weights for each value of each key from the significant fields in the event. The weights obtained allow to form vectors for further analysis and search for anomalies. The parametrisation of the statistics collection allows flexible adjustment of the feature vector correlation [22].

In the software implementation, the statistics are presented in the form of map  $M$ , containing such values as,  $m[key|value] = 0.1$ , where *key|value* is a component key consisting of polar events and a specific value.

Statistics are calculated periodically by time or on batch of events. At each step, the statistics merge with the previous statistics through a forgetting ratio. This approach allows the system to implement memory and gradual adaptation to the current situation, as well as to analyse events on the stream, close to real time.

Necessary steps to get feature vectors [22]:

1. Getting the occurrence frequency of value in the chunk of events.
2. Obtaining weights for each key-value based on the previous and current statistics using the “averaging” algorithm and using the forgetting coefficient (taking into account the current and prior values using the ratios). This step allows implementing a system with memory and flexible adaptation to changes in input data.
3. Normalising of weights and refusing from weights by a threshold value.

The following are the steps for calculating statistics that are calculated periodically from the time window:

1. Getting the frequency of occurrence of the value in the chunk of events is according to the formula (1).

$$v^{k_i} = \frac{\mathit{count}_i^k}{T} \quad (1)$$

where  $\mathit{count}_i^k$  is the number of events with the same value for a specific key,  $T$  is the window period for which events collect.

2. Obtaining weights for each key value based on the previous, current statistics and normalisation of weights and refusal from weights by a threshold value is according to the formula (2).

$$\omega_i^k = \omega_{i-1}^k * k_f + \frac{v_i^k}{\mathit{median}(v_i^k) * \mathit{count\_uniq}_i^k(v_i^k)} * (1 - k_f), \quad (2)$$

$$\omega_0^k = \frac{v_0^k}{\mathit{median}(v_0^k) * \mathit{count\_uniq}_0^k(v_0^k)}$$

where  $\omega_{i-1}^k$  is the value of the normalised weight at the previous step (previous chunk),  $k_f$  is the coefficient of forgetting the weight for the previous step,  $v_i^k$  is the frequency of occurrence of the value for the current step,  $\mathit{count\_uniq}_i^k(v_i^k)$  is the number of enqueueing value for a specific key.

The strengths of statistics are calculated periodically by the time:

- The approach takes into account time specificity of events and correlation of activities by the time window.
- It allows us to get a generalised characteristic of the flow, which in the future makes it possible to catch not only anomalies in events but also anomalies in the behaviour of their sources.

Weaknesses of this calculation are the following:

- Due to the data collection specificity, the events not sorted. Because of that, statistics validity distorted with a large data stream. To bring the analysis closer to real time, we refused to sort events. If desired, one can use several approaches to windows: fix the window in time from above, from below or from two sides. In this case, events that have a timestamp below the lower or upper threshold of the window can skip. It also becomes possible to skip events due to their delay from certain sources due to their temporary unavailability. The second approach is to use only the upper limit for events. There may be the possibility of incorrect calculation of statistics due to the presence of events not from the current time window.

Examples and a more detailed description of the calculation of statistics can be found in [22].

### 4 Example of Cases

Two simple examples were chosen to demonstrate the approach. The data on the employees from the YouTrack system and the analysed of the following metrics by day for nine months of work were studied: the number of active tasks in work and working time logging by employees.

Figure 2 shows a histogram of task distribution for an individual employee, which shows the distribution of tasks by days. Here you can see seasonality, weekends and vacations.

Figure 3 shows the anomalies for this user based on his history and deviation from predictions.

Figure 4 shows a histogram of working time logging for an individual employee, which shows the distribution of tasks by days.

Figure 5 shows the anomalies for this user based on his history and deviation from predictions for worklog data.

This simple example shows for an individual employee an analysis of two metrics based on which an anomaly detection methods can be used to detect abnormal behaviour of an employee on certain days and to analyse the effectiveness of the employee, paying particular attention to the anomaly. This method can be easily extended to more complex metrics of employee performance by combining the rest of the data, as well as by grouping employees by behaviour, or directly by department, profession.

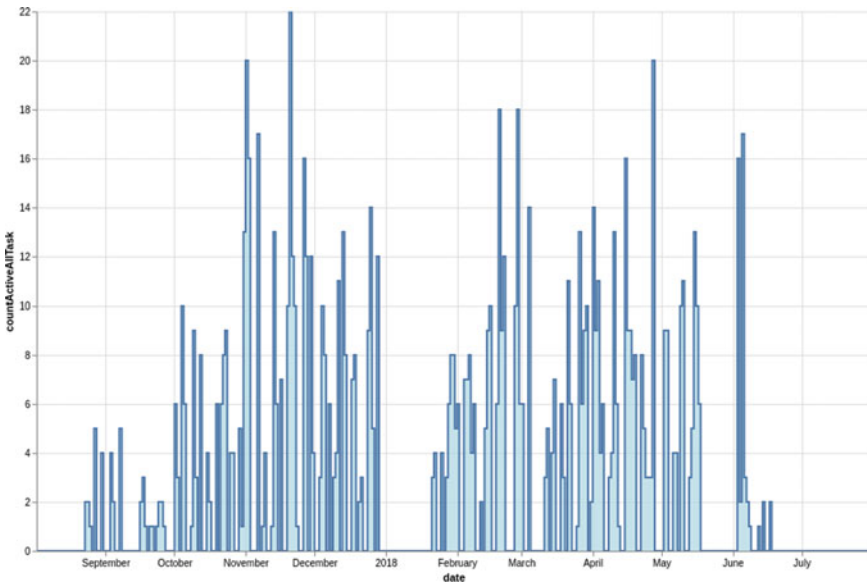
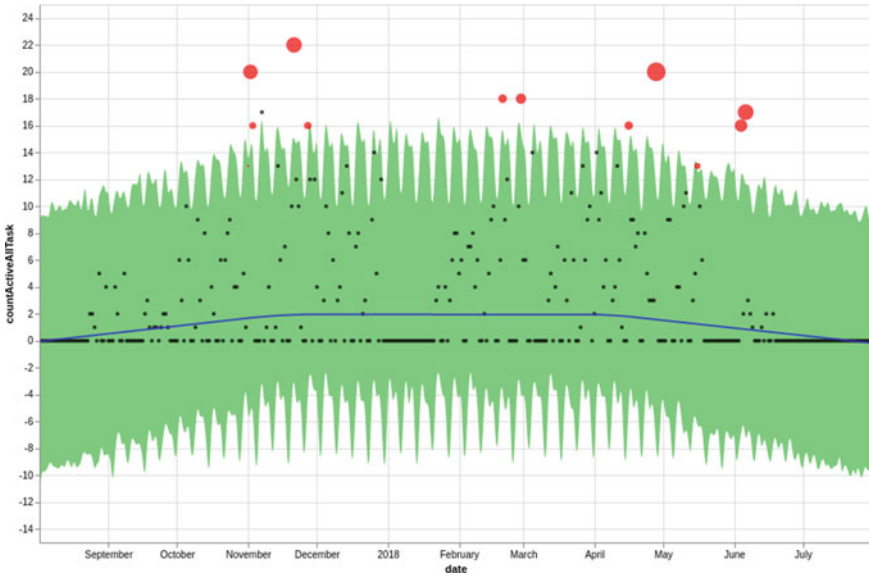
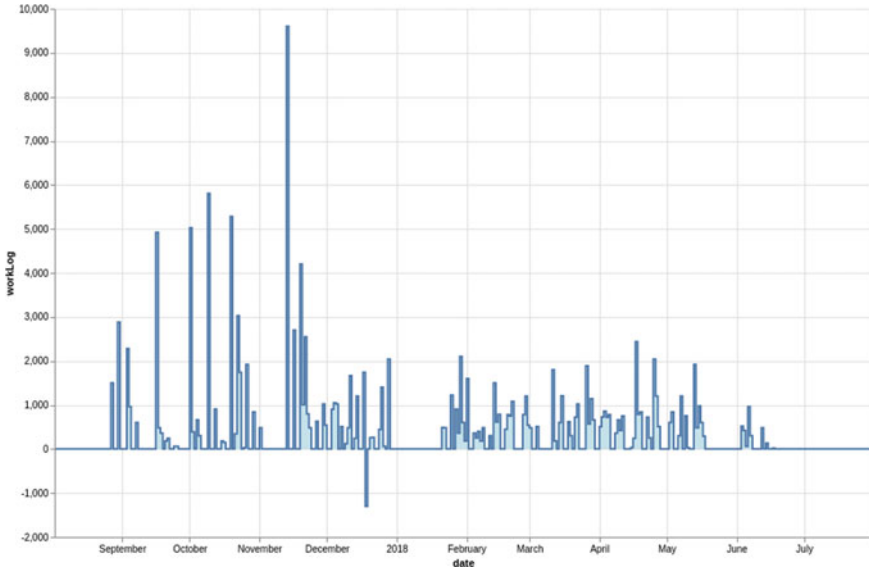


Fig. 2 Histogram of task distribution for an individual employee

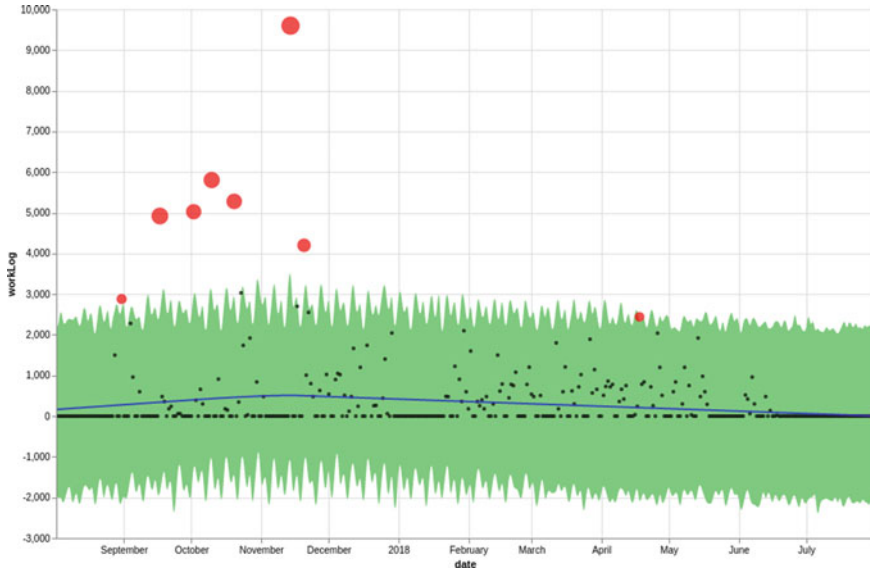


**Fig. 3** Anomalies of the user on the basis of his history of active tasks in work and deviations from predictions



**Fig. 4** Histogram of working time logging for an individual employee





**Fig. 5** Anomalies of the user on the basis of his history of working time logging and deviations from predictions

## 5 Conclusion

In this paper, we described an approach to the employee performance analytics based that relies upon tracking the activity and psychophysiological state of the workers and anomaly detection on this data. We also presented a software platform that implements this approach, which takes into account processing unstructured data from different sources of user activity.

In the future work, we will extend a set of analysers and experiment with different machine learning methods including:

- Research of possibility to use LSTM method for anomaly detection;
- Research of survival analysis methods for evaluating the probability of performing effectiveness of the employees' work;
- Adding other types of events related to a person's actions to expand the analysers' library.

The proposed architecture allows to extend the prototype by adding new analysers to the digital library and perform the different experiments on the same data.

**Acknowledgements** This research is a part of the joint project by Intelin LLC (Moscow, Russia) and Peter the Great St. Petersburg Polytechnic University (St. Petersburg, Russia). This work is financially supported by the Ministry of Education and Science of the Russian Federation (state contract 03.G25.31.0247 from 28.04.2017).

## References

1. Litan, A.: Market guide for user and entity behavior analytics. Gartner (G00276088), p. 22 (2015)
2. Van der Aalst, W.M.: Process Mining: Discovery, Conformance and Enhancement of Business Processes. Berlin (2011)
3. Rabiner, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proc. IEEE* **77**(2), 257–286 (1989)
4. Fine, S., Singer, Y., Tishby, N.: The hierarchical Hidden Markov Model: analysis and applications. *Mach. Learn.* **32**(1), 41–62 (1998)
5. Yu, S.-Z.: Hidden semi-Markov models. *Artif. Intell.* **174**(2), 215–243 (2010)
6. Timofeev, D., Samochadin, A.: An unified representation of source code authoring workflows. In: Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 3, KMIS, pp. 228–232 (2018)
7. Goryushko, S.M., Samochadin, A.V.: Tools for cognitive load evaluation in the educational process. *Comput. Tools Educ.* **4**, 35–44 (2018) (In Russian)
8. Berta, R., Bellotti, F., De Gloria, A., Pranantha, D., Schatten, C.: Electroencephalogram and physiological signal analysis for assessing flow in games. *IEEE Trans. Computat. Intell. AI Games* **5**(2), 164–175 (2013)
9. Klimesch, W.: EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis. *Brain Res. Rev.* **29**(2–3), 169–195 (1999)
10. Krigolson, O.E., Williams, C.C., Norton, A., Hassall, C.D., Collingo, F.L.: Choosing MUSE: validation of a low-cost, portable EEG system for ERP research. *Front. Neurosci.* **11**(109) (2018)
11. Hou, X., Liu, Y., Sourina, O., Tan, Y.R.E., Wang, L., Mueller-Wittig, W.: EEG based stress monitoring. In: IEEE International Conference on Systems, Man, and Cybernetics, pp. 3110–3115. IEEE (2015)
12. Filonov, P., Lavrentyev, A., Vorontsov, A.: Multivariate industrial time series with cyber-attack simulation: fault detection using an lstm-based predictive data model. arXiv preprint [arXiv:1612.06676](https://arxiv.org/abs/1612.06676) (2016)
13. Utkin, L.V.: A framework for imprecise robust one-class classification models. *Int. J. Mach. Learn. Cybern.* **5**(3), 379–393 (2014)
14. Utkin, L., Zhuk, J.: Robastnaja model' obnaruzhenija anomalij s ispol'zovanijem modeli zasorenija. *Vestnik Komp'juternyh I Informacionnyh Tehnologij* **7**, 47–51 (2013)
15. Rieger, C., Manic, M.: On critical infrastructures, their security and resiliencetrends and vision. arXiv preprint [arXiv:1812.02710](https://arxiv.org/abs/1812.02710) (2018).
16. Kumari, R., Singh, M.K., Jha, R., Singh, N.K.: Anomaly detection in network traffic using K-mean clustering. In: 2016 3rd International Conference on Recent Advances in Information Technology (RAIT), pp. 387–393. IEEE (2016)
17. Inuwa, M.: Job satisfaction and employee performance: an empirical approach. *Millennium Univ. J.* **1**(1), 90 (2016)
18. Taylor S.J., Letham, B.: Forecasting at scale. *Am. Statist.* **72**(1), 37–45 (2018)
19. Bucur, S.L., Moldoveanu, F.: Anomaly detection for time series infrastructure metric data. In: 2019 22nd International Conference on Control Systems and Computer Science (CSCS), pp. 170–175. IEEE (2019)
20. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: Eighth IEEE International Conference on Data Mining, pp. 413–422. IEEE (2008)
21. Sun, L., Versteeg, S., Boztas, S., Rao, A.: Detecting anomalous user behavior using an extended isolation forest algorithm: an enterprise case study. arXiv preprint [arXiv:1609.06676](https://arxiv.org/abs/1609.06676) (2016)
22. Lukashin, A., Popov, M., Bolshakov, A., Nikolashin, Y.: Scalable data processing approach and anomaly detection method for user and entity behavior analytics platform. In: International Symposium on Intelligent and Distributed Computing, pp. 344–349. Springer, Cham (2019)

# Deep Predictive Control



Dmitry Baskakov  and Vyacheslav Shkodyrev 

**Abstract** Modern control systems are characterized by high complexity and hierarchies of many orders. In the case of a large number of nodes, real-time control becomes a non-trivial task and requires new concepts and paradigms. This paper discusses the concept of using model predictive control (MPC) and deep predictive control (DPC) in the tasks of control complex objects. The ability of deep networks to generalize allows using them to build effective control systems of increased complexity, working in conditions of uncertainty and limited data.

**Keywords** Deep Q-networks · Model predictive control · Neural network · Real time control · Reinforcement learning · Temporal difference learning

## 1 Introduction

Modern control systems work with a large number of parameters that do not allow the construction of real-time systems due to the computational complexity of a number of procedures. The hypothesis is made in the work that the use of deep networks will make it possible to build promising high-loaded control systems with a powerful degree of generalization that will remain operational even in case of failures of entire hierarchies. This ability is provided primarily by the a priori properties of deep networks. The paper proves the possibility of reducing the number of nodes and loops of the control system while maintaining all the parameters of the target control function. Model predictive control (MPC) is a dynamic optimization technique widely used in industrial process. MPC is used in robotics for the control of ground [1] and humanoid [2]. The transition from the process industry to robotics brings additional challenges since the computation time is reduced from hours to milliseconds. Model predictive control is considered by a number of authors as a very promising concept for the development of traditional control systems [3]. End-to-end learning is attractive for the realization of autonomous cyber-physical systems, thanks to the appeal

---

D. Baskakov (✉) · V. Shkodyrev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [dmitry.e.baskakov@gmail.com](mailto:dmitry.e.baskakov@gmail.com)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021  
N. Voinov et al. (eds.), *Proceedings of International Scientific Conference on Telecommunications, Computing and Control, Smart Innovation, Systems and Technologies* 220, [https://doi.org/10.1007/978-981-33-6632-9\\_29](https://doi.org/10.1007/978-981-33-6632-9_29)

333

of control systems based on a pure data-driven architecture. By taking advantage of the current advances in the field of reinforcement learning, several works in the literature showed how a well-trained deep neural network that is capable of controlling cyber-physical systems to achieve certain tasks [2]. A key problem in control complex systems is the use of large amounts of data online. At the same time, the computational load on such systems grows in terms of constructing and calculating the objective control function. It is additionally important to understand that it is not always possible not only to obtain this data, but also to process it in some way. A feature of the proposed solution is that neural networks can work with data that does not arrive synchronously and also have the ability to perfectly cope with omissions of such data. In addition, and the taste with the possibility of generalization, this approach allows us to talk about the construction of a new type of control systems that allows you to work in a space of limited dimension, incomplete data, filling in the possible gaps and computational limitations of deep learning control [4]. Together with the latest achievements in the field of deep learning using advanced computing resources, we can safely begin to talk about breaking the paradigm of the traditional approach to control and the transition to new-type control models.

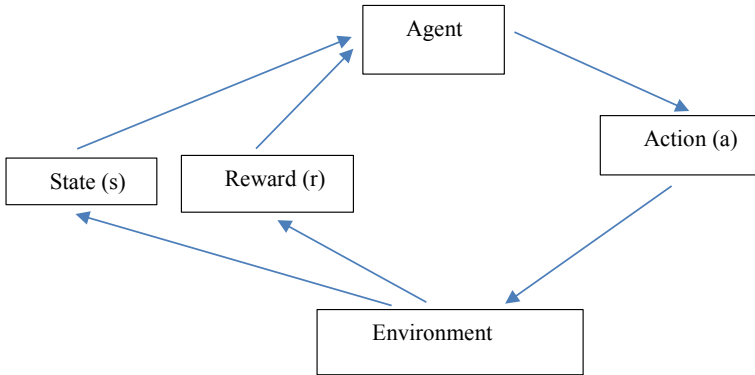
### ***1.1 Related Work***

The idea of training neural networks to mimic the behavior of model predictive controllers can be traced back to the late 1990s where neural networks trained to imitate MPC controllers were used to navigate autonomous robots in the presence of obstacles and to stabilize highly nonlinear systems [5]. The use of neural networks as control modules is not new in itself and has been proposed more than once [6]. A lot of work was also devoted to the tasks of predictive analytics or the use of neural networks in such structures with varying degrees of success, which indicates a high problem of the task [7]. New solutions and proposals for various problems appear, because the paradigm of deep learning and control complex systems based on it, nevertheless, is more relevant to engineering solutions, if we talk about creating specific working systems [8].

## **2 Materials and Methods**

### ***2.1 Markov Decision Processes***

The control system goes from state to state depending on input data, as well as control algorithms. At the same time, it is extremely important to understand that it is not always possible how the system turned out to be in one state or another and what



**Fig. 1** Reinforcement learning process

exactly affected it? This is called the Markov decision-making process [6]. A Markov decision process is a 4-tuple  $(S, A, P_a, R_a)$ , where:

- $S$  is a finite set of states,
- $A$  is finite set of actions (alternatively,  $A_s$  is the finite set of actions available from state  $s$ ),
- $P_a(s, s') = \Pr(s_{t+1} = s' | s_t = s, a_t = a)$  is the probability that action  $a$  in state  $s$  at time  $t$  will lead to state  $s'$  at time  $t + 1$ ,
- $R_a(s, s')$  is the immediate reward (or expected the immediate reward) received after transitioning from state  $s$  to state  $s'$ , due to action (Fig. 1).

In the most general case, reinforced learning is somewhat similar to the optimal control problem, in which the function of the control object is known, it produces some effect on the object and the task is to find the very optimal effects. The basis in control theory is primarily the Pontryagin’s principle<sup>1</sup> and the Bellman’s optimality<sup>2</sup>. It states that it is necessary for any optimal control along with the optimal state trajectory to solve the so-called Hamiltonian system, which is a two-point boundary value problem, plus a maximum condition of the Hamiltonian. These necessary conditions become sufficient under certain convexity conditions on the objective and constraint functions.

Pontryagin’s maximum principle is used in optimal control theory to find the best possible control for taking a dynamical system from one state to another, especially in the presence of constraints for the state or input controls:

$$\hat{L}_u = 0 \tag{1}$$

<sup>1</sup>[https://en.wikipedia.org/wiki/Pontryagin%27s\\_maximum\\_principle](https://en.wikipedia.org/wiki/Pontryagin%27s_maximum_principle)

<sup>2</sup>[https://en.wikipedia.org/wiki/Bellman\\_equation](https://en.wikipedia.org/wiki/Bellman_equation).

where  $\widehat{L}_u$ —stationarity<sup>3</sup> by  $u$ . According to the Pontryagin’s maximum principle, the optimal control value is equal to the control value at one of the ends of the allowable range. The Pontryagin’s equations are written using the Hamilton function  $H$ , defined by the relation:

$$H = F(t, x(t), u) - \lambda(t)a(t, x(t), u) \quad (2)$$

Bellman’s principle of optimality or the dynamic programming method breaks this decision problem into smaller subproblems. Richard Bellman’s principle of optimality describes how to do this: An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision.

In reinforcement learning, the main difficulty is not finding optimal control, but to study and understand the environment. We already know the environment and the control object (in the theory of optimal control, this is called system identification), usually finding the optimal action is wrong so hard.

The key to reinforcement learning is the existence of a so-called reward function, as well as a state value function (value function,  $V(s)$ ). It will be the total expected reinforcement, which can be obtained by starting with this state. The essence of many teaching methods with reinforcement—in evaluating and optimizing the function of values. In fact, our task is to choose moves that lead to a state with the maximum value of  $V(s)$ . For Markov processes, we can formally determine:

$$V^\pi(s) = E_\pi[R_t | s_t = s] = E_\pi \left[ \sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s \right] \quad (3)$$

The  $\pi$  is the strategy that the agent follows. Strategy  $\pi$  is a function which for a given state  $s$  gives the probability distribution on the set action  $s$ . We will also denote by  $\pi(a, s)$  the probability of choosing the action of  $a$  in state  $s$ , and if we want to emphasize that the strategy is given parametrically with the parameter vector  $\theta$ , we write  $\pi(a, a; \theta)$ . If the strategy is deterministic, it just means that for each  $s$  all the probabilities  $s$  are equal to 0, except for one, which equal to 1.

## 2.2 Temporal Difference Learning (TD-Learning)

Almost all modern teaching approaches in deep learning are based on a very simple, but a very powerful principle called TD-learning<sup>4</sup>, from the words temporal difference. General TD Learning Principle<sup>5</sup> this is: let’s train states based on the grades we

<sup>3</sup>[https://en.wikipedia.org/wiki/Stationary\\_process](https://en.wikipedia.org/wiki/Stationary_process).

<sup>4</sup>[https://en.wikipedia.org/wiki/Temporal\\_difference\\_learning](https://en.wikipedia.org/wiki/Temporal_difference_learning).

<sup>5</sup><https://web.stanford.edu/group/pdplab/pdphandbook/handbookch10.html>.

have already trained for subsequent states. It is extremely important to understand that in this case we get the opportunity to use a neural network and deep learning just when working with incomplete, missing or incorrect data, which we will return to later.

Learning algorithm TD (0) for estimating  $V^\pi$  training looks like:

---

```

Initialize function V(s) arbitrarily
Initialize randomly strategy  $\pi$  to the policy to be evaluated
Repeat (for each episode):
    Initialize s
    Repeat (for each step of the episode):
        a ← action given by  $\pi$  for s
        Take action a; observe reward, r, and next
state, s'
        V(s) ← V(s) +  $\alpha[r + \gamma V(s') - V(s)]$ 
        s ← s'
Until s is terminal
    
```

---

where:

$$TD(\text{TemporalDifference}) = \gamma V(s') - V(s) \tag{4}$$

The key idea of using the TD method is that we use already trained patterns to search for even deeper patterns. First, we teach fortunes that already lead to the famous  $r$ . And then we use these states for teaching previously unknown states [9]. The TD method has various further modifications and extensions that have a significant impact on the quality of the algorithm.

### 2.3 Advantages of TD Prediction Methods

We list main advantages TD methods:

- TD methods do not require a model of the environment, only experience,
- methods can be fully incremental,
- less memory.

The culmination of the development of this approach is the teaching of the Q function, which is often called the Q-Learning. Here we immediately solve the Bellman equations:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha(r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)) \quad (5)$$

## 2.4 Deep Q-Networks

The key problem of managing complex objects is the large number of elements in the control loop and data. And if you imagine the real number of nodes with the number of possible states? There is no such opportunity to build, train and control so many values  $Q(s, a)$ . Usually, the following algorithm is used when teaching such deep models:

- Correlate inputs (states)  $s \in A$ , action  $a \in A$  in the form of some characteristic signs to reduce the dimension;
- Function  $Q(s, a)$  in which previously the values at different inputs were independent from each other, imagine as some kind of parametric model of a deep learning  $Q(s, a; \theta)$  at the input of which signs describing  $s$  and  $a$ ;
- $Q(s, a; \theta)$  complex function from signs to one real number;
- Learning inputs according to the TD-learning each successive transition  $(s_t, a_t, r_{t+1}, s_{t+1})$ .

The learning algorithm then looks as follows: Agent makes a move  $a$  out of state to  $s'$ , then receives a reward for this move  $r$ ; function training takes place  $Q(s, a; \theta)$  with input  $(s, a)$  and output  $(\max_{a'} Q(s', a', \theta) + r)$ , usually  $r = 0$ . The agent can also take such steps in relation to the previous positions, updating weights not only for the last entry, but also for several previous ones. An excellent first example in this regard was the Gammon TD network, which played essentially with itself without any learning set [10, 11]. This approach was developed and eventually became known as Deep Reinforcement Learning, and networks that are trained in this way are called Deep Q-networks (DQN) [12].

The learning algorithm of DQN is as follows at every step:

- Choose the next action  $a$  (in the  $\epsilon$ -greedy strategy we choose random action with probability  $\epsilon$  or  $a_t = \operatorname{argmax} Q(s_t, a; \theta)$ );
- Getting reward  $r_t$  and go to the next state  $s_{t+1}$ . Get a new unit important experience  $(s_t, a_t, r_t, s_{t+1})$ , which is stored in memory;
- Randomly select from memory a mini-batch of such units of experience for deep learning  $(s_j, a_j, r_j, s_{j+1})$ ;
- Count the network output  $y_j$ , take a step of gradient descent for the error function  $L = (y_j - Q(s_j, a_j, \theta))^2$ ; take a step of gradient descent:

$$\nabla_{\theta} L = 2(y_j - Q(s_j, a_j; \theta)) \nabla_{\theta} Q(s, a; \theta) \quad (6)$$



## 2.5 Algorithm

Reinforcement training can be divided into several almost independent parts that should share news with each other only at certain far enough apart moments' time, and between them can work completely parallel and independently:

- There is a central process, a server that stores current parameter values, updates them as necessary and distributes to everyone else;
- The first kind of processes is actually the “players” who interact with the outside world and gain new experience; they need from time to time receive updates from the server model parameters (they are used when choosing actions), and they themselves simply accumulate experience units in the form those fours  $(s_t, a_t, r_t, s_{t+1})$  and transmit the accumulated experience to the general memory storage;
- the second type of processes, “educators,” gain experience from the storage of memory in the form of mini-batches of experience units and consider the gradients of the error function; for this they need a network that generates target values, and the current one, so that “educators” are in closer contact with the server; but note that they are still completely independent, each of them considers his own gradient value and own custom updates for weights models;
- finally, the server itself collects all these updates, applies them to the stored he has models, and at some point (usually regular, but enough rare) distributes the updated model back to “players” and “educators”, and also updates the model that generates target values; it turns out that synchronization in such an architecture, of course, is needed, but it can be to do relatively rarely [11, 13].

## 3 Results

### 3.1 General Reinforcement Learning Architecture

Consider an architecture General Reinforcement Learning Architecture that is just right for tasks Deep Predictive Control. This architecture, shown in Fig. 2 contains the following components [14]:

**Actors.** Any reinforcement learning agent must ultimately select actions at to apply in its environment. We refer to this process as acting. This  $N_{\text{act}}$  corresponding instantiations of the same environment.

**Experience replay memory.** Experience tuples  $(s_t^i, a_t^i, r_t^i, s_{t+1}^i)$  generated by the actors are stored in a replay memory.

**Learners.** Each learner contains a replica of the Q-network, and its job is to compute desired changes to the parameters of the Q-network.

**Parameter server.** The general reinforcement learning architecture uses a central parameter server to maintain a distributed representation of the Q-network.

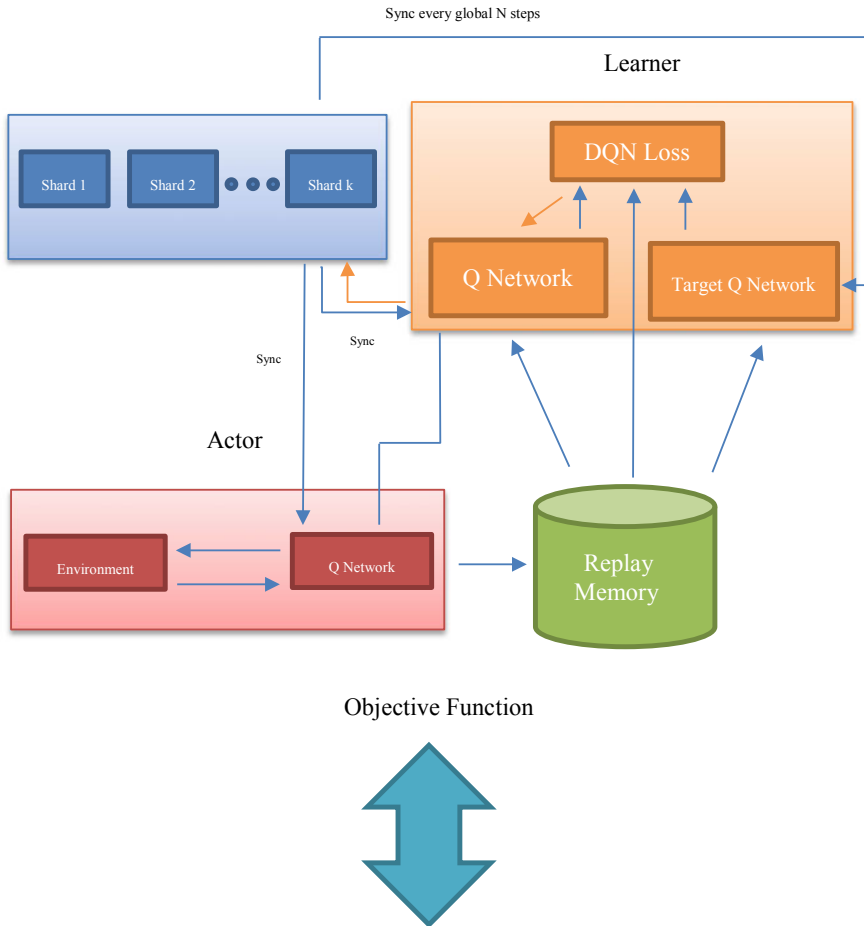


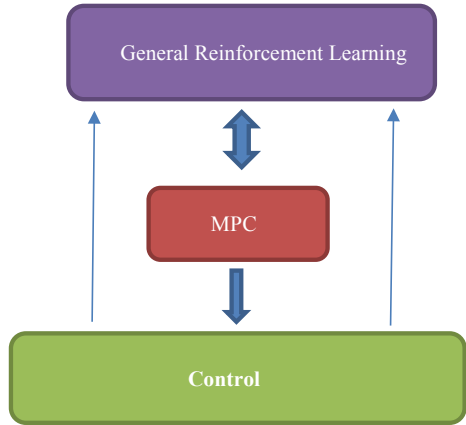
Fig. 2 General reinforcement learning architecture for control

**MPC.** Model predictive control [15].

### 3.2 Deep Predictive Control

Model predictive control (MPC) describes an advanced control method that has found a wide range of applications in industry and artificial intelligence [16]. MPC employs an explicit dynamic model of the plant to determine a finite sequence of control actions to take at each sampling time. MPC is widely used in various control tasks [17]. But this concept has, in our opinion, a number of significant drawbacks, for example, not very high performance. The key idea is to reduce the amount

**Fig. 3** Deep predictive control



of data processed by the module using deep learning with general reinforcement learning architecture. Thus, we can get a more efficient control system that will not be overloaded with unnecessary and not always useful data that comes online. The architecture of such a solution deep predictive control will look like this (Fig. 3):

## 4 Discussion

The considered technique offers a completely updated look at the control tasks, especially in real time. In this case, it is important to understand that the proposed concept proposes to consider the management task not only as a multi-criteria optimization task, which is often simply not solvable in a reasonable amount of time, but also use deep learning models to develop the deep predictive control concept [18].

## 5 Conclusion

The proposed deep predictive control architecture is planned to be used in the future to identify possible emergency situations during the operation of energy facilities. The key factor in this case is the reduction in computational complexity for calculating the control function, as well as the elimination of unnecessary or excessively noisy data, missing data for decision-making purposes [19].

**Acknowledgements** In the framework of this work, I would like to express my gratitude to the colleagues of the Institute of Computer Science and Technology of SPbPU Peter the Great, who provided comprehensive support in the framework of this work. We are at the beginning of a fairly large path, within which there will be a very significant change in concepts and approaches to managing complex objects and systems.

## References

1. Richter, C., Vega-Brown, W., Roy N.: Bayesian learning for safe high-speed navigation in unknown environments. In *Robot. Res.* pp. 325–341 (2018)
2. Erez, T., Lowrey, K., Tassa, Y., Kumar, V., Koley, S., Todorov, E.: An integrated system for real-time model predictive control of humanoid robots. In: *IEEE International Conference on Humanoid Robots*, pp. 292–299 (2013)
3. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., et al.: End to End Learning for Self-driving Cars 2016. <https://arxiv.org/abs/1604.07316> (дата обращения: 21.10.2019)
4. Vallon, C., Borrelli, F.: Task Decomposition for Iterative Learning Model Predictive Control. <https://arxiv.org/pdf/1903.07003v3.pdf> (дата обращения: 10.10.2019)
5. Cavagnari, L., Magni, L., Scattolini, R.: Neural network implementation of nonlinear receding-horizon control. *Neu. Comput. Appl.* **8**(1), 86–92 (1999)
6. Beneš, P.M., Cejnek, M., Kalivoda, J., Bukovsky, I.: Neural network approach to railway stand lateral skew control <https://arxiv.org/ftp/arxiv/papers/1402/1402.7136.pdf> (дата обращения: 10.24.2019).
7. Wang, Y., Zhang, J., Zhu, H.: Active neural network for learning higher-order non-stationarity from Spatiotemporal Dynamics. <https://arxiv.org/pdf/1811.07490v3.pdf> (дата обращения: 14.10.2019).
8. Pasini, M.L., Yin, J., Li, Y.W., Eisenbach, M.: A greedy constructive algorithm for the optimization of neural network architectures <https://arxiv.org/pdf/1909.03306v1.pdf> (дата обращения: 7.10.2019).
9. Mean S.: *Control Techniques for Complex Networks*. Cambridge University (2007) 615pp
10. Tesauro, T.: Temporal difference learning and TD-gammon. *Commun. ACM.* **38**(3), 58–68 (1995)
11. Николенко, С., Кадурич, А., Архангельская, Е.: Глубокое обучение. Погружение в мир нейронных сетей. Санкт-Петербург: Питер, 2018. 480 pp.
12. Mnih, V., et al.: Human-level control through deep reinforcement learning. *Nature* **518**(518), 529–533 (2015)
13. Nair, A., et al.: Massively Parallel Methods for Deep Reinforcement Learning URL: <https://arxiv.org/abs/1507.04296> (дата обращения: 11.10.2019).
14. Veerapaneni, R., Co-Reyes, J.D., Chang, M., Janner, M., Finn, C., Levine, S.: Entity abstraction in visual model-based reinforcement learning. In: *3rd Conference on Robot Learning (CoRL 2019)*. Osaka, Japan (2019)
15. Bradford, E., Imsland, L., Zhang, D., del Rio Chanona, E.A.: Stochastic data-driven model predictive control using Gaussian processes. <https://arxiv.org/pdf/1908.01786v1.pdf> (дата обращения: 28.09.2019).
16. Kamthe, S., Deisenroth, M.P.: Data-efficient reinforcement learning with probabilistic model predictive control. *AISTATS. Lanzarote, Spain*, vol. 87 (2018)
17. Amos, B., Jimenez Rodriguez, I.D., Sacks, J., Boots, B., Kolter, J.Z.: Differentiable MPC for end-to-end planning and control. <https://arxiv.org/pdf/1810.13400v3.pdf> (дата обращения: 16.10.2019)
18. Erickson, Z., Clever, H.M., Turk, G., Liu, C.K., Kemp, C.C.: Deep haptic model predictive control for robot-assisted dressing (2019). <https://arxiv.org/pdf/1709.09735v3.pdf> (дата обращения: 10.15.2019)
19. Franke, J.K.H., Koehler, G., Awad, N., Hutter, F.: Neural architecture evolution in deep reinforcement learning for continuous control <https://arxiv.org/pdf/1910.12824v2.pdf> (дата обращения: 11.10.2019)

# On the Computational Complexity of Deep Learning Algorithms



Dmitry Baskakov  and Dmitry Arseniev 

**Abstract** The paper analyzes current research and the state of the industry to assess the complexity of machine learning algorithms. The tasks of deep learning are associated with an extremely high degree of computational complexity, which requires the use, first of all, of new algorithmic methods and an understanding of the assessment of the complexity of the calculations. This area of research is not given due attention for various reasons, but primarily because of the novelty of this paradigm, as well as the use of other advanced methods, which is briefly analyzed in this paper.

**Keywords** Artificial intelligence · Fine-Grained reduction · Machine learning · Optimization

## 1 Introduction

Many books and articles have been written on algorithms for machine learning and deep learning [1]. Entire monographs are devoted to the description of different algorithms and methods for working with data of various types. Books of completely different levels are from fundamental works [2, 3]. An excellent example of research on the mathematical foundations of algorithms and asymptotic is the publication [4]. There are monographs that have already become classical in the theory of algorithms, their application and analysis [5]. Excellent works are devoted to the use of various algorithms with implementation in specific programming languages such as Java or C [6]. The presentation in each of the books differs both in level and in the availability of material starting from a fairly simple and elegant level [7] to very complicated and non-trivial level works [8]. Large research teams work in serious research centers in the field of algorithms and optimization <sup>1</sup>. But at least some serious research in the field of complexity of algorithms of deep and machine learning is not given so much

---

<sup>1</sup><https://www.mpi-inf.mpg.de/departments/algorithms-complexity/>.

---

D. Baskakov (✉) · D. Arseniev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [baskakov.de@edu.spbstu.ru](mailto:baskakov.de@edu.spbstu.ru)

time. This is due to a number of reasons, which include both the sufficient youth of the studied area and the current work, which are only in their infancy. It is important to note that in addition to algorithmic solutions for machine learning problems, technical properties in the form of libraries also appear CUDA<sup>2</sup>. Moreover, various advanced technical solutions such as supercomputers from NVIDIA DGX<sup>3</sup> give a multiple increase in productivity in deep learning problems, but do not conceptually solve the problems of the polynomial complexity of these calculations. The subject area under study is so vast and promising that in this short review it will be possible to show in the most general light those problems and possible solutions to the complexity of deep learning algorithms.

Machine learning consists of designing efficient and accurate prediction algorithms. As in other areas of computer science, some critical measures of the quality of these algorithms are their time and space complexity. But, in machine learning, we will need additionally a notion of sample complexity to evaluate the sample size required for the algorithm to learn a family of concepts. More generally, theoretical learning guarantees for an algorithm depend on the complexity of the concept classes considered and the size of the training sample [1].

It should be noted a very curious, but not a new trend in the analysis of the complexity of networks using topology [9]. This line of research seems very promising and is called topological data analysis (TDA). Note that the complexity of networks increases significantly from the amount of data, and at the same time, the computational complexity of the algorithms used increases.

Since the success of a learning algorithm depends on the data used, machine learning is inherently related to data analysis and statistics. More generally, learning techniques are data-driven methods combining fundamental concepts in computer science with ideas from statistics, probability and optimization [1]. The construction of an algorithm approximating rather complex objective functions in the mathematical plane is a very non-trivial task, which requires primarily high mathematical and algorithmic training [10]. It is extremely important to understand that the use of algorithms allows you to repeatedly improve the speed of solving a problem. Often the use of advanced and unique algorithms leads to a decrease [5]. Modern concepts are often used in deep learning, to which engineers have very serious expectations, for example parallel computing<sup>4</sup>. While sometimes the use of the law is completely overlooked Amdahl's law or Amdahl's argument<sup>5</sup>. Amdahl's law is often used in parallel computing to predict the theoretical speedup when using multiple processors. Amdahl's law can be formulated in the following way:

$$S = \frac{1}{(1 - P) + \frac{P}{N}} \quad (1)$$

---

<sup>2</sup><https://developer.nvidia.com/cuda-zone>.

<sup>3</sup><https://www.nvidia.com/en-us/data-center/dgx-1/>.

<sup>4</sup>[https://en.wikipedia.org/wiki/Parallel\\_computing](https://en.wikipedia.org/wiki/Parallel_computing)

<sup>5</sup>[https://en.wikipedia.org/wiki/Amdahl%27s\\_law](https://en.wikipedia.org/wiki/Amdahl%27s_law).

where  $S$ —is the theoretical speedup of the execution of the whole task,  $N$ —is the speedup of the part of the task that benefits from improved system resources and  $P$ —is the proportion of execution time that the part benefiting from improved resources originally occupied. Amdahl's law just shows that the possible benefit of using parallel computing is predetermined by the properties of the methods or algorithms used in the program [11]. In the conditions of a lack of new ideas, specialized calculators often become a factor of growth at present. The greatest success of Graphics Processing Unit (GPU)<sup>6</sup> is precisely due to the fact that not many new methods and approaches at least somehow bring us closer to solving NP (non-deterministic polynomial time) -complexity problems<sup>7</sup> [11]. GPUs are most effective in solving problems that have a high degree of parallelism for melons, the number of arithmetic operations in which is large in comparison with operations on memory. Deep learning algorithms relate to such tasks, which determine the high prevalence of using GPU when working with neural networks. The use of memory for storing data and calculation results is twofold and is often associated with an increase in the computational complexity of the algorithm used with the simultaneous disappearance of a number of tolerances and limitations that lead to less accurate results. The article discusses the main problems in the field of computational complexity of deep learning algorithms and discusses the main subject areas for overcoming it.

## 2 Materials and Methods

### 2.1 Deep Learning Complexity. Motivation

It is extremely important to understand that the modern theory of computational complexity of modern algorithms is, to some extent, an established scientific field with its own terminology, tasks, methods and approaches [12]. In the context of deep learning, we get several significant differences that are completely ignored sometimes. We note the most important from a computational point of view, problems and tasks that allow us to talk about the extremely high usefulness of this article:

- Deep learning algorithms have many hierarchies that require, perhaps, the use of several different asymptotic notations.
- Deep learning algorithms are often function of many variables. This requires the use of  $O$  notation as a function of many variables, thereby multiplying the complexity of these algorithms [12].
- In the case of using deep and machine learning algorithms, it makes sense to consider the complexity of such algorithms, in our opinion, as a function of *Data Complexity* (DC), and *Learning Complexity* (LC) [13].

---

<sup>6</sup>[https://en.wikipedia.org/wiki/Graphics\\_processing\\_unit](https://en.wikipedia.org/wiki/Graphics_processing_unit).

<sup>7</sup>[https://en.wikipedia.org/wiki/NP\\_\(complexity\)](https://en.wikipedia.org/wiki/NP_(complexity))

In this case, we move on to the computational complexity of the function of many variables in the following analytical form:

$$O(a) = DC(a) + LC(a) \quad (2)$$

Even with a cursory examination, you can notice that the complexity of deep learning algorithms depends directly on the type of data and the properties of their storage and directly on the learning algorithm itself.

The attached notation has some inaccuracies related to the fact that the amount of data involved in the training is not equal to the amount of data involved in the training. Given the possible pre-processing, the amount of data can be reduced by several times and many times, which allows you to significantly change the quality of the algorithm taking into account computational complexity.

## 2.2 Features of Deep Learning Algorithms

Deep learning algorithms include optimization in a wide variety of contexts. For example, to perform inference in models such as the principal component method, it is necessary to solve the optimization problem. We often use analytical optimization a proof for proving correctness or designing algorithms. Of all the many optimization tasks solved in deep learning are the most difficult arise when training a neural network. Often you have to spend on how many days to several months of work on hundreds of machines to solve everything one task of training a neural network. It is so expensive, special optimization methods have been developed.

It is important to note that the computational complexity of the deep learning algorithm is often a function of the architecture of the neural network. Such an architecture directly depends on the minimum number of weight parameters that need to be generalized by a neural network of power  $N$ :

$$N = O\left(\frac{W}{\varepsilon}\right) \quad (3)$$

where  $W$ —number of network parameters,  $\varepsilon$ —permissible part of incorrectly classified objects [14]. Note that there is a universal approximation theorem<sup>8</sup> that tells us about the number of hidden layers needed to approximate a sample of a given size. Below we list the main features of deep learning algorithms and the tasks that they must solve in theory and in practice:

- Ability to learn complex, highly varying functions, i.e., with a number of variations much greater than the number of training examples.

---

<sup>8</sup>[https://en.wikipedia.org/wiki/Universal\\_approximation\\_theorem](https://en.wikipedia.org/wiki/Universal_approximation_theorem).



- Ability to learn with little human input the low-level, intermediate, and high-level abstractions that would be useful to represent the kind of complex functions needed for Artificial Intelligence (AI) tasks.
- Ability to learn from a very large set of examples: computation time for training should scale well with the number of examples, i.e., close to linearly.
- Ability to learn from mostly unlabeled data, i.e., to work in the semi-supervised setting, where not all the examples come with the “right” associated labels.
- Ability to exploit the synergies present across a large number of tasks, i.e., multi-task learning. These synergies exist because all the AI tasks provide different views on the same underlying reality.
- In the limit of a large number of tasks and when future tasks are not known ahead of time, strong unsupervised learning<sup>9</sup> (i.e., capturing the statistical structure in the observed data) is an important element of the solution [15].

Calculation time is one of the most important indicators of deep learning algorithms and is much better if this time tends to linear approximation. One long-term goal of machine learning research is to produce methods that are applicable to highly complex tasks, such as perception (vision, audition), reasoning, intelligent control, and other artificially intelligent behaviors. We argue that in order to progress toward this goal, the machine learning community must endeavor to discover algorithms that can learn highly complex functions, with minimal need for prior knowledge, and with minimal human intervention [16]. We present mathematical and empirical evidence suggesting that many popular approaches to nonparametric learning, particularly kernel methods, are fundamentally limited in their ability to learn complex high-dimensional functions [16].

Deep learning algorithms should have some key features that distinguish them qualitatively from traditional computational algorithms:

1. *Computational complexity.* Number computations (training, hyperparameter optimization<sup>10</sup>, recognition).
2. *Statistical properties.* Examples for training should have generalization properties in the theoretical part and have markup properties at least in a limited area.
3. *Human involvement.* A person must possess the ability to work with this algorithm, be able to use it, generalize. In this case, it is important to consider the amount of human labor that will participate in the operation of the algorithm.

To create a high-quality algorithm of deep learning, work and research should be concentrated in the following critical areas and have the following properties:

- The deep learning algorithm should be flexible and fast enough, able to work with a large number of architectures and data.
- The algorithm must be able to work with deep architectures with many levels and concepts.

---

<sup>9</sup>[https://en.wikipedia.org/wiki/Unsupervised\\_learning](https://en.wikipedia.org/wiki/Unsupervised_learning).

<sup>10</sup>[https://en.wikipedia.org/wiki/Hyperparameter\\_optimization](https://en.wikipedia.org/wiki/Hyperparameter_optimization)

- A deep algorithm should be able to work with many functions with millions and even more parameters.
- A learning algorithm that can be trained efficiently even when the number of training examples becomes very large. This excludes learning algorithms requiring to store and iterate multiple times over the whole training set, or for which the amount of computations per example increases as more examples are seen. This strongly suggests the use of online learning [16].
- A learning algorithm that can discover concepts that can be shared easily among multiple tasks and multiple modalities (multi-task learning) and that can take advantage of large amounts of unlabeled data (semi-supervised learning) [16].

### 2.3 Universal Approximation Theorem

The universal approximation theorem can be expressed mathematically. Let  $\varphi : \mathbb{R} \rightarrow \mathbb{R}$ , be a non-constant, bounded, and continuous function (activation function). Let  $I_{m_0} - m_0$ —dimensional hypercube  $[0, 1]^{m_0}$ . Then the space of real-valued continuous function on  $I_{m_0}$  is denoted by  $C(I_{m_0})$ . If any  $\varepsilon > 0$  and any function  $f \in C(I_{m_0})$ , there exist an integer  $m_1$ , real constants  $a_i, b_i \in \mathbb{R}$  and real vectors  $\omega_{i,j} \in \mathbb{R}^{m_0}$  for  $i = 1, \dots, m_1$  and  $j = 1, \dots, m_0$ , such that we may define:

$$F(x_1, \dots, x_{m_0}) = \sum_{i=1}^{m_1} a_i \varphi \left( \sum_{j=1}^{m_0} \omega_{i,j} + b_i \right) \quad (4)$$

As an approximate realization of the function  $f$ ; that is,

$$|F(x) - f(x)| < \varepsilon \quad (5)$$

For all  $x_1, \dots, x_{m_0}$  from any dense  $D = [x_1, \dots, x_{m_0}, f(x_1, \dots, x_{m_0})]$ . According to this neural network theorem with one hidden size layer  $m_0$  and output layer size  $m_1$  with sigmoid activation functions are sufficient to approximate the free sampling with any accuracy. However, this theorem only speaks of approximation and says nothing about the specific size of the hidden layer. To estimate the size of the hidden layer, we recommend using the classical Theorem Barron [17]. Key to the advantageous approximation and estimation properties of artificial neural networks is the fact that the model is not linear in all its parameters (activation weights). The adjustment of the scale, direction, and location parameters of the sigmoidal basis functions permits them to be adapted to the estimation of the target function. Nonlinear adjustment of sinusoidal, polynomial, spline, and wavelet basis functions is also possible, and it is anticipated that similar approximation and estimation bounds can be obtained in each of these cases by the same technique as used here for sigmoidal basis functions [17]. But Theorem Andrew R. Barron does not say anything about the generalizing abilities constructed in such a way neural network at once. The problem of building

a network with optimal generalizing ability is considered in papers [18]. In this case, the use of greedy algorithms is very often [19].

## 2.4 Training Features

We define here the features which could influence the prediction of execution times when performing training and categorize these features into layer features, layer specific features, implementation features and hardware features. Each of these categories can contain an almost endless list of features. As such we define here a core subset of those features but argue that other features could easily be added. A full analysis of all available features and the impact they have on the accuracy of prediction is beyond the scope of this work [20].

### *Layers Features*

These relate to those features of a particular layer within the neural network and in particular to the hyperparameters related to that layer. These include, but are not limited to:

- Activation function. These may include ReLU, softmax, sigmoid and tanh.
- Optimizer (Adadelta, Gradient Descent, Adagrad, Momentum, Adam and RMS Prop).
- Batch size (number of training samples which are processed together as part of the same batch).

### *Layer Specific Features*

Here we discuss the features which are unique to a particular type of layer within the neural network:

- Number of inputs to the layer. There can be many layers; each layer can contain many inputs.
- Numbers of outputs within this layer which is equivalent to the number of neurons of the layer.
- Matrix size is the size of the input data to be trained on.
- Kernel size representing the size of the filter applied, for example, to the image data.
- Input and output depths are the number of channels or payers in the input and output data.
- Hidden layers.
- Neural network type.

### *Hardware Features*

- GPU technology.
- GPU count the number of GPUs in the system.

- Memory.
- GPU clock speed.
- GPU core count is the number of processing units.
- GPU peak performance.

Certainly, this list is not exhaustive. Nevertheless, all these parameters and conditions have an all-sided impact on deep learning algorithms and its final computational complexity.

## ***2.5 Shallow Architectures, Deep Architectures, and Learning Model***

In machine learning, two types of architectures are currently most widely used, which we will consider below.

### ***2.6 Shallow Architectures***

The best example of shallow architecture would be, for example, modern kernel machines [21] or Support Vector Machines [22]. Such architectures consist of one layer. Further, taking into account some mathematical operations, we transform the input data into some result, which is both a linear combination of the input data and possibly a more complex mathematical transformation. The only components subject to supervised training are the coefficients of the linear combination [16].

### ***2.7 Deep Architectures***

Deep architectures are perhaps best exemplified by multi-layer neural networks with several hidden layers. In general terms, deep architectures are composed of multiple layers of parameterized nonlinear modules [16]. While shallow architectures have advantages, such as the possibility to use convex loss functions, we show that they also have limitations in the efficiency of the representation of certain types of function families. Although a number of theorems show that certain shallow architectures (Gaussian kernel machines, 1-hidden layer neural nets, etc.) can approximate any function with arbitrary precision, they make no statements as to the efficiency of the representation. Conversely, deep architectures can, in principle, represent certain families of functions more efficiently (and with better scaling properties) than shallow ones, but the associated loss functions are almost always non-convex.

## 2.8 Learning Model

The training model should include three main points:

- The representation of the data: pre-processing, feature extractions, etc.
- Optimization functions, objective functions, parameters that must be achieved in the learning process.
- The loss function, regularize, hyperparameters.

## 2.9 Kolmogorov Complexity

For example, consider a Turing machine  $U$  with input alphabet  $\{0, 1\}$  and tape alphabet  $\{0, 1, \_ \}$ , where  $\_$  is the blank symbol. A binary string  $p$  is a program for the Turing machine  $U$ , if and only if  $U$  reads the entire string and halts [13]. For a program  $p$ , we use  $|p|$  to denote its length in bits, and  $U(p)$  the output of  $p$  executed on the Turing machine  $U$ . It is possible to have an input string  $x$  on an auxiliary tape [13]. In that case, the output of a program  $p$  is denoted as  $U(p, x)$ . The Kolmogorov complexity measures the algorithmic complexity of an arbitrary binary string  $s$  by the length of the shortest program that outputs  $s$  on  $U$ . Kolmogorov complexity  $K_U(s)$  is defined as:

$$K_U(s) = \min\{|p| : U(p) = s\} \quad (6)$$

Kolmogorov complexity can be regarded as the length of the shortest description or encoding for the string  $s$  on the Turing machine  $U$ . Conditional Kolmogorov complexity measures how many additional bits of information are required to generate  $s$  given that  $x$  is already known. The Kolmogorov complexity is a case of the conditional one where  $x$  is empty [13]. The Kolmogorov complexity  $K(s)$  is a universal measure for the amount of information needed to replicate  $s$ .

The conditional Kolmogorov complexity  $K(s|x)$  is defined as the length of the shortest program that outputs  $s$  given the input string  $x$  on the auxiliary tape:

$$K(s|x) = \min\{|p| : U(p, x) = s\} \quad (7)$$

## 2.10 Learning and Optimization. Differences

The last problem is especially but difficult because effective capacity is limited by algorithm capabilities optimization, and we have few theoretical results on the general problems of non-convex optimization found in deep learning.

The optimization algorithms used to train deep models differ from traditional optimization algorithms in several respects. Machine learning usually doesn't work directly, which is a measure of quality  $P$ , which is defined relative to the test set and can to be computationally impregnable. Therefore, we optimize  $P$  indirectly [23]. From a mathematical point of view, these nuances are extremely important, because all this very seriously affects the final performance of deep learning algorithms [24]. Main idea—we reduce another cost function  $J(\theta)$  in the hope that  $P$  will also improve. This is very different from pure optimization, where minimization of  $J$  is the final target. In addition, optimization algorithms for teaching deep models usually include specializations for a specific structure of objective functions. In theory, all this should greatly facilitate multi-criteria optimization tasks, but often this is not at all. A typical objective function can be represented as the average of the training set:

$$J(\theta) = E_{(x,y) \sim \hat{p}_{\text{data}}} L(f(x; \theta), y) \quad (8)$$

where  $L$ —loss function in one example,  $f(x; \theta)$ —predicted output for input  $x$  and  $\hat{p}_{\text{data}}$ —empirical distribution. Case of study with supervised learning  $y$ —label associated with input [25].

Equation (1) determines the objective function relative to the training set. But we usually prefer to minimize the corresponding objective function, in which the mathematical expectation is taken from the  $p_{\text{data}}$  distribution generating the data, and not just from the final training set:

$$J^*(\theta) = E_{(x,y) \sim p_{\text{data}}} L(f(x; \theta), y) \quad (9)$$

It is in such a transition that the main computational difficulties of deep learning algorithms can lie in the most general case.

## 2.11 Fine-Grained Reduction

Complexity theory traditionally distinguishes whether a problem can be solved in polynomial-time (by providing an efficient algorithm) or the problem is NP-hard (by providing a reduction). For practical purposes however, the label “polynomial-time” is too coarse: It may make a huge difference whether an algorithm runs in say linear, quadratic, or cubic time. In this course, we explore an emerging subfield at the intersection of complexity theory and algorithm design which aims at a more fine-grained view of the complexity of polynomial-time problems. We present a mix of upper and lower bounds for fundamental polynomial-time solvable problems, often by drawing interesting connections between seemingly unrelated problems.<sup>11</sup>

---

<sup>11</sup><https://www.mpi-inf.mpg.de/departments/algorithms-complexity/teaching/summer16/poly-complexity/>.

## 2.12 Empirical Risk Minimization

Empirical risk minimization (ERM) has been highly influential in modern machine learning. ERM underpins many core results in statistical learning theory and is one of the main computational problems in the field [26]. Several important methods such as support vector machines (SVM), boosting, and neural networks follow the ERM paradigm [27]. In this brief work, we will not dwell on this issue in detail because there are a sufficient number of approaches and methods that allow us to solve these problems, thereby optimizing the very computational complexity of the deep learning algorithms.

## 3 Results

### 3.1 Decisions

Let us suppose a part of solutions that can significantly reduce the computational complexity of deep learning algorithms and increase the efficiency of the neural network as a whole.

**Background information.** Deep knowledge and understanding of the subject area, the study of modern and promising methods and approaches allows even in the existing technological stack to solve problems at an outstanding level without involving expensive solutions.

**Cooperating learning algorithms.** Firstly, it is necessary to involve industry experts in the subject area. Experts often seem able to greatly reduce the time required to learn by communicating and working together in groups. It would be interesting to define a formal model of learning algorithms that are allowed to communicate their hypotheses and/or other information in an attempt to converge on the target more rapidly [28].

**Learning more expressive representations.** Data preparation, pre-processing, in-depth examination, understanding of the essence of what is happening and so on allow you to qualitatively change the approach to the subject area being solved and may even completely abandon any bulky and difficult decisions [29].

**Learning with mistakes.** The use of deliberate errors in the data at the stage of testing and preliminary fitting of hyperparameters allows you to quickly identify the weaknesses in the network and deep learning models, which avoids significant difficulties in the future [30].

## 4 Discussion

The complexity model of deep learning algorithms is a fairly modern problem, the solution of which lies in many disciplinary planes. As part of a small material, it is extremely difficult to talk about all the principles and approaches in this direction. Nevertheless, it is important to understand that the development of this area is a key task for the direction of deep learning in general.

## 5 Conclusion

The key idea of writing this short review was that the problem of the complexity of deep learning algorithms is very non-trivial in nature, and many researchers for various reasons get around. We tried to demonstrate that this task is extremely promising both in terms of theory and in terms of the development of further hardware solutions. In our opinion, researchers should apply various techniques and approaches described in this paper in practice. Personally, it seems that the use of fine-grained reduction helped to significantly solve some of the problems, especially from the point of view of constructing optimization algorithms on graphs. After all, as you know, many control and optimization problems are well reduced to tasks on the graph, some of which are related to NP-complexity [31].

**Acknowledgements** We thank the staff of SPbPU Peter the Great and the Institute of Computer Science and Technology for their support in the preparation of this material. We drew thoughts and ideas at joint seminars within the framework of the institute, as well as during fruitful communication with colleagues.

## References

1. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning, p. 427. The MIT Press, Cambridge, Massachusetts, London, England (2012)
2. Knuth D.E.: The Art of Computer Programming. 3rd ed. vol 1. Addison-Wesley (2013) 672pp
3. Knuth, E.D.: The Art of Computer Programming, Seminumerical Algorithms, 3rd edition ed., vol 2. Addison-Wesley, 762pp
4. Graham, R.L., Knuth, D.E., Patashnik, O.: Concrete Mathematics. 2nd ed. Addison-Wesley Publishing Company (1994) 604pp
5. Cormen, T.H., Leiserson, C.E., Rivest, R.L.: Introduction to Algorithms, 3rd ed. The MIT Press, 1292pp
6. Sedgewick, R., Wayne, K.: Algorithms. Addison-Wesley (2013), 848pp
7. Dasgupta, S.: Algorithms, 1st ed. McGraw-Hill Education, 336pp
8. Knuth, D.E.: The Art of Computer Programming. Generating All Trees-History of Combinatorial Generation, vol. 4 (2007), 160pp
9. Rieck, B., Togninalli, M., Bock, C., Moor, M.: Neural persistence: a complexity measure for deep neural networks using algebraic topology. ICLR (2019)



10. Williamson, D.P., Shmoys, D.B.: *The Design Approximation Algorithms*. Cambridge University Press (2010)
11. Боресков, А.В., Харламов, А.А., Марковский, Н.Д., Микушин, Д.Н., Мротников, Е.В., Мылльцев, А.А., Сахарных, Н.А., Фролов, В.А.: *Параллельные вычисления на GPU. Архитектура и программная модель CUDA*, Издательство Московского университета (2012)
12. Graham, R.L., Knuth, D.E., Patashnik, O.: *Concrete Mathematics*. Reading. Addison-Wesley Publishing Company, Inc., Massachusetts (1994, 1989), 604pp
13. Li, L., Abu-Mostafa, Y.S.: *Data Complexity in Machine Learning*, Caltech Computer Science Technical Report CaltechCSTR:2006.004, May 2006 (2006)
14. Widrow, B.: *Adaptive Signal Processing*, 1st ed. (1985), 486pp
15. Bengio, Y.: *Learning Deep Architectures for AI*, Technical Report 1312. [Online]. <https://www.imo.umontreal.ca/~lisa/pointeurs/TR1312.pdf>
16. Bengio, Y., LeCun, Y.: *Scaling Learning Algorithms towards AI*. In: *Large-Scale Kernel Machines*, P. 408. The MIT Press (2007)
17. Barron, A.R.: *Approximation and estimation bounds for artificial neural network*. *Mach. Learn.* **14**, 115–133 (1994)
18. Hassibi, B., Stork, D.G.: *Second order derivatives for network pruning: optimal brain surgeon*. *Adv. Neural Inf. Proc. Syst.* **5**(NIPS 1992) (1992)
19. Bengio, Y., Lamblin, P., Popovici, D., Larochelle, H.: *Greedy layer-wise training of deep networks*. <https://papers.nips.cc/paper/3048-greedy-layer-wise-training-of-deep-networks.pdf> (дата обращения: 15.09.2019)
20. Justus, D., Brennan, J., Bonner, S., McGough, A.S.: *Predicting the computational cost of deep learning models*. <https://arxiv.org/pdf/1811.11880.pdf> (дата обращения: 1.11.2019)
21. Scholkopf, B., Christopher, Smola, A.J.: *Advances in Kernel Methods*. The MIT Press (1999)
22. Corinna Cortes, Vladimir Vapnik. *Support-Vector Networks*. <http://homepages.rpi.edu/~bennek/class/mmlnd/papers/svn.pdf> (дата обращения: 02.10.2019)
23. Dall'Anese, E., Simonetto, A., Becker, S., Madden, L.: *Optimization and Learning with Information Streams: Time-varying Algorithms and Applications*. <https://arxiv.org/pdf/1910.08123v1.pdf> (дата обращения: 22.10.2019)
24. Colin, I., Dos Santos, L., Scaman, K.: *Theoretical limits of pipeline parallel optimization and application to distributed deep learning* [http://arxiv-sanity.com/search?q = learning+and+optimization](http://arxiv-sanity.com/search?q=learning+and+optimization) (дата обращения: 15.10.2019)
25. Liu, S., Vicente, L.N.: *The stochastic multi-gradient algorithm for multi-objective* The stochastic multi-gradient algorithm for multi-objective. <https://arxiv.org/pdf/1907.04472v2.pdf> (дата обращения: 28.09.2019)
26. Backurs, A., Indyk, P., Schmidt, L.: *On the fine-grained complexity of empirical risk minimization: kernel methods and neural networks*. <https://papers.nips.cc/paper/7018-on-the-fine-grained-complexity-of-empirical-risk-minimization-kernel-methods-and-neural-networks.pdf> (дата обращения: 02.10.2019)
27. Shalev-Shwartz, S., Ben-David, S.: *Understanding Machine Learning. Cambridge University Press, From Theory to Algorithms* (2014)
28. Fujimoto, S., Conti, E., Ghavamzadeh, M., Pineau, J.: *Benchmarking Batch Deep Reinforcement Learning Algorithms*. <https://arxiv.org/pdf/1910.01708v1.pdf> (дата обращения: 10.7.2019)
29. Glasser, I., Sweke, R., Pancotti, N., Eisert, J., Cirac, J.I.: *Expressive power of tensor-network factorizations for probabilistic modeling*. <https://arxiv.org/pdf/1907.03741v1.pdf> (дата обращения: 03.10.2019)

30. Sheriff, M.R., Chatterjee, D.: Dictionary Learning With Almost Sure Error Constraints. <https://arxiv.org/pdf/1910.08828v1.pdf> (дата обращения: 10.17.2019)
31. Jungnickel, D.: Graphs, Networks and Algorithms. 3rd ed. Springer (2008)

# Enactivism in the Conceptual Basis of the Non-classical Theory of Management of Ergatic Systems



Sergey Sergeev, Vladimir Ivanov , and Oleg Ipatov

**Abstract** The article discusses the problems of managing complex ergatic systems containing symbiotic and environment-oriented forms of control and orientation by interacting agents. The difference in the control methods of complex ergatic systems created in the framework of classical and non-classical ergonomics is shown. The conceptual basis of non-classical and post-non-classical ergonomics is presented. The article discusses the prospects of using the concept of enactivism in the conceptual basis of ergatic systems management, which allows implementing a project of continuously updated ergatic environment, which focuses on the processes of continuous updating, operational control and correction of the parameters of the technical and human parts of the system, taking into account the cyclic processes of self-organization in the actor's environment of the acting subject achievement of business goals. The forms and properties of intelligent entities embodied in organized environments are presented. A number of general definitions of intelligence and intelligent symbionts are embodied in existing ergatic systems that arise in the process of combining artificial and natural self-organizing systems of the environment of activity. The prospects of using conceptual representations of enactivism and constructivism in the management of complex ergatic systems are shown.

**Keywords** Complex systems · Non-classical management · Post-non-classical management · Intelligent symbionts management · Ergonomics of immersive environments · Enactivism · Ergonomics

## 1 Introduction

The term “enactivism” is close in meaning to “enaction,” defined as “the manner in which a subject of perception creatively matches its actions to the requirements of its situation” [1]. The development of anthropogenic civilization and the related introduction of computer and network control systems in all spheres of human life

---

S. Sergeev · V. Ivanov (✉) · O. Ipatov  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [ssfpost@mail.ru](mailto:ssfpost@mail.ru)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021  
N. Voinov et al. (eds.), *Proceedings of International Scientific Conference on Telecommunications, Computing and Control*, Smart Innovation, Systems and Technologies 220, [https://doi.org/10.1007/978-981-33-6632-9\\_31](https://doi.org/10.1007/978-981-33-6632-9_31)

357

lead to the emergence of a complex world problem that fundamentally changes the methodology for managing complex systems that implement intelligent control functions in the form of symbiotic interaction with operators/users. The problems of interacting intelligent systems that solve the target tasks of managing elements of the technological environment are solved at the border of engineering and psychological disciplines within the framework of engineering psychology and ergonomics (discipline for the consideration of the human factor).

The objects of study of classical engineering psychology and ergonomics are the “man–machine interaction” (“man–machine interaction”–environment) systems considered in the paradigm of informational interaction of a person with a machine controlled by him in the environment and conditions of professional activity [2]. The subject of these disciplines includes all forms of human interactions with the world, mediated by technique and technology with the goal of creating and operating efficient ergatic systems and environments. The main task of the classical disciplines in taking into account the problems of the human factor is to ensure maximum management efficiency in the system by distributing functions between the person and the cybernetic part of the system and providing an informational basis for decision making [3, 4]. The implementation of this task is considered in the framework of the opposition “man–environment,” where the leading role is played by the adaptive properties of man, which are used in the implementation of management functions in the technical system. At the same time, a person’s mental properties and features of consciousness are considered only through the prism of his behavior and experience and are interpreted as information–physical interactions between elements of the ergatic system and the environment [4]. This approach limits the consideration of the ergatic environment to the properties of the human consciousness with the objective world reflected in it in a subjective form. Classical ergonomics is the ergonomics of common sense, subjective reality, arbitrary and simplified cognitive interpretations of the contents of everyday consciousness. Naturally, the effectiveness of this discipline in its classical version decreases due to the appearance of complex communication systems and control modes, which are fundamentally irreducible by consciousness. A barrier of subjective complexity arises, which does not allow the operator to conduct adequate activities for the adoption and implementation of reasonable and adequate decisions.

## 2 Materials and Methods

The main management methods within the framework of classical ergonomics are associated with the solution of the problem of relations “subject–object of management” and are associated with the classical theory of activity (Vygotsky L.S., Leontyev A.N., Rubinstein S.L.). In this case, activity is understood as a specific type of human activity aimed at cognizing and creatively transforming the surrounding world and oneself. The presence in the human mind of a neuropsychological model of the external world is postulated in accordance with which the governing structural

components of activity—skills, abilities, and knowledge—are built. The theory of activity has a number of limitations associated with not taking into account in its postulates the self-organizing nature of the work of the organism and the human psyche, which leads to the problem of reduction of activity restricting the work of the human operator in complex control contexts [5]. Control in the classical control paradigm is formed on the basis of the implementation of programs (algorithms) that correct deviations of the system from a given state using feedbacks [6].

## ***2.1 Non-classical Management and Ergonomics of Immersive Systems and Environments***

The advent of non-classical ergonomics was a natural reaction of the scientific community to the emergence of new objects of engineering psychology and ergonomics, including complex technogenic environments endowed with artificial intelligence and global communication information networks. Their functioning cannot be described in the classical language of causal relationships and information interactions within the framework of distinguished hierarchical systems. The management paradigm is changing to a new non-classical form of scientific rationality [7]. It includes the interaction of the individual with people, given their self-organizing nature. The subject-subject relationships are considered, and in ergonomics, the communication, social, and symbiotic interactions of a human operator with similar and artificial intelligence systems [8].

Non-classical ergonomics of immersive environments, based on the ideas of radical and social constructivism, synergetic, second-order cybernetics, autopoietic self-organization and complexity theory (E.N. Knyazeva, M. Eigen, K.J Gergen, E. von Glasersfeld, H. Haken, H. von Foerster, N. Luhmann, H. Maturana, I. Prigogine, F. Varela, C.H. Waddington, P. Watzlawick) [4].

The following views on the concept of “environment” are used:

- the environment of the ergatic system is a product of the constructive activity of the psyche of the human operator and cannot be considered outside its mental content;
- the environment reflects the phenomenon of the dynamic integrity of cyclically forming chains of a person’s relationship with physical and social realities in the process of ensuring his life, speaks to the subject simultaneously in the form of subjective reality and as an external objective, objective structure of the world in which the subject operates.

In this process, selectively, in the logic of reflexive consciousness, various elements of the external and/or internal environment are involved in order to ensure: autopoiesis of the body, personality stability and the continuity of its history.

The concept of “knowledge” in non-classical ergonomics also makes sense different from the concepts of “knowledge” adopted in traditional instrumental theories:

- knowledge, unlike information, cannot be extracted from a person in whom it exists in an implicit form; it cannot be transmitted directly from person to person;
- knowledge originates and develops with a person, improves in the process of life, acquires properties that take into account the experience of the subject;
- knowledge does not have a material form, operations similar to operations with physical, material objects are not applicable to it;
- knowledge is associated with the work of the mechanism of understanding;
- knowledge bears the features of a social construct, reflecting the interpretations generated and shared by members of society;
- language acts as a means of constructing knowledge [9, 10].

## ***2.2 Post-non-classical Management in Ergonomics***

Post-non-classical rationality forms the basis of post-non-classical science, including complex self-organizing evolving systems in its consideration [11]. In post-non-classical science, various scientific theories (understood as models and subjective realities) constitute a conceptually interconnected network of self-organizing systems. This ensures the synergistic effect of applying the methodological principles of subjectivity to the tasks of subject-oriented design of self-developing polysubject media that create various dynamic contexts that control the behavior of control subjects. As an example of the post-non-classical methodology, we can consider the technologies of “controlled chaos” oriented to control through the environment with the goal of destroying human communications, deformation of the subject medium [12].

In the context of post-non-classical scientific rationality, new disciplines are considered, and in particular, neocybernetics or cybernetics of the second order, which is associated with the processes of controlled self-organization and is focused on the development of a methodology for the formulation and solution of problems of analysis and synthesis of intelligent processes and control systems of complex objects of arbitrary nature with selectivity and operational closeness [13]. Introduction to the ergonomics of the concepts of cyclic self-organization allows you to expand the methodological capabilities of this discipline in relation to objects of organized complexity, which include ergatic systems, social communication, group and collective operator activities, etc.

### ***2.3 Enactivism in the Basis of Management in the Disciplines of the Human Factor***

The main theoretical concepts of enactivism were formulated by F. Varela, E. Roche, and E. Thompson in the book “The Embodied Mind” (1991) [14]. Enactivism is a holistic form of views on human cognition, the activity of a cognitive agent and is the development of the ideas of radical and epistemological constructivism, evolutionary epistemology. The subject and object, body and mind, organism and environment, life and cognition, real and virtual, are in mutual cyclic determination, condition each other, make up a single process, in which both these sides are involved each time [15].

## **3 Results**

The subject of cognition, or the cognitive agent, is seen as active and interactive. It actively integrates into the environment; its cognitive activity is accomplished through its active “incorporation” into the environment or its activation. Cognition, perception, thinking, and imagination are associated with action. In this concept, a holistic picture of cognitive processes is built, in which the brain as a part of the body, the body as an instrument of cognition, seeking and knowing the mind and the environment it knows, are considered in a mutually conditioning bunch of autopoietic systems that are mutually oriented with respect to each other. A significant role in the processes of self-organization of the human psyche is played by the mechanism of consciousness, which provides reflection processes in the cycles of self-organization of complex systems [16]. Thus, we can consider human consciousness, the human mind as an organizing force due to the features of its functioning as a self-developing historical system. Such representations blur the line between subject and object, internal and external.

Enactivism allows introducing the concept of “continuously updated ergatic environment” into the design process of ergatic complexes, which focuses on the processes of continuous updating, operational control and correction of the parameters of the technical part of the system, taking into account the cyclic processes occurring in the actor’s environment of the acting subject when achieving the goals of the activity. The stage of ensuring the inactivation and maintenance of the cognitive subject in the process of its integration into the environment constructed by him is introduced. The role of artificial intelligence in the processes of optimizing the operator’s environment is emphasized.

Enactivism allows us to consider mental functions as embodied in the human body and at the same time independent entities existing in it. For example, intelligence loses its specificity of a purely human property and can be represented as an emergent property embodied in a complex system; we can talk about the distribution of intelligence in a complex environment, its artificial and diffuse nature [17, 18]. Table

**Table 1** Forms of intelligent entities embodied in organized environments

Intelligence education	Relations between components	Activity and control center, mechanism	Relations with the environment, boundaries
Natural intelligence	Self-organization autopoietic system	Consciousness, the ego system of man	Continuous active transformation of the world, borders are dynamically changing
Artificial intelligence	Fixed or variable firmware	Program, algorithm in a environment yet structured or in environment being in a process of structuring	The algorithm is implemented, situational control, fixed boundaries
Hybrid intelligence	Symbiosis, adaptation of organized and autopoietic components to the environment, associations at the macro-level with the priority of consciousness	Person in a structured environment	Mutual adaptation of natural and artificial intelligence, boundaries are variable
Diffuse intelligence	Selective communication at all levels of an autopoietically organized and organized environment and human	Arises in an organized environment	Synergetic association, borders are formed for the task

1 presents intelligent formations as properties inherent in organized environments [19].

## 4 Discussion

Enactivism uses the metaphor of embodiment and continuous dynamic integration of the system into the environment. Cognition is a form of active construction, a constant search based on sensory-motor contacts of a person with the world [20]. All this distinguishes this concept from the popular cognitive model considering computer metaphors in the work of the human brain and the activity of consciousness based on rules and logical inference. In terms of enactivism, knowledge of the system reflects its current repertoire of possible actions. In this case, the action covers not only physical, but also as a subset of its mental activity.



## 5 Conclusion

A number of general definitions of intelligence and intelligent symbionts can be given, embodied and acting in an ergatic system, arising in the process of combining artificial and natural intelligences and the environment of activity.

- Intelligence is a form of active self-organization of a complex system, involving the user immersed in the environment in creating changes.
- Intelligence is associated with the environment as a mechanism of its organization, providing processes of self-organization of the system endowed with it.
- Intelligence is distributed in the “system-environment” continuum and is embodied in the cycles of self-organization of a system operating in the environment.
- Natural intelligence is organizing complexity in an organized environment, and artificial intelligence is organized complexity in an organized environment.
- Hybrid and diffuse intelligences are symbionts, including the organizing and organized complexity of systems in their synergistic and symbiotic interactions as a tool for achieving an actor’s goal in organized and organized environments.
- Intelligence reflects the results of selection and application of a self-organizing system of effective ways to achieve goals in an organized environment.

When creating complex ergatic systems, it is necessary to take into account the emergent properties that arise and exist due to the complex organization of the environment. These are the effects of intellectualization, the emergence of cooperative and hybrid forms of combining the cognitive mechanisms of a person and the intellect and its symbiotic forms distributed among the environment. The inclusion in a complex technogenic environment of a person is also associated with the effects of techno-modification of his personality and cognitive systems, which leads to the emergence of techno-psychological symbionts into which resources sufficient to achieve the goals of the system are activated [21].

## References

1. Protevi, J. (ed.): «Enaction». A Dictionary of Continental Philosophy. Yale University Press, pp. 169–170 (2006)
2. Lomov, B.F.: Man and technology. Essays on Engineering Psychology. Publishing House Soviet Radio, Moscow (1966)
3. Sergeev, S.F.: Engineering Psychology and Ergonomics. Research Institute of School Technologies, Moscow (2008)
4. Sergeev, S.F.: Psychological aspects of the interface problem in the technogenic world. *Psycholog. J.* **35**, 88–98 (2014)
5. Sergeev, S.F.: Ideological prolegomens of activity theory. *Philosoph. Sci.* **62**(5), 44–61 (2019)
6. Lepsky, V.E.: Analytics of Assembly of Development Subjects. “Kogito-Center”, Moscow (2016)
7. Lepsky, V.E.: Methodological and Philosophical Analysis of the Development of Management issues. “Kogito-Center”, Moscow (2019)

8. Sergeev, S.F.: Intelligent symbionts of organized technogenic controls for moving objects. *Mech. Auto. Control.* **9**, 30–36 (2013)
9. Sergeev, S.F.: *Educational and Professional Immersive Environments*. Public education, Moscow (2009)
10. Sergeev, S.F.: *Ergonomics of immersive environments: methodology, theory, practice: author. dis. ... Dr. Psychol. Sciences.* (2010)
11. Stepin, V.S.: *Classics, Non-classics, Post-non-classics: Criteria for Distinguishing. Post-non-classics: Philosophy, Science, Culture*. Publishing House Mir, St. Petersburg, pp. 249–295 (2009)
12. Lepsky, V.E.: *Management technologies in information wars (from classics to post-classics). “Kogito-Center”, Moscow* (2016)
13. Sokolov, B.V., Yusupov R.M.: *Neocybernetics: Opportunities and Development Prospects*. Central Research Institute “Elektropribor”, St. Petersburg (2008)
14. *The Embodied Mind: Cognitive Science and Human Experience*, by Francisco Varela, Evan Thompson, and Eleanor Rosch, Cambridge. MIT Press, MA (1991)
15. Knyazev, E.N.: *Enactivism: A Conceptual Turn in Epistemology*. *Philosophy Issues.* **10**, 91–104 (2013)
16. Lefebvre, V.A.: *About self-organizing and self-reflective systems and their research*. *Prob. Res. Syst. Struct.* pp. 61–68 (1965)
17. Lepsky, V.E.: *Subjective-reflexive analysis of management paradigms. Reflexive approach: from methodology to practice* (2009)
18. Sergeev, S.F.: *Artificial and natural intelligence in technogenic educational environments*. *Open Education.* **2**(97), 52–60 (2013)
19. Sergeev, S.F.: *Intelligent symbionts in ergatic systems*. *Sci. Tech. J. Inf. Technol. Mech. Opt.* **2**(84), 149–154 (2013)
20. Ivanov, D.V.: *Enactivism and the problem of consciousness*. *Epistemology Philo. Sci.* **49**, 88–104 (2016)
21. Sergeev, S.F.: *Psychological problems of technogenic modification of a person*. *World Psychol.* **4**(96), 77–86 (2018)

# The Solution of “If-Problem” in Computations with Multi-valued Variables Based on Operator Overloading



Vyacheslav Sal'nikov and Konstantin Semenov

**Abstract** This paper presents the exhaustive solution of the “if-problem” in computations with multi-valued variables based on the technique of operation overloading that is widely used in object-oriented programming. The proposed approach addresses the problem of the conditional branches when due to the uncertainty of the operand the result of the logical operation is not determined explicitly and therefore both conditional branches should be executed simultaneously. For the purpose of the study, the code for C/C++ was developed and tested. The results of the performed tests are provided. The proposed solution makes it possible to transfer previously written C/C++ program code for calculations with single-valued types of variables (float, double, etc.) to calculations with any formalism representing the multi-valued variables using minor code modifications. This offers the opportunity to expand the applicability of earlier developed program code to the wider scope of application problems: from the calculations with input variables accurate to round-off errors to the computations with inaccurate or uncertain data.

**Keywords** Inaccurate Data Processing · Conditional Branches · If-Problem · Operator Overloading · Interval Arithmetics

## 1 Introduction

In many practical situations, we deal with inaccurate data: during measurement results processing, during mathematical modeling with rough or uncertain input data, during calculations with multi-valued variables, etc. Due to inaccurate input data, all final and intermediate results of computations are also inaccurate. If we do not take into account this circumstance, then we can make wrong or ineffective decisions based on calculations results. To estimate their uncertainty characteristics automatically, we can resort to one of two approaches:

---

V. Sal'nikov · K. Semenov (✉)

Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia

e-mail: [semenov.k.k@iit.icc.spbstu.ru](mailto:semenov.k.k@iit.icc.spbstu.ru)

- to develop the new program code especially for the specific application of processing the inaccurate data (with taking into account the information on the type and nature of input data uncertainty),
- to take the already developed code for the problem to be solved that deals with real numbers interpreted as values accurate to round-off errors and to expand it to calculations with multi-valued variables using the object-oriented approach in programming – this will keep the algorithm and code safe from major modifications.

The first approach is not flexible. We can not extend the developed code to other formalisms of input data uncertainty representation: for expanding the code applicability to new variants of input data types, we need to make a profound change in code and maybe change the logic of its functioning.

The second approach is preferable because of its specific advantages. We can take well-developed code for single-valued variables and extend it to the case of the multi-valued variables with including some new library that realizes all the logic of dealing with inaccurate data and with replacing variables type from “float” or “double” to the new one – for example, “interval”. The library contains overloaded arithmetic operators for variables of this specified type [1, 2]. The mentioned approach is widely used in practice because it does not require code rewriting and ensure code considerable flexibility.

The only “underwater rock” of this approach is the so-called “if-problem” stated in [3–5] and mentioned in [6].

## 2 The “If-Problem” Statement

The procedural or object-oriented programming languages mostly have a strict logic for conditional operators. For example, if someone has “if...else...” operator, it always runs either “if” branch or “else” one and never both of them simultaneously. This circumstance is the consequence of the historical way of computational means developing: since the appearance of the first computers, we have dealt with single-valued variables during each computational operation. Now, the computations in many real-life applications migrate from processing data with specified accurate values to calculations with uncertain data presented in the form of multi-valued variables (using intervals [7], fuzzy variables [8, 9], Dempster-Shafer variables [10, 11], or other formalism [12–14]). But, once we start such computations, then the so-called “if-problem” appears that lies in the impossibility of using strict logic for conditional operators in this case.

For example, let the interval variable be defined. Its real value is unknown, and all information we have is that this value lies in a specified range. If we need to test this interval variable on the satisfaction of some condition expressed by the “if...else...” operator, then we can face the problem: one part of the interval may fit the “if”

branch and the second fits an “else” one. So both branches should be executed to meet interval logic. The example showing this circumstance is in Table 1.

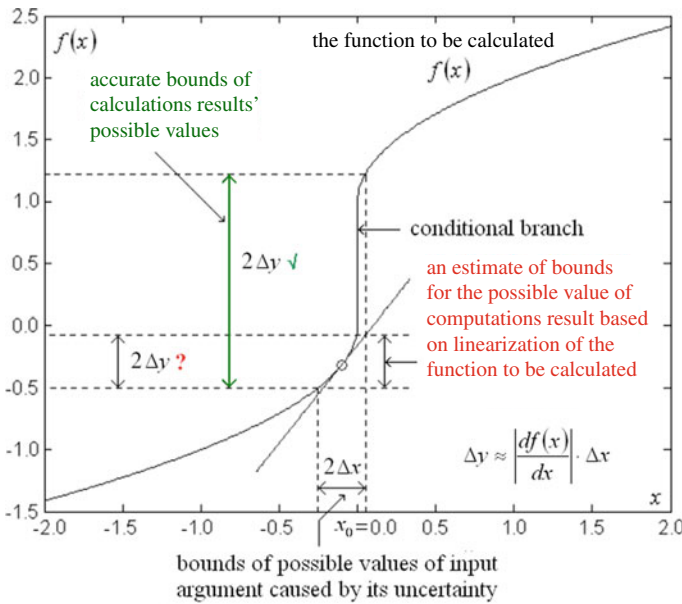
In Table 1, the variable type “interval” describes the interval, and the corresponding constructor uses as inputs its left and right borders.

The graphical explanation of the “if-problem” is shown in Fig. 1. The following notation is used:  $f$  is a function to be calculated,  $x$  is the value of its argument that is known with an error which absolute value is not more than  $\Delta x$ ,  $y$  is a value of the function. The function  $f$  has the jump in values when  $x = x_0 = 0$ :

$$f(x) = \begin{cases} f_1(x), & x \leq x_0, \\ f_2(x), & x > x_0. \end{cases}$$

**Table 1** An example illustrating the “if-problem” for inaccurate data processing

The C/C++ code to be executed	The initial data	The obtained result	The expected result	
/* absolute value computing */ if (x >= 0) return(x); else return(-x);	double x = 1.0;	1.0	1.0	✓
	double x = -1.0;	1.0	1.0	✓
	interval x(1, 2);	(1;2)	(1;2)	✓
	interval x(-3, 4);	(-3;4) or (-4;3) depending on operator “>=” overloading	(0;4)	?



**Fig.1** The graphical explanation of the “if-problem” nature [15]

This circumstance represents the conditional branch: the values of  $y$  for  $x \rightarrow x_0 - 0$  and for  $x \rightarrow x_0 + 0$  differ. To find out, what the possible error of  $y$  can be for the uncertain argument value satisfying an inequality  $x - \Delta x < x_0 < x + \Delta x$ , we need to calculate the values of  $f_1$  and  $f_2$  – i.e. both conditional branches should be evaluated. In practice, the maximum possible absolute value  $\Delta y$  of error of  $y$  is usually estimated using the derivative of function  $f$ :  $\Delta y \approx |df(x)/dx| \cdot \Delta x$ . This ratio suggests only one conditional branch performing: if  $x \leq x_0$  then  $df(x)/dx = df_1(x)/dx$  and if  $x > x_0$  then  $df(x)/dx = df_2(x)/dx$ . In Fig. 1, we see that the result of such estimating  $\Delta y$  is not correct.

The “if-problem” naturally appears when we try to extend our code from calculations with real numbers to calculations with more sophisticated objects using the operator overloading technique. This allows us to avoid algorithm changes and to make it fully independent from the used data representation formalism. In the case when we develop a new code to handle the multi-valued variables from the very beginning – for example, with fuzzy variables – then we can construct a special code for fuzzy variables and can use fuzzy logic conditional operators [16, 17] instead of classical “if...else...” constructions [18].

### 3 The Proposed Solution of the “If-Problem”

Let us try to find a solution for a powerful object-oriented language like C++ . We can create classes and introduce class-specific arithmetic that includes the set of arithmetic (+, -, \*, /) and conditional (>, <, =) operators. Taking into account preprocessor directives (mainly #define), we can try to create more or less convenient syntax for the interval-oriented conditional operators.

How should the overloaded “if...else...” operator work ideally? If we have a usual integer or float variable, it should work as a natural operator (run “if” branch or “else” one). But if we have an interval variable, it should be checked if the whole variable’s range fits the “if” condition (and run only the “if” branch), the “else” condition (so run only the “else” branch), or partially fits both cases (so run both the “if” and “else” branches for corresponding parts of the interval variable range). And being ideal, its syntax should look like a standard “if...else...” C/C++ operator notation. This requirement is important: if no corrections or modifications in the main part of the program code are made, then we will provide code and algorithmic compatibility during migration from calculations with single-valued variables (numbers) to computations with multi-valued variables (intervals, etc.). This allows us to keep the logic of the substantial part of the software regardless of the nature of variables we are dealing with and save time for software developers.

Is it possible? Yes and no. We made it works as described above, but it looks not exactly like the standard C/C++ notation. Take a look at Table 2.

The second column of Table 2 contains an example of the code that is not ideal but is very close to the standard C/C++ syntax presented in the table first column. The only difference is that we use “BEG” in place of “{” and “END” for “}” for

**Table 2** An example showing the code free from the “if-problem”

How it should look	How it will look
<pre>if (x &lt;= 5) { x -= 10; } else { x += 100; }</pre>	<pre>if (x &lt;= 5) BEG x -= 10; END else BEG x += 100; END</pre>

bounds of code corresponded to conditional operators branches. Also both branches in the “if...else...” operator should be placed in “{...}” or in the “BEG...END” section.

Such corrections are minor on the one hand since we always can automatically replace in our code the pre-defined terms “BEG” and “END” to “{” and “}” correspondingly, but significant for hand-driven code corrections on the other hand since we need to replace by hand only that braces “{” and “}” which relate to conditional jumps. To be fair, it should be noted that the last specified replacement can be made automatically using a simple parser of program code. The presented construction looks more suitable and comfortable to be used in comparison with earlier published code constructions [3, 4] for the “if-problem” solution (partial).

Let us describe how the proposed approach works. To overcome the “if-problem,” we have two classes with redefined comparison and arithmetic operators and three defines:

```
#define BEG {if(m.b()){
#define END }m.e();}
#define else if(m.el()).
```

As one can see, we do not have a natural “else” operator but another one “if” instead. It means that both branches of “if...else...” may be executed for the interval or other multi-valued variables. It should be noted that being dependent on used formalism to represent multi-valued variables, the overloading of the conditional operators should be performed carefully because some implementations may raise the issue. The user can calculate the value of an ordinary (single-valued) variable in a multivalued-oriented “if...else...” conditional operator and get the incorrect value because both branches run.

The code was written for classical interval arithmetic [19, 20] and tested in Visual Studio 2012 but it looks like almost any modern C++ compiler should work fine with this code.

The described minor code modifications are the minimum-cost way to overcome the problem of conditional branches for calculations with inaccurate data. The proposed “BEG-END” construction is explained by the preprocessor and C++ compiler limitations.

## 4 The Tests

The presented software construction was tested using a large number of code examples. The results of code executions were successful. In Table 3, some examples of the performed tests are presented.

In Table 3, the class “interval” representing the interval-valued variables is used. All arithmetic operations were overloaded as it is supposed for classical interval arithmetic [19, 20]. The special method “print” is used to print the interval borders to the console.

**Table 3** The code of tests and their results

#	Code example	Output result and comments
1	<pre>// Example 1. Input value = 6, condition is (&lt;= 5) double x = 6; printf("\nInput x = %0.1f",x); if(x &lt;= 5) BEG x-= 10; END else BEG x += 100; END printf("\nOutput x = %0.1f",x);</pre>	<p>Input x = 6.0 Output x = 106.0 “else” branch was run</p>
2	<pre>// Example 2. Input value = 3, condition is (&lt;= 5) double x = 3; printf("\nInput x = %0.1f",x); if(x &lt;= 5) BEG x-= 10; END else BEG x += 100; END printf("\nOutput x = %0.1f",x);</pre>	<p>Input x = 3.0 Output x = -7.0 “if” branch was run</p>
3	<pre>// Example 3. Input value is interval [2; 7], // condition is (&lt;= 0) interval x(2,7); printf("\nInput x = "); x.print(); if(x &lt;= 0) BEG x-= 10; END else BEG x += 100; END printf("\nOutput x = "); x.print();</pre>	<p>Input x = [2.0; 7.0] Output x = [102.0; 107.0] “else” branch was run</p>
4	<pre>// Example 4. Input value is interval [2; 7], // condition is (&lt;= 10) interval x(2,7); printf("\nInput x = "); x.print(); if(x &lt;= 10) BEG x-= 10; END else BEG x += 100; END printf("\nOutput x = "); x.print();</pre>	<p>Input x = [2.0; 7.0] Output x = [-8.0; -3.0] “if” branch was run</p>
5	<pre>// Example 5. Input value is interval [2; 7], // condition is (&lt;= 5) interval x(2,7); printf("\nInput x = "); x.print(); if(x &lt;= 5) BEG x-= 10; END else BEG x += 100; END printf("\nOutput x = "); x.print();</pre>	<p>Input x = [2.0; 7.0] Output x = [-8.0;-5.0], (105.0;107.0) “if” and “else” branches were both run</p>



## 5 Conclusions

This paper presents the elegant solution of the “if-problem” in computations with multi-valued variables by modern means of program languages like C/C++ . The proposed software construction allows performing both conditional branches for inaccurate compared operand if necessary.

The proposed approach solves the 25 years old problem and allows the transfer of earlier written C/C++ program code for calculations with single-valued types of variables (float, double, etc.) to calculations with any formalism representing the multi-valued variables with minor code modifications.

**Acknowledgements** The study was funded by grant No 19-71-00127 of the Russian Science Foundation.

## References

1. Piponi, D.: Automatic differentiation, C++ templates, and photogrammetry. *J. Graph. Tools* **9**(4), 41–55 (2004)
2. Holmqvist, K., Migdalas, A.: A C++ class library for interval arithmetic in global optimization. In: *State of the Art in Global Optimization*, pp. 213–226. Springer, Boston, MA (1996)
3. Beck, T., Fischer, H.: The if-problem in automatic differentiation. *J. Comput. Appl. Math.* **50**(1–3), 119–131 (1994)
4. Kearfott, R.B.: Automatic differentiation of conditional branches in an operator overloading context. In: *Computational Differentiation: Techniques, Applications, and Tools*, pp. 75–81. SIAM (1996)
5. Fischer, H.: Automatic Differentiation: Root Problem and Branch Problem. In: *Encyclopedia of Optimization*, pp. 176–181. Springer (2009)
6. Kubota, K., Iri, M.: Estimates of rounding errors with fast automatic differentiation and interval analysis. *J. Inf. Proc.* **14**(4), 508–515 (1991)
7. Hyvönen, E.: Constraint reasoning based on interval arithmetic: the tolerance propagation approach. *Artif. Intell.* **58**(1–3), 71–112 (1992)
8. Mauris, G., Lasserre, V., Foulloy, L.: A fuzzy approach for the expression of uncertainty in measurement. *Measurement* **29**(3), 165–177 (2001)
9. Abebe, A.J., Guinot, V., Solomatine, D.P.: Fuzzy alpha-cut vs. Monte Carlo techniques in assessing uncertainty in model parameters. In: *Proceedings of the 4th International Conference on Hydroinformatics*. Iowa City, USA (2000)
10. Yager, R.R.: Arithmetic and other operations on dempster-shafer structures. *Int. J. Man Mach. Stud.* **25**(4), 357–366 (1986)
11. Ferson, S., Kreinovich, V., Grinburg, L., Myers, D., Sentz, K.: Constructing probability boxes and Dempster-Shafer structures. Sandia National Lab.(SNL-NM), Albuquerque, NM (USA). Report No. SAND-2015-4166J (2015)
12. Hunter, A., Parsons, S.D. (ed.): *Applications of Uncertainty Formalisms*. Springer (2003)
13. Parsons, S., Hunter, A.: A review of uncertainty handling formalisms. In: *Applications of Uncertainty Formalisms*, pp. 8–37. Springer, Berlin, Heidelberg (1998)
14. Villanueva, M.E., Rajyaguru, J., Houska, B., Chachuat, B.: Ellipsoidal arithmetic for multi-variate systems. *Comput. Aided Chem. Eng.* **37**, 767–772 (2015)
15. Semenov, K.K.: Metrological auto-tracking of calculation programs in informational measuring systems. PhD thesis. St.Petersburg State Polytechnical University (2011) (in Russian)

16. Zadeh, L.A.: Fuzzy logic. *Computer* **21**(4), 83–93 (1988)
17. Hu, B.Q., Wong, H.: Generalized interval-valued fuzzy rough sets based on interval-valued fuzzy logical operators. *Int. J. Fuzzy Syst.* **15**(4), 381–391 (2013)
18. Mari, L.: A computational system for uncertainty propagation of measurement results. *Measurement* **42**(6), 844–855 (2009)
19. Moore, R.E., Kearfott R.B., Cloud M.J.: *Introduction to interval analysis*. SIAM (2009)
20. Hickey, T., Ju, Q., van Emden, M.H.: Interval arithmetic: from principles to implementation. *J. ACM* **48**(5), 1038–1068 (2001)

# The Interval Method of Bisection for Solving the Nonlinear Equations with Interval-Valued Parameters



Konstantin Semenov  and Anastasia Tselishcheva 

**Abstract** The article dwells on the interval extension of the bisection approach for solving nonlinear equations with interval-valued parameters, i.e. the ones that might have values from the specified bounds. It is shown that such a procedure allows to obtain an interval of possible values for equation root that is entirely determined by the equation parameters inaccuracy and does not depend on any other factor. The proposed interval bisection method can be easily implemented. All the differences from the traditional bisection approach for solving equations have a clear meaning. The simple stopping rule is proposed. It is shown that considering the interval nature of equation parameters makes it possible to finish the iterative process of equation solving earlier in full accordance with known information on the equation parameters. The proposed approach keeps the important bisection method property—all the intermediate estimates of the bounds of the root's possible values interval include the exact boundaries. The article provides an illustrative example of how to use the interval bisection.

**Keywords** Inaccurate data · Nonlinear equations · Equation solving · Interval bisection

---

K. Semenov (✉) · A. Tselishcheva  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [semenov.k.k@iit.icc.spbstu.ru](mailto:semenov.k.k@iit.icc.spbstu.ru)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021  
N. Voinov et al. (eds.), *Proceedings of International Scientific Conference on Telecommunications, Computing and Control*, Smart Innovation, Systems and Technologies 220, [https://doi.org/10.1007/978-981-33-6632-9\\_33](https://doi.org/10.1007/978-981-33-6632-9_33)

373

## 1 Introduction

Many problems in data processing require solving nonlinear equations to obtain the desired result: parametrical identification, fitting data using a specific model, some cases of indirect measurements, etc. In most cases, the data to be processed aren't accurate and, consequently, the equations to be solved have uncertain parameters. That's the reason why the roots of such equations cannot be found precise: we always will face the residual uncertainty inherited from the inaccuracy of equation parameters. So, the trials to interpret the roots estimates as quantities, which values are only corrupted with round-off errors and errors caused by using an iterative numerical algorithm of equation solving, aren't correct—if we solve the required equation analytically, calculate the roots directly using corresponding expressions and don't allow round-off errors to occur, then it still does not make the root estimate absolutely accurate. If we don't consider this circumstance, then we overestimate the preciseness of our knowledge on obtained results which in turn leads to an increase in the risks of making erroneous decisions in the future. As an example, we can consider the root of the function  $f(x, a, b) = a \cdot x + b$ , where values of  $a$  and  $b$  aren't known accurately. If we know that the value of  $a$  is inside the interval  $I_a = [1; 2]$  and the value of  $b$  is inside the interval  $I_b = [-2; -1]$  then the root of the function  $f$  is inside  $I_x = [0.5; 2]$ . Applying any numerical method to solve the equation  $f(x, a, b) = 0$  considering the uncertainty of  $a$  and  $b$  will bring us to the interval wider than the mentioned interval  $I_x$ . So, these bounds for the equation root's possible value are the accurate limits that can be reached for some combination of values  $a$  from  $I_a$  and  $b$  from  $I_b$ .

This paper discusses the interval bisection method for solving nonlinear equations in which parameters are known approximately—in practice, usually, all we know is the interval of possible values of these parameters.

## 2 Solving Nonlinear Equations for Indirect Measurements

All the data that we encounter in real-world conditions are uncertain—from world constants to measurement results. In some cases, we can neglect this uncertainty in our calculations and reasoning, but, in other cases, we cannot because of the big price we will pay: decisions made without taking into account the uncertainty of source data may be ill-founded. The absolute accurateness is the distortion of reality, and, in practice, we always need to know the quality of our results. The natural scientific area, in which there are the regulatory requirements to accompany each result with individual characteristics of its uncertainty, is metrology and science on measurements. So, the most relevant example of applications where we deal with equations with inaccurate parameters is the case of indirect measurements when we calculate the value of interest from the measurement results of the related quantities connected with it with the known dependence. Without loss of generality, we can

consider the problem of solving indirect measurements equations to illustrate the approach proposed and describe details.

Many quantities cannot be measured directly for various reasons. In these cases, indirect measurements can be used. Firstly, the mathematical model is constructed and tested that describes the interconnection between quantities measured directly and values to be found out. After performing all corresponded measurements, the necessary values are computed as the root of the equation or the solution of the equations system relating the participating quantities. The calculation of the result is an essential part of the indirect measurements and should be taken into account during metrological characteristics estimation. It should be metrologically verified like any other measurement procedure or conversion. Indeed, as it was mentioned earlier, all results of calculations with uncertain data are always inaccurate too—the uncertainty inherited from the input data cannot be overcome. The only thing that can be performed is to analyze how the uncertainty transforms during the computations and to estimate the value of the final calculations results error.

To date, there are many approaches and methods to support metrologically computations—including solving equations with inaccurate parameters (usually being direct measurement results). These methods can be grouped into two big sets: methods that use randomization (Monte Carlo approach [1], the Cauchy deviate method [2, 3] and related techniques) and methods that perform automatic analysis of the computational algorithm by overloading the operations performed during calculations—assuming a linear approximation (automatic differentiation of first order [4, 5] as the most valuable approach, finite differences and complex step derivatives estimates [6, 7] and similar techniques for sensitivity analysis) and in the common case (interval arithmetic [8, 9] and its extensions and modifications like affine arithmetic [10, 11] or others [12, 13] joined with random variables processing approaches like probability boxes framework [14, 15]). All of these techniques were developed for the wide class of computational problems with inaccurate input data and can be used in computational metrology.

Numerous different methods are developed and used to search the roots of nonlinear equations. The most popular of them are Newton and Newton–Raphson methods, secant method, bisection method, etc [16]. The first three methods need the initial guess as a first estimation of the root, and the last-mentioned one needs the interval of root localization. If the initial guess is unsuccessful, then the iterative process can diverge, and no root can be obtained at all. In many cases, it is impossible to determine if the guess is acceptable or not before the iterative process starts. In contrast, the bisection method guarantees that the final result of root estimation will be got and supposes a very simple procedure to test if the initial localization interval is acceptable or not.

We should consider the additional sides of the issue to determine what method for solving equations from the listed above is better for metrological practice. So for this, we analyze the details of the metrological supporting the corresponding computational procedures. Approaches for processing initial data uncertainty using randomization might cause the situation when the Newton, Newton–Raphson, and secant methods will diverge—so, we should recognize corresponded iterations during

the Monte Carlo method execution or similar approaches applying and should stop timely and throw out the wrong results from consideration. This complicates the procedure of root finding. The same situation may occur if we use operator overloading—the procedure that converges being executed with individual numbers as input variables may begin to diverge if it is executed with intervals or interval-valued quantities like probability boxes. Thus, this can bring us to an unacceptable situation if the equation to be solved is the equation of indirect measurements—we will not obtain any measurement result at all because of the computational procedure. To address these shortcomings, we suggest using the bisection method that always ensures the final result obtaining. Besides, supporting the bisection with one of the discussed approaches for estimating uncertainty inherited from the initial data doesn't bring us to the iterative process divergence. So, combining the bisection method with any kind of procedure of metrological supporting is the preferable way to solve nonlinear equations of indirect measurements.

This paper presents the interval version of the bisection method for solving nonlinear equations with interval-valued parameters that are commonplace in metrological practice. The proposed method is fully in line with metrological requirements that is an advantage in comparison with traditionally used approaches [17, 18]. The way is proposed for taking into account the uncertainty of initial data during equation solving and reasonably set the moment to stop the iteration process.

### 3 The Interval Bisection

Let  $\vec{x}^T = (x_1, x_2, x_3, \dots, x_n)$  be the direct measurement results, and  $f(y, \vec{x}) = 0$  be the equation that connects these measurands with quantity  $y$  that is supposed to be measured indirectly. Let  $\Delta x_1, \Delta x_2, \Delta x_3, \dots, \Delta x_n$  be the absolute errors of  $x_1, x_2, x_3, \dots, x_n$  correspondingly, and let it be known that their maximum possible values satisfy the restrictions:  $|\Delta x_i| < \Delta_i, i = 1, 2, \dots, n$ .

The traditional bisection method [19] ignores that quantities  $x_1, x_2, x_3, \dots, x_n$  are inaccurate and treats them as the only possible values of parameters of the equation to be solved. Let the interval  $I_1 = [a, b]$  be the localization bounds for  $y$ . In metrology during indirect measurements, we usually face equations representing zeros of the monotonic functions  $f$ . So we have the only root because, for one set of direct measurement results, we must have the only one corresponding value of the indirect measurement. If the values  $f(a, \vec{x})$  and  $f(b, \vec{x})$  have different signs, then interval  $I$  will contain the only root. For problems from other fields, we should start with such an interval of values of argument  $y$  that will provide different signs of function  $f$  values obtained at the interval's left and right bounds. Then, we can be sure that not less than one root is inside this interval.

For each step of bisection, the current interval of root localization is divided into two equal parts, and the one that contains the root is chosen. To determine what half should be preferred, the sign of value  $f(0.5 \cdot (a + b), \vec{x})$  in the middle of interval  $I$  should be calculated. The obtained narrowed interval is new localization bounds

for equation root, and then the new iteration starts, and all described operations are repeated. So, let the localization interval for  $i$ -th iteration step be  $I_i = [a_i, b_i]$ . If  $f(a_i, \bar{x}) \cdot f(0.5 \cdot (a_i + b_i), \bar{x}) > 0$ , then  $a_{i+1} = 0.5 \cdot (a_i + b_i)$ ,  $b_{i+1} = b_i$ . If  $f(0.5 \cdot (a_i + b_i), \bar{x}) \cdot f(b_i, \bar{x}) > 0$ , then  $a_{i+1} = a_i$ ,  $b_{i+1} = 0.5 \cdot (a_i + b_i)$ . The interval  $I_{i+1} = [a_{i+1}, b_{i+1}]$  is the localization interval for the next iteration.

The situation stops to be unambiguous if taking into account the information on the uncertainty of initial data. Since some iteration, we will not be able to determine exactly the sign of the value  $f(0.5 \cdot (a_i + b_i), \bar{x})$  because of the influence of uncertainty of direct measurement results  $\bar{x}$  acting as equality parameters. So, we will not be able to decide what half of the current localization interval contains the root of the equation to be solved: for some possible values of  $\bar{x}$ , it will be in the left half, and for other possible values – in the right half.

The solution allowing to overcome this obstacle is to use one of the methods discussed in the previous section of the paper that provides each calculation of the function  $f(y, \bar{x})$  with its individual uncertainty estimate  $\Delta f(y, \bar{x})$ . Then we will be able to determine the moment when the traditional bisection method faces at current iteration  $i$  such center  $c_i = 0.5 \cdot (a_i + b_i)$  of the current root localization interval  $I_i = [a_i, b_i]$  that satisfies the condition  $\Delta f(c_i, \bar{x}) > |f(c_i, \bar{x})|$ . This inequality indicates the situation described above when we cannot choose half of the localization interval for the next bisection iteration. Really, if it holds, then there are no reasons to consider the value  $f(c_i, \bar{x})$  differing from zero. The equivalent form of the inequality is  $0 \in (f(c_i, \bar{x}) \pm \Delta f(c_i, \bar{x}))$ , so we see that any positive or negative values  $f(c_i, \bar{x})$  lying inside the interval determining by mentioned inequality could be formed by distorting the true value equal to zero by measurement errors. In this paper, we suggest using the moment when the analyzed inequality holds as the transition to the second stage of the modified bisection method.

So, the following simple algorithm can describe the first stage of the proposed approach.

```

Input:  $f(y, \bar{x}) = 0$  // the equation to be solved
        $I_0 = [a_0, b_0]$  // the initial root localization interval
Do:    $i := 1$ ; // assign the value to the iteration index
       $c_i := 0.5 \cdot (a_i + b_i)$ ; // calculate the center of localization interval
      While  $\Delta f(c_i, \bar{x}) < |f(c_i, \bar{x})|$  // the condition to end the method first stage
        // choose the half of localization interval for next iteration
        If  $f(a_i, \bar{x}) \cdot f(c_i, \bar{x}) < 0$ 
          Then  $a_{i+1} := a_i$ ;  $b_{i+1} := c_i$ ;
          Else  $a_{i+1} := c_i$ ;  $b_{i+1} := b_i$ ;
        End
       $I_{i+1} := [a_{i+1}, b_{i+1}]$ ; // the new root localization interval
       $c_{i+1} := 0.5 \cdot (a_{i+1} + b_{i+1})$ ; // its center
       $i := i + 1$ ; // increment the iteration index
End
Output:  $I_i$  // the interval of root localization obtained at the last iteration
    
```

On the second stage of the interval bisection, we need to narrow the last obtained interval  $I_i = [a_i, b_i]$  of root localization that provides the unambiguous sign of function  $f$  at its bounds:

$$\Delta f(a_i, \vec{x}) < |f(a_i, \vec{x})| \text{ and } \Delta f(b_i, \vec{x}) < |f(b_i, \vec{x})|$$

The goal of each iteration of the second stage of the proposed method is to narrow these bounds to such an interval  $I_{i+1} = [a_{i+1}, b_{i+1}]$  that ensures holding the condition  $I_{i+1} \subseteq I_i$  and guarantees at the same time that the sign of function  $f(y, \vec{x})$  at  $y = a_{i+1}$  and  $y = b_{i+1}$  isn't ambiguous:

$$\Delta f(a_{i+1}, \vec{x}) < |f(a_{i+1}, \vec{x})| \text{ and } \Delta f(b_{i+1}, \vec{x}) < |f(b_{i+1}, \vec{x})|.$$

Surprisingly, the traditional bisection approach can be easily applied for this purpose. We can reformulate the problem to be solved in the following manner:

- to find the root's minimum possible value, we need to solve equation  $\Delta f(y_{\min}, \vec{x}) = |f(y_{\min}, \vec{x})|$  for  $y_{\min}$  within the localization interval  $[a_i, c_i]$ ;
- to find the root's maximum possible value, we need to solve equation  $\Delta f(y_{\max}, \vec{x}) = |f(y_{\max}, \vec{x})|$  for  $y_{\max}$  within the localization interval  $[c_i, b_i]$ .

Here, as before,  $c_i := 0.5 \cdot (a_i + b_i)$  is the center of interval  $I_i$  that is obtained on the last iteration of the first stage of interval bisection.

Thus, at every new iteration, we need to examine the left and right bound of the interval that localizes the equation root separately. To finish the iterative process, we propose the following rule. It is rational to stop improving the interval estimating when the interval length refining on the next iteration is less than the given constant  $\varepsilon > 0$ :

$$\|I_i\| - \|I_{i+1}\| < \varepsilon.$$

Solving the metrological problems, the uncertainty bounds for the obtained root's value should be rounded – this circumstance is the natural opportunity to determine the best moment to stop the interval bisection method. If the rounded bounds of the interval of root localization obtained on the previous iteration are the same as the rounded bounds of the interval of root localization obtained on the current iteration, then we should finish. The rounding is suggested to be performed in a metrological sense.

The algorithm of the second stage of the interval bisection is the following.



```

Input:   $f(y, \vec{x}) = 0$  // the equation to be solved
         $I_i = [a_i, b_i]$  // the interval of root localization that was obtained
        // on the last iteration  $i$  of the method's first stage
         $\varepsilon$  // user-defined positive constant
Do:     $c_i := 0.5 \cdot (a_i + b_i)$ ; // calculate the center of localization interval  $I_i$ 
         $a'_i := a_i$ ;  $b'_i := c_i$ ; // the interval that localizes  $y_{\min}$ 
         $a''_i := c_i$ ;  $b''_i := b_i$ ; // the interval that localizes  $y_{\max}$ 
Do
    // refining of the left border of the root localization area
     $c'_i := 0.5 \cdot (a'_i + b'_i)$ . // the center of the interval that localizes  $y_{\min}$ 
    If  $\Delta f(c'_i, \vec{x}) < |f(c'_i, \vec{x})|$ 
    Then  $a'_{i+1} := c'_i$ ;  $b'_{i+1} := b'_i$ ;
    Else  $a'_{i+1} := a'_i$ ;  $b'_{i+1} := c'_i$ ;
    End
    // refining of the right border of the root localization area
     $c''_i := 0.5 \cdot (a''_i + b''_i)$ . // the center of the interval that localizes  $y_{\max}$ 
    If  $\Delta f(c''_i, \vec{x}) < |f(c''_i, \vec{x})|$ 
    Then  $a''_{i+1} := a''_i$ ;  $b''_{i+1} := c''_i$ ;
    Else  $a''_{i+1} := c''_i$ ;  $b''_{i+1} := b''_i$ ;
    End
     $I_{i+1} := [a'_{i+1}, b''_{i+1}]$ ; // the new root localization interval
     $i := i + 1$ ; // increment the iteration index
While  $(b''_{i-1} - a'_{i-1}) - (b'_i - a'_i) > \varepsilon$  // condition:  $\|I_{i-1}\| - \|I_i\| < \varepsilon$ 
Output:  $I_i$  // the interval of root localization obtained at the last iteration
    
```

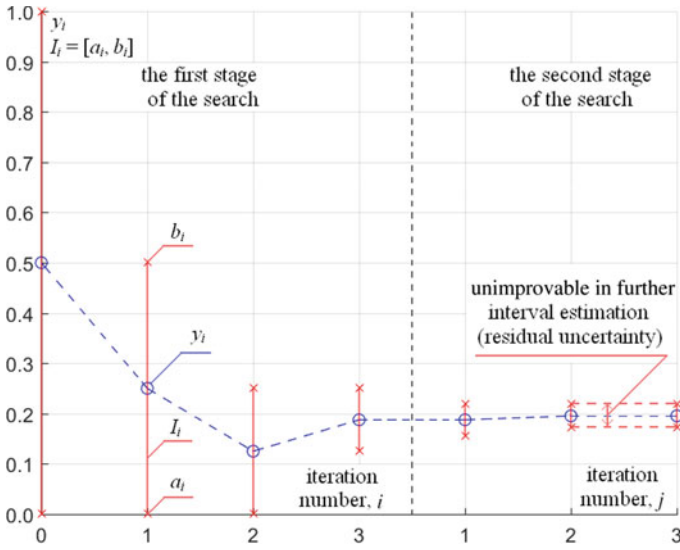
### 4 Illustrating Example

To make the proposed ideas of the interval bisection method clearer, let us examine some function  $f(y, \vec{x}) = \exp(x_1 \cdot y) + x_2 \cdot y$  depending on the variable of our interest  $y$  and a set of parameters  $\vec{x}$  that are known with uncertainty. Let us find the root of the equation  $f(y, \vec{x}) = 0$  using the discussed approach.

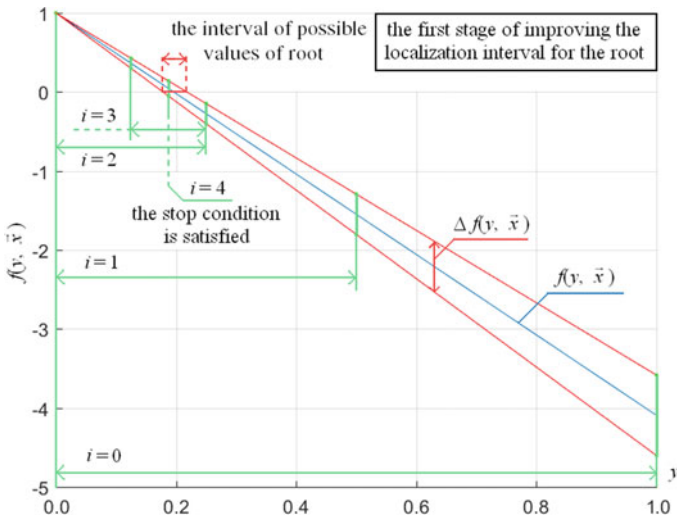
From the physical sense, this equation models the environmental pollution caused by the point source. The parameters  $\vec{x}$  describe the characteristics of the environment and the pollution. From the mathematical viewpoint, this problem is equivalent to calculating the standard Lambert W-function [20].

The values  $\vec{x}^T = (x_1, x_2)$  aren't known exactly. All we know about values  $\vec{x}$  is that  $x_1 \in [-0.11, -0.09]$  and  $x_2 \in [-5.05, -4.95]$ . So,  $(x_1, x_2) = (-0.10, -5.00)$  and  $(\Delta x_1, \Delta x_2) = (0.01, 0.05)$ . Let us try as the start root's localization interval the interval  $I_0 = [a_0, b_0] = [0, 1]$ . These bounds satisfy the condition of the bisection method applicability condition:  $f(a_0, \vec{x}) \cdot f(b_0, \vec{x}) < 0$ .

The entire iterative process that corresponds to using the interval bisection for the mentioned equation is presented in Fig. 1. We can see that, as a result, we obtain further unimprovable interval estimation of the root that cannot be narrower because

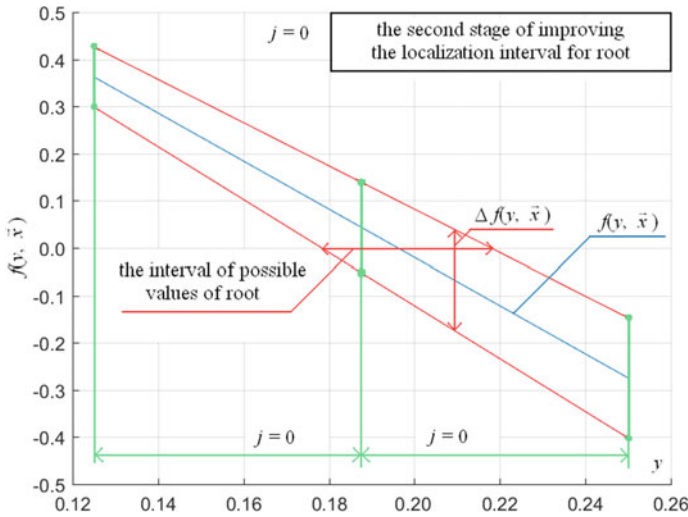


**Fig. 1** Intervals of root localization for different stages of the proposed algorithm



**Fig. 2** Illustration of the first stage of the interval bisection algorithm

of the uncertainty of the solved equation parameters. Figure 1 also illustrates that, during the first stage of the interval bisection method, this approach reproduces the traditional scheme of the bisection and that the second stage essence is in narrowing the last localization interval obtained at the first stage.



**Fig. 3** Results of the first stage of the interval bisection algorithm and the transition to the second stage

In Fig. 2, the results of the first stage execution of the proposed method are illustrated. The stop condition is satisfied on the 4th iteration when we cannot, for the first time, determine the sign of the function  $f$  in the center of the root localization interval. So, we go to the second part of the method.

The results obtained on the several iterations of the second stage of interval bisection are illustrated in Figs. 3, 5, and 6. We can see how the left and right bounds of the localization interval are refined. For convenience, in Figs. 3, 5, and 6, the independent indexing of iterations is used: index  $j = 0$  corresponds to the beginning of the second stage of interval bisection when dealing with the localization interval obtained on the last iteration of the method's first stage.

In Fig. 6, we see the final iteration of the proposed approach. It corresponds to the stopping rule taken from the metrological nature of the solving problem: if we round the uncertainty bounds of the root's estimate at the current iteration, then the new iteration will not bring the refining, and we should finish. The obtained bounds are  $[0.172, 0.219]$ .

## 5 Conclusions

In this paper, the interval extension of the bisection method is proposed for solving nonlinear equations with inaccurate parameters. A simple and effective algorithm is presented that brings with the guarantee to the root estimation. The clear stopping rules are proposed that naturally follow from the problem and allow to finish the

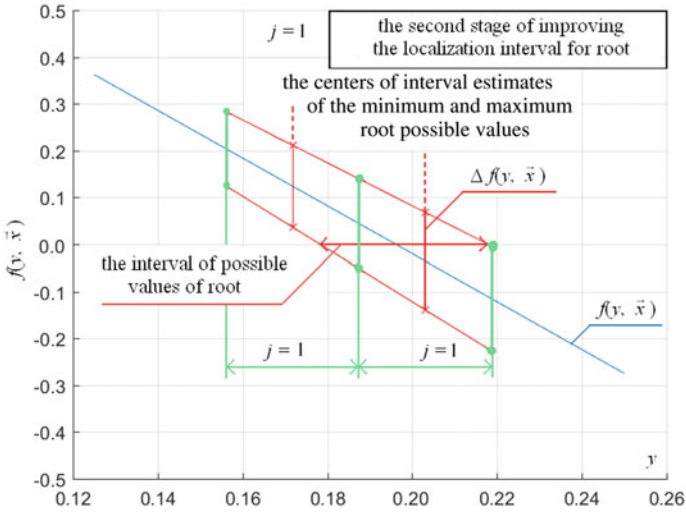


Fig. 4 Results of the first iteration of the second stage of the interval bisection algorithm

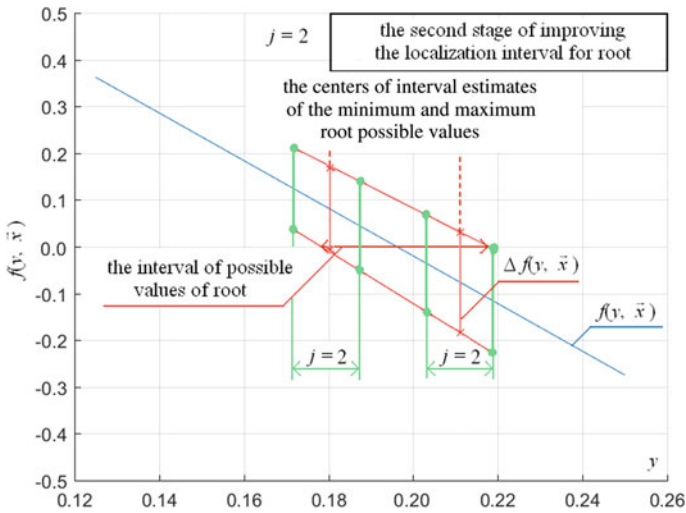


Fig. 5 Results of the second iteration of the second stage of the interval bisection algorithm

iterative process of equation solving earlier in full correspondence with known initial data on the equation to be solved. The proposed approach remains the important property of the bisection method—all the intermediate interval estimates of the root possible values contain the exact bounds.

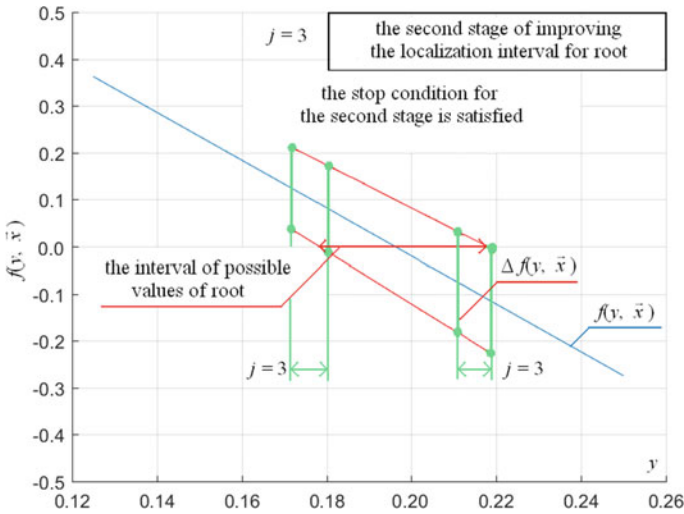


Fig. 6 Final results of the interval bisection algorithm

**Acknowledgements** The reported study was funded by the Russian Foundation for Basic Research (RFBR), project number 19-31-90165.

## References

1. Kreinovich, V., Beck, J., Ferregut, C., Sanchez, A., Keller, G.R., Averill, M., Starks, S.A.: Monte-Carlo-type techniques for processing interval uncertainty, and their potential engineering applications. *Reliable Comput.* **13**(1), 25–69 (2007)
2. Kreinovich, V., Nguyen, H.T.: Towards intuitive understanding of the cauchy deviate method for processing interval and fuzzy uncertainty. In: *IFSA World Congress/EUSFLAT Conference (IFSA-EUSFLAT 2009)*, pp. 2336–2341. Lisbon, Portugal (2009)
3. Kreinovich, V., Ferson, S.A.: A new Cauchy-based black-box technique for uncertainty in risk analysis. *Reliab. Eng. Syst. Saf.* **85**(1–3), 267–279 (2004)
4. Griewank, A.: On automatic differentiation. In: Iri, M., Tanabe, K. (eds.) *Mathematical Programming: recent developments and applications*, pp. 83–107. Kluwer Academic, Dordrecht, the Netherlands (1989)
5. Bücker, H.M., Corliss, G., Hovland, P., Naumann, U., Norris, B. (eds.): *Automatic differentiation: applications, theory, and implementations*. 50. Springer Science & Business Media (2006)
6. Martins, J.R.R.A., Sturdza, P., Alonso, J.J.: The complex-step derivative approximation. *ACM Trans. Math. Software (TOMS)* **29**(3), 245–262 (2003)
7. Lai, K.L., Crassidis, J.: Generalizations of the complex-step derivative approximation. In: *AIAA Guidance, Navigation, and Control Conference and Exhibit*, paper 6348 (2006)
8. Moore, R.E., Kearfott, R.B., Cloud, M.J.: *Introduction to Interval Analysis*. Siam, Philadelphia (2009)
9. Hickey, T., Ju, Q., van Emden, M.H.: Interval arithmetic: from principles to implementation. *J. ACM* **48**(5), 1038–1068 (2001)

10. de Figueiredo, L.H., Stolfi, J.: Affine arithmetic: concepts and applications. *Numer. Algorithms* **37**(1–4), 147–158 (2004)
11. Shou, H., Lin, H., Martin, R., Wang, G.: Modified affine arithmetic is more accurate than centered interval arithmetic or affine arithmetic. In: *Mathematics of Surfaces*, pp. 355–365. Springer, Berlin, Heidelberg (2003)
12. Nedialkov, N.S., Kreinovich, V., Starks, S.A.: Interval arithmetic, affine arithmetic, Taylor series methods: why, what next? *Numer. Algorithms* **37**(1–4), 325–336 (2004)
13. Barrio, R., Rodríguez, M., Abad, A., Blesa, F.: Breaking the limits: the Taylor series method. *Appl. Math. Comput.* **217**(20), 7940–7954 (2011)
14. Ferson, S., Kreinovich, V., Grinzburg, L., Myers, D., Sentz, K.: Constructing probability boxes and Dempster-Shafer structures. Sandia National Lab (SNL-NM), Albuquerque, NM (United States). Report No. SAND-2015-4166J (2015)
15. Troffaes, M.C.M., Miranda, E., Destercke, S.: On the connection between probability boxes and possibility measures. *Inf. Sci.* **224**, 88–108 (2013)
16. Kelley, C.T.: Solving nonlinear equations with Newton's method. SIAM, Philadelphia (2003)
17. Semenov, K.K., Tselishcheva, A.A.: Interval method of bisection for a metrologically based search for the roots of equations with inaccurately specified initial data. *Meas. Tech.* **61**(3), 203–209 (2018)
18. Semenov, K.K., Tselishcheva, A.A.: Generalized interval method of bisection for metrologically based search for solutions of systems of equations with inaccurately specified initial data. *Meas. Tech.* **62**(3), 193–201 (2019)
19. Wu, X.: Improved Muller method and bisection method with global and asymptotic superlinear convergence of both point and interval for solving nonlinear equations. *Appl. Math. Comput.* **166**(2), 299–311 (2005)
20. Corless, R.M., Gonnet, G.H., Hare, D.E., Jeffrey, D.J., Knuth, D.E.: On the Lambert W function. *Adv. Comput. Math.* **5**(1), 329–359 (1996)

# Complex Monitoring Systems for Landfills



Aleksandr Titov , Sergey Krasnov , Andrey Timofeev ,  
and Victor Denisov 

**Abstract** The problem of monitoring the state of landfills is described in the article. There are a lot of such objects in the world. At the same time, there are no standard solutions for monitoring the state of these facilities both in Russia and abroad. It is proposed to develop a technical solution based on autonomous sensors for measuring the concentrations of hazardous fumes, radiation background, geotechnical factors and other environmental variables. Such system can be easily installed at any site and can work offline for a long time. The design and implementation of this system are undoubtedly connected with the issue of investment analysis. Positive economic results of the use of such systems can be an important target component in concept of green economy.

**Keywords** Environmental economics · Green economy · Landfills · LoRaWAN · Intelligent control systems · Big data

## 1 Introduction

Nowadays, the issues of ecological using of resources of big cities and human's influence on the environment are discussed widely [1, 2]. The task of monitoring the status of municipal solid waste dumps and landfills has become an acute importance in the Russian Federation. Hundreds of thousands of legal and illegal dumps and landfills are registered in the country. At the same time, there are no standard solutions for monitoring the state of these facilities both in Russia and abroad. The situation is aggravated by the fact that each object has a unique form and a unique list of hazards.

---

A. Titov (✉) · S. Krasnov  
Peter the Great St.Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [titov\\_ab@spbstu.ru](mailto:titov_ab@spbstu.ru)

A. Timofeev  
LLP "EqualiZoom", Astana, Kazakhstan

V. Denisov  
Flagman Geo Ltd, St. Petersburg, Russia

Therefore, it is proposed to develop a technical solution on the basis of autonomous sensors for measuring the concentration of hazardous fumes, radiation background and geotechnical factors. It is necessary to form a list of parameters of the landfill, which are the objects to constant monitoring. Each sensor must be equipped with a battery and communication module. This will allow you to create a monitoring system, configurable by both the composition of the sensors and the dimensions. Such a system can be easily installed at any site and will be able to work offline for several years, transferring data to remote control centres.

## 2 Background

Studies of the current state of the problem of ensuring environmental safety of landfills showed that the effective use of economic instruments of environmental management in the activities of landfills has been constrained by the lack of well-developed approaches to evaluation of the risk of adverse effects of these objects on the environmental components [3, 4], unlike other risks in the field of environmental management, for which currently there is a detailed scientific and methodological apparatus [5, 6].

Landfills, like any industrial enterprise, have limits of permissible emissions/discharges, limits on waste disposal and for a certain fee perform a set of measures for waste management, including reception and placement for sorting, processing and disposal. In the vast majority, landfills arose spontaneously, without regard to environmental requirements, in waste pits, other types of pits, etc. The average square of each landfill is significant—from several tens to 120 hectares. The time resource of a landfill is developed in three to four years, while the amount of waste does not decrease from year to year, but, on the contrary, increases.

As a rule, to estimate the probability of damage resulting from the occurrence of environmental risks at the landfill, and hence the total cost of losses, several parameters are measured, such as the square of the landfill, the year of its completion and the limit of waste disposal.

However, these works do not consider the possibility of implementing technical systems for early warning of environmental risks. The results of their hazard assessment can play an important role in the management of municipal solid waste and in the development of an economic environmental management mechanism that can solve the problem of environmental pollution in the operation of municipal solid waste landfills.

Environmental risk is the probability of negative changes in the environment under the influence of adverse effects on the environment [4]. The following significant environmental risks inherent in the field of solid waste management can be identified:

1. Air pollution due to ignition of waste disposal facilities [7, 8].
2. Soil contamination with heavy metals.
3. Groundwater pollution in the locations of landfills [9].



4. Formation and emission of harmful gases into the atmosphere [10–13].
5. Increasing the proportion of uncultivated waste landfills.
6. High probability of occurrence of infectious diseases centre at the waste disposal facilities.

As a rule, in this context the main principles of prevention of irreversible consequences for the environment are the following:

1. The correct choice of location for the polygons.  
Polygons are placed outside the settlements in compliance with the size of the sanitary protection zone established by normative documents.
2. Creation the technical design of landfills that prevent the penetration of pollutants into the components of the environment.
3. Proper operation of landfills. In the process of filling, the landfill waste should be provided possibility for garbage trucks and construction equipment, as well as the overall stability of the construction of landfill soils. It is forbidden to accept certain types of waste to landfills.
4. Rationing in the field of waste management. In order to ensure the protection of the environment and human health, to reduce the amount of waste in relation to enterprises as a result of economic and other activities of which waste is generated, standards for waste generation and limits for their disposal are established.
5. Reducing the number of unauthorized landfills.
6. Quality control of stored waste and monitoring of the environment should be organized.

The monitoring system is an information basis for determining the effectiveness of environmental measures, as well as a database for the development of technical and technological solutions to improve landfill operations [14, 15].

Monitoring programs can monitor.

- chemical composition and amount of filtrate formed in the landfill body;
- changes in the quality of groundwater outside the landfill (groundwater monitoring at landfills is carried out with the help of wells);
- air pollution, both in the working area on the territory of the landfill and beyond;
- compliance of the waste entering the landfill with the declared degree of danger [4].

The last one deserves the special attention.

With the active development of automatic monitoring in the last decades, new opportunities for intelligent collection and analysis of information about the state of landfills are appeared. For these purposes, subsystems of the lower level which collect data can be used. And the middle and upper levels allow to assess, analyse and predict the development of adverse events in complex systems. Last times, the Industry 4.0 has penetrated into the most various spheres of our life [16, 17].

### 3 Methods

It is proposed to use a complex system of monitoring of landfills in order to minimize environmental risks. The main objectives of the integrated monitoring of polygons are the following:

- timely detection of exceedance of critical levels of hazardous fumes of methane (CH<sub>4</sub>), carbon dioxide (CO<sub>2</sub>), and methanol, petrol, toluene and ethanol;
- control of internal deformation of the landfill, such as landslides and subsidence;
- control of radiation situation on its territory;
- perimeter control;
- working in stand-alone mode (from internal power sources) and transfer of information via wireless communication lines to the data collection centre in accordance with the adopted regulations;
- construction of a dynamic map of dangerous incidents in the area of the landfill, which is modified as the next data from the spatially distributed data collection system.

As a result, a detailed database of recorded events should be maintained.

Figure 1 shows the block diagram of the integrated monitoring system of landfill. It is necessary to consider each of the subsystems separately.

1. Gas analysers subsystem is a spatially distributed network of gas analysers of various types, capable of detecting dangerous vapours from the target list with high reliability.

The location of the network sensors is determined by the structure of the polygon and its geotechnical parameters. Each sensor network is integrated with LoRaWAN

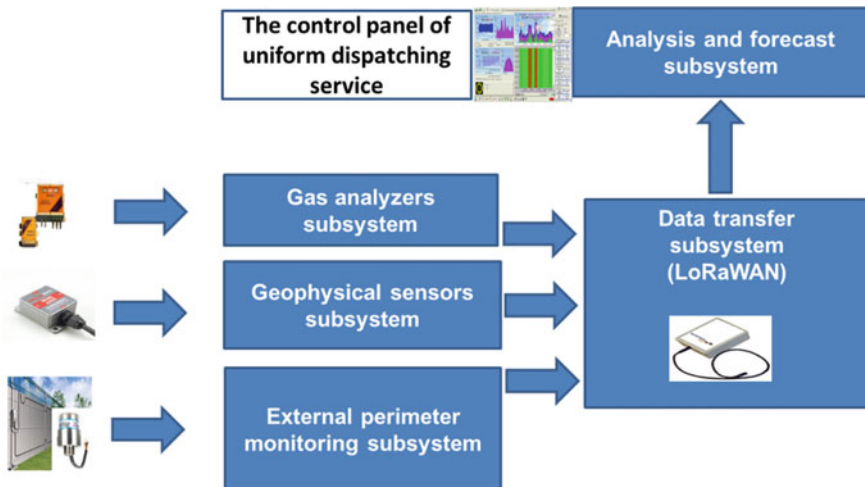


Fig. 1 Structure of the proposed landfill monitoring system

modem for data transmission and to control its functioning [18]. The form factor of each sensor must ensure the normal operation of the sensor in severe weather conditions. The energy supply of the sensors is supplied from an independent power supply.

2. The subsystem of geophysical sensors is a spatially distributed network of seismic sensors inclinometers, designed to monitor the state of the grounds of landfill and control their internal deformations. Inclinometric complex of soil control provides monitoring of soil movements and allows to assess the condition of the foundations of engineering structures, dams, quarries and other structures. The equipment of the complex provides a stationary installation, automatic operation and wireless data transmission during the entire calibration period. Energy supply of sensors is provided by batteries, fuel cells or combined power plants. The subsystem of geotechnical sensors can be equipped with additional systems for building monitoring and earthquakes monitoring.
3. External perimeter monitoring subsystem is designed to monitor activities in the area of the geometric boundaries of the landfill. It consists of a seismosensitive C-OTDR system [19], designed to monitor seismic activity in the area of the controlled perimeter, as well as a network of long-range surveillance cameras. This subsystem is a typical bimodal perimeter control system (modes: “video” and “seismic”). The video mode is provided by the use of long-range video monitoring systems (up to 3 km), which are placed on special towers and are powered from the central power plant or from the autonomous power subsystem (diesel generator and solar panels). Seismic fashion is achieved by using C-OTDR system vibrosensors type. A sensitive sensor of this system is a standard optical fibre SMF-28, buried along the perimeter of the landfill to a depth of 30–50 cm. This system provides detection of a pedestrian at a distance of 5 m from the sensor when the value of the spatial resolution along the cable length is from 5 to 10 m. One analyser system is capable of servicing a sensor length of up to 40 km, while ensuring from 4 000 to 8 000 channels. Such systems have a reputation for being very reliable and relatively inexpensive solutions optimized for perimeter control of extended facilities [19].
4. Data transfer subsystem collects information from sensors of different types, the data transmission based on wireless technologies (LoRaWAN). Each sensor of gas analysers and geotechnical sensors subsystems is equipped with LoRaWAN modem operating in unlicensed 868 MHz range. Depending on the mode of operation of the monitoring system, each sensor is assigned an individual mode of operation, which depends on the type of sensor, the season, the situation at the monitoring site, etc. In accordance with the concept of LoRaWAN, each modem is equipped with a battery that guarantees up to 8 years of operation of the modem without recharging the battery. Data from LoRaWAN modems is collected at LoRaWAN base stations. The number of these stations depends on the square of the landfill, and one base station is able to serve up to 5000 modems at a distance of 3–5 km. Each base station is equipped with a conventional GSM modem, which is connected via a mobile network type 2G (or higher) with a

mobile operator to access the Internet via a secure channel. If the location of the landfill is not mobile or it is unstable, connection LoRaWAN base station with mobile operators can operate under a special radio channels with encrypted traffic. This is a relatively inexpensive and very effective solution. The proposed topology of this network is “star”.

- 5. Analysis and forecast subsystem of intellectual analysis and prediction of the landfill state dynamics are based on complex accounting of information about the geological state of the polygon body, a network of precedent events of emission of target gases and spatiotemporal dynamics of the event flow.

This subsystem is designed to solve the short-term predictive problem of the landfill state on the basis of a set of data collected from subsystems of gas analysers and geotechnical sensors, as well as initial data on the geophysical structure of the landfill.

Decision-making technology is based on using modern achievements of machine learning [20].

### 4 Discussion and Results Analysis

Figure 2 shows an illustration of the implemented system superimposed on the map of the existing landfill.

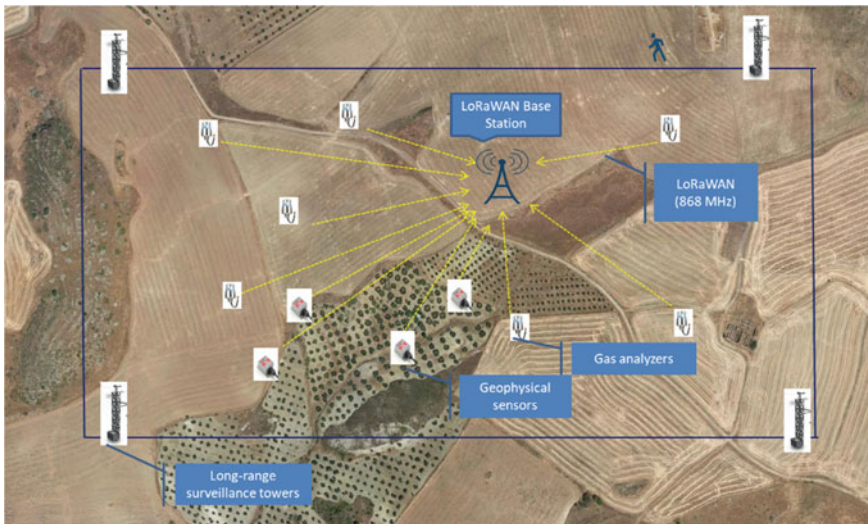


Fig. 2 An example of system implementation

Since the creation, implementation and operation of an complex monitoring system of landfills require an investment costs, it is reasonable to use classical approaches to estimating the return on investment.

The integral efficiency should reflect the synergetic effect of obtaining new opportunities for management decisions for Environmental enterprise (landfill). These management decisions can potentially reduce or even eliminate the costs of liquidation of consequences of emergency situations, investigative measures, repairs, etc. (economy effect):

$$NPV = \sum_{i=0}^n \frac{E}{(1+r)^i} - \sum_{i=0}^n \frac{I}{(1+r)^i}, \quad (1)$$

where NPV is net present value;

- $E$  the general economy effect in the corresponding period;
- $r$  the discount rate;
- $I$  investments in the creation and supporting the system;
- $i$  period number (varies from 1 to  $n$ ).

Thus, the cumulative effect in the long-term period can be determined using the  $NPV > 0$  criterion.

The presented model provides a clear quantitative assessment of the feasibility of investment in the design, construction and operation of the system. It should be noted that the scaling of the system to related areas of monitoring can bring additional economic benefits, which can be evaluated by similar methods. Synchronous measurement of values of various parameters of controlled objects is the most actual direction of modern technology, based on increasing amounts of data. So, developing the lower level of big data management systems can achieve a qualitatively new effect in emergency management in the frame of green economy concept.

## 5 Conclusion

Nowadays, the proposed monitoring system does not have effective serial solutions both in Russia and abroad.

The proposed concept can be implemented as in Russia as abroad. It will be a network of autonomous wireless sensors capable of transmitting information in the LoRa format to its own base station, which will transmit the received data via mobile communication channels at any distance to the control centre.

The maximum distance from the sensors to the base station is not more than 15 km. The frequency of the survey is at least once an hour. Battery life without battery replacement should be up to 5 years.

The use of wireless technologies will allow to reconfigure the network of sensors for objects of different size and shape.

The composition of the sensors used can easily vary depending on the conditions of a particular landfill.

A generalized model for assessing the economic feasibility of investments in the design, creation and supporting the system of permanent monitoring of landfills is proposed in this paper.

**Acknowledgements** This paper was financially supported by the Ministry of Education and Science of the Russian Federation on the programme to improve the competitiveness of Peter the Great St.Petersburg Polytechnic University (SPbPU) among the world's leading research and education centres in the 2016–2020.

## References

1. Sergeev, S., Kirillova, T., Krasnyuk, I.: Modelling of sustainable development of megacities under limited resources. *E3S Web of Conf.* **91**, 05007 (2019)
2. Didenko, N., Skripnuk, D., Mirolyubova, O.: The effects of human behavior on fresh water resources international multidisciplinary scientific geo conference surveying geology and mining ecology management. *SGEM* **17**(53), 901–910 (2017)
3. Mudretsov, A., Tulupov, A.: Estimation of ecological risks of landfills. *Reg. Prob. Trans. Econ.* **3**(37), 242–247 (2013)
4. Derkacheva, E., Razinkov, N.: Emergency situations related to the fire of solid waste disposal sites and the way to solve the problem. *Complex problems of technosphere safety: materials international. Scientific Conference 12 Nov. 2015, Voronezh. Part 9.* pp. 138–142 (2016)
5. Kotlyarov, I., Petrov, S.: Risk assessment procedure for economic-geological and cost estimate of mineral deposits. *Gornyi Zhurnal.* **9**, 94–99 (2014)
6. Borisovich, V., Kurbanov, N., Zaernyuk, V., Sefullaev, B.: Practical risk management at gold mining companies. *Gornyi Zhurnal.* **11**, 70–75 (2018)
7. Vaverková, M., Adamcová, D.: Long-term Temperature monitoring of a municipal solid waste landfill. *Polish J. Environ. Stud.* **24**, 1373–1378 (2015)
8. Shi, J., Zhang, T., Zhang, J., Ai, Y., Zhang, Y.: Prototype heat exchange and monitoring system at a municipal solid waste landfill in China. *Waste Manag.* **78**, 659–668 (2018)
9. Lopes, D.D., et al.: Geophysical technique and groundwater monitoring to detect leachate contamination in the surrounding area of a landfill—Londrina (PR – Brazil). *J. Environ. Manage.* **113**, 481–487 (2012)
10. Jovanov, D., Vujić, B., Vujić, G.: Optimization of the monitoring of landfill gas and leachate in closed methanogenic landfills. *J. Environ. Manage.* **216**, 32–40 (2018)
11. Beaven, R., Scheutz, C.: Landfill gas emission monitoring. *Waste Manag.* **87**, 833–834 (2019)
12. Kormi, T., Mhadhebi, S., Bel Hadj Ali, N., Abichou, T., Green, R.: Estimation of fugitive landfill methane emissions using surface emission monitoring and genetic algorithms optimization. *Waste Manag.* **72**, 313–328 (2018)
13. Xing, Z., et al.: Real-time monitoring of methane oxidation in a simulated landfill cover soil and MiSeq pyrosequencing analysis of the related bacterial community structure. *Waste Manag.* **68**, 369–377 (2017)
14. Iacoboaia, C., Petrescu, F.: Landfill monitoring using remote sensing: a case study of Glina. Romania. *Waste Manag. Res.* **31**, 1075–1080 (2013)
15. Simões, G.F., Catapreta, C.A.A.: Monitoring and modeling of long-term settlements of an experimental landfill in Brazil. *Waste Manag.* **33**, 420–430 (2013)
16. Glukhov V., Ilin I., Iliashenko O.: Improving the efficiency of architectural solutions based on cloud services integration Lecture Notes in Computer Science (including subseries Lecture

- Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9870 LNCS, pp. 512–524 (2016)
17. Borisoglebskaya, L., Provotorova, E., Sergeev, S., Khudyakov, A.: Automated storage and retrieval system for Industry 4.0 concept. International Scientific Workshop «Advanced Technologies in Material Science, Mechanical and Automation Engineering» MIP: Engineering-2019. IOP Conference Series: Materials Science and Engineering, vol. 537 pp. 032036 (2019)
  18. Gusev, O.: Experiment to create a monitoring system of economic objects using LORAWAN. *Wireless Technol.* **2**(43), 72–76 (2016)
  19. Timofeev, A.: Comparison of various approaches to multi-channel information fusion in C-OTDR systems for remote monitoring of extended objects. *Sci. Tech. J. Inf. Technol. Mech. Opt.* **15**(1), 122–129 (2015)
  20. Flah, P.: *Machine Learning*. DMK Press, Moscow (2015)

# Modeling the Control Object in the Management System of the Regional Socioeconomic System



Elena Averchenkova 

**Abstract** The structural and functional model of the control system of the regional socioeconomic system is presented. The author proposes a model of the control object of the regional socioeconomic system in the general context of managing the subject of the Russian Federation of the type “region”. It is shown how the vector matrix calculus apparatus can be used to describe the control object in the control system of the regional socioeconomic system. The task of managing the regional socioeconomic system is formulated as choosing a vector of the controlling action for transition to the desired state of the state matrix of the control object, for which the corresponding subdomain in the range of permissible values is determined. The features of dynamic processes inherent to the controlling action and output coordinates of the control object are considered.

**Keywords** Regional socioeconomic system · Modeling the control object · Dynamics of managerial impact · Output coordinates of the system

## 1 Introduction

Considering the regional socioeconomic system, hereinafter RSES, from the perspective of choosing alternatives under the situation with uncertainty of influencing the external environment made it possible to use the apparatus of control theory to describe it. So, the RSES is considered as a control object, experiencing a control action formed under a certain influence. On the other hand, the information produced by the external environment of the RSES is characterized by increased complexity, heterogeneity and inconsistency. In addition, there is a need for an integrated approach to managing the RSES on the basis of using modern instruments of state influence on the regions and the country as a whole, namely using national projects.

---

E. Averchenkova (✉)  
Bryansk State Technical University, Bryansk, Russia  
e-mail: [lena\\_ki@inbox.ru](mailto:lena_ki@inbox.ru)



## 2 Relevance of the Article

A systematic presentation of the theory structure of managing socioeconomic systems was given in the works of Russian scientists at the Institute of Control Sciences of the Russian Academy of Sciences named after Trapeznikov V.A. An initial idea of the tasks and methods of the theory of managing organizational systems was given in the works of Burkov [1]. The current state of the theory of managing organizational systems was developed in the works of Novikov [2, 3], Voronin et al. [4].

The results of a theoretical study of models and methods of managing organizational systems find their application in solving a wide range of practical problems in various fields, for example, to describe the processes of effective management of enterprises, corporations and regions, which we can see, for example, in the works of Burkov et al. [5] and Chkhartishvili [6]. Scientists such as Ajzerman and Aleskerov developed basic theory of options choice [7], and in the works of domestic [8, 9] and foreign authors [10–17] it shows the development of the theory of decision making. Expert assessments were considered in the works of Litvak and Orlov [18, 19].

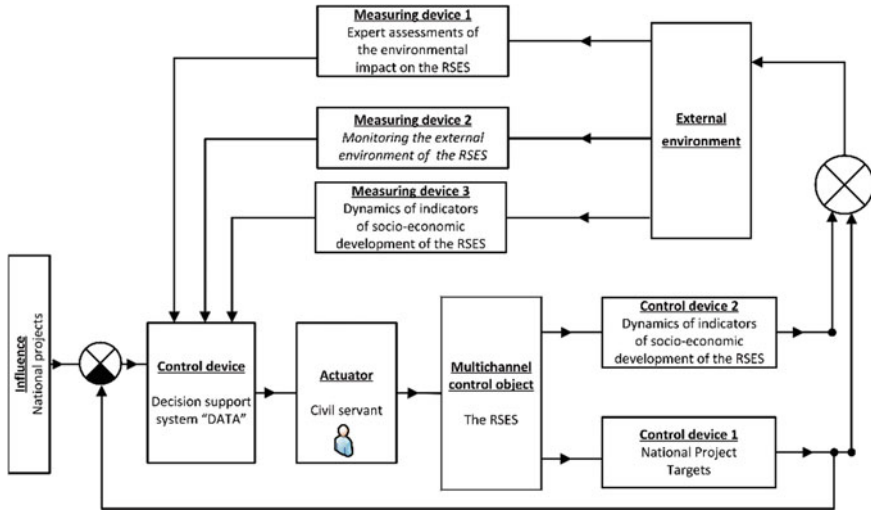
On the other hand, issues of effective regional management are an important area of research for business scientists and practitioners. In the works of Butrin et al. [20, 21] and Tatarkin [22] regions are considered as objects of management, taking into account their economic, political, natural and other features.

However, a review of work in the field of formalizing the control system of the regional socioeconomic system (MS RSES) has revealed certain reserves for improvement in this area. So, the peculiarity of this study is applying modern instruments of state influence on the regions and the country as a whole, namely using national projects in the developed MS RSES. Under this formulation, the conditions for the effective development of the region of the Russian Federation to achieve the targets set by the set of National Projects, as well as the estimated controlling actions to achieve the desired state of the regional socioeconomic system are determined.

Thus, the purpose of this work can be defined as applying the principles and concepts of the control theory to describe the control object that is a regional socioeconomic system from both a structurally functional and mathematical point of view.

## 3 Problem Statement

Consider the structure and functional components of the MS RSES in the context of the authors' works [23, 24]. MS RSES (Fig. 1) includes external environment, a control object (RSES), a control device (DATA decision support system), an executive device (abstract civil servant), three measuring devices and two control devices. From a functional point of view, this model uses a combined control principle that takes into account the negative feedback loop and the compensation circuit for errors and environmental disturbances.



**Fig. 1** Control system of the regional socioeconomic system

The preset impact  $\overline{g(t)}$  in the developed MS RSES is represented by the target settings of the National Projects of the Russian Federation. The control action  $\overline{g(t)}$  is supplied to the control device (decision support system “DATA”) and is corrected by the error  $\varepsilon(t)$  by means of the negative feedback system.

The external environment gives rise to external disturbing influences of various nature. The main perturbations  $\overline{f(t)}$  are taken into account (compensated) by the control device of the system represented by the decision support system “DATA” (DSS “DATA”). Information (signal) coming from the external environment to the control object is redirected to measuring devices N 1, 2 and 3.

The control device generates a controlling action  $\overline{u(t)}$  on the control object (the RSES) and is presented by DSS “DATA”. Features of functioning DSS “DATA”, as well as the algorithms of its operation, were previously described in [11]. The control device processes the current environmental information obtained using measuring devices 1, 2 and 3.

The executive device in the developed MS RSES is represented by an abstract civil servant—the manager of the lower and middle level who in practice is a user of the DSS “DATA”. He addresses it in the process of making managerial decisions to form controlling actions  $\overline{u(t)}$ .

The vectors of the output coordinates  $\overline{y_\varphi(t)}$  are summed up and affect the external environment in the form of a single vector  $\overline{Y(t)}$ : This is how the dualistic nature of the RSES manifests itself as both a control object and a control subject.

## 4 Theoretical Part

### 4.1 Mathematical Model RSES

To study the control object (RSES) in the context of the control theory, it is necessary to clearly formulate the structural components of the RSES, for which this study uses the method of hierarchical classification. The basis for separating many elements of the RSES is a sign of a socially oriented influence of the National Projects of the Russian Federation on the subjects of the Russian Federation. As a result of the hierarchical classification, mnemonic codes are generated that facilitate the process of forming the controlling action on the RSES, as well as assessing the output parameters of the RSES after the controlling action.

Let us represent the control object in the control system of the RSES as a union of two matrices of indicators describing the RSES. On the one hand, these are target indicators of the national projects, and on the other hand, these are indicators of form 2P under the title “Key Indicators for Forecasting the Socio-Economic Development of the Russian Federation for the Mid-Term Period”, which are presented by the executive authorities of the constituent entities of the Russian Federation to the Ministry of Economic Development of the Russian Federation:

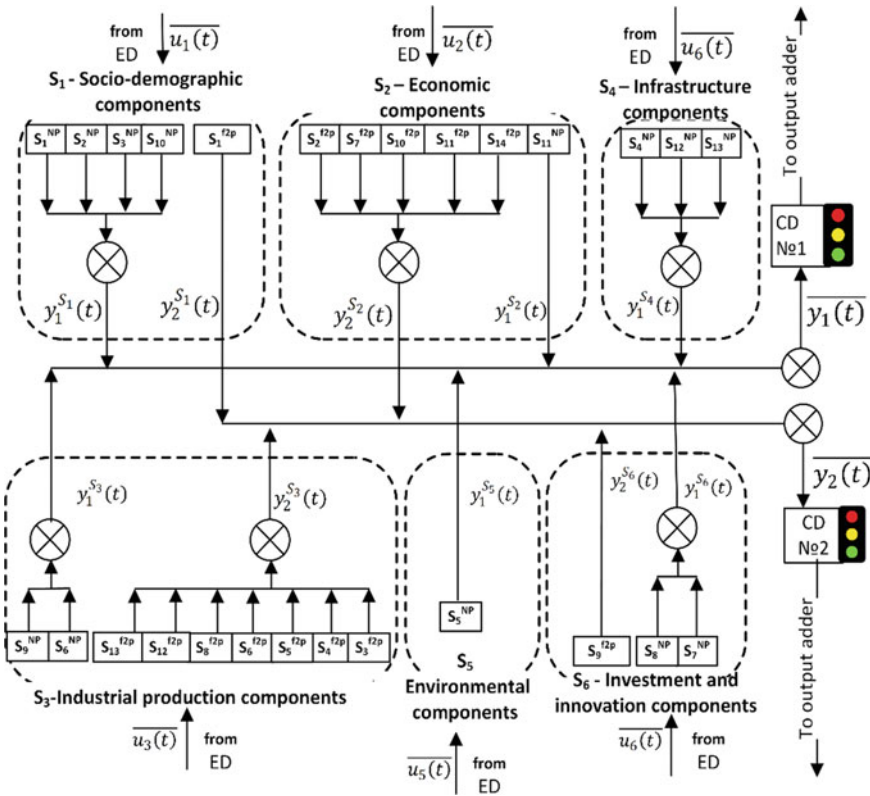
$$S = S^{NP} \cup S^{f2P} = \parallel S_{\beta\gamma}^{NP} \parallel_{\beta=1, \gamma=1}^{13 \times 31} \cup \parallel S_{\rho\epsilon}^{f2P} \parallel_{\rho=1, \epsilon=1}^{14 \times 38} = \parallel S_{mn} \parallel_{m=1, n=1}^{14 \times 69} \quad (1)$$

where  $S^{NP}$  is a matrix of the national projects targets,  $S_{\beta\gamma}^{NP}$  is a matrix element representing the target of the  $\beta$ -th national project;  $\beta = 1, 2, \dots, 13$ ,  $\gamma = 1, 2, \dots, 31$ ;  $S^{f2P}$  is a matrix of key indicators presented for forecasting socioeconomic development of the Russian Federation (for subjects of the Russian Federation);  $S_{\rho\epsilon}^{f2P}$  is a matrix element representing the  $\epsilon$ -th target indicator of the  $\rho$ -th group of indicators presented for forecasting socioeconomic development of the Russian Federation;  $\rho = 1, 2, \dots, 14$ ,  $\epsilon = 1, 2, \dots, 38$ ;  $S_{mn}$  are elements of the new state matrix of the control object;  $m = 1, 2, \dots, 14$ ;  $n = 1, 2, \dots, 69$ .

### 4.2 Structural and Functional Model of the Control Object—RSES

To describe the RSES, vector matrix calculus apparatus is used. The RSES is represented as a multiply connected control object with several interconnected vectors of the controlling actions  $\overline{u(t)}$  and output coordinates  $\overline{y(t)}$ . A visual representation of the control object, the RSES, as a combination of input and output parameters, is shown in Fig. 2.

It is determined that S1 is a group of socio-demographic components, S2 is a group of economic components, S3 is a group of industrial and production components, S4



**Fig. 2** Structural and functional diagram of the control object—the regional socioeconomic system (RSES)

is a group of infrastructure components, S5 is a group of environmental components, and S6 is a group of investment and innovative attractiveness components of the region.

Consider the input and output parameters in the structure of RSES. Each group of the components of RSES  $S_m$  corresponds to its own control action  $\overline{u_m(t)}$ , where  $m = 1, 2, \dots, 6$ , comes from the executive device (ED). Under their influence, two output coordinates are formed in each group  $S_m$ :

- arriving at the control device №1 (CD №1)  $y_1^{S_m}(t)$ , где  $m = 1, 2, \dots, 6$ ;
- arriving at the control device No. 2 (CD №2)  $y_2^{S_m}(t)$ , где  $m = 1, 2, \dots, 6$ .

In other words, a vector of output coordinates comes from each group  $S_m$  of components of the RSES:  $\overline{y^{S_m}(t)} = (y_1^{S_m}(t), y_2^{S_m}(t))$ .

We define the vector of the RSES output coordinates arriving at CD №1 as the sum of the corresponding signals from each group  $S_m$ :

$$\overline{y_1(t)} = \sum_{m=1}^6 y_1^{S_m}(t) \tag{2}$$

As each group of the components  $S_m$  of the RSES forms its output coordinate  $\overline{y^{S_m}(t)} = (y_1^{S_m}(t), y_2^{S_m}(t))$ , the signals arriving at CD №1 can be represented as the following system:

$$y_1^{S_m}(t) = \begin{cases} y_1^{S_1}(t) = y_1^{S_1^{NP}}(t) + y_1^{S_2^{NP}}(t) + y_1^{S_3^{NP}}(t) + y_1^{S_{10}^{NP}}(t) \\ y_1^{S_2}(t) = y_1^{S_{11}^{NP}}(t) \\ y_1^{S_3}(t) = y_1^{S_6^{NP}}(t) + y_1^{S_9^{NP}}(t) \\ y_1^{S_4}(t) = y_1^{S_4^{NP}}(t) + y_1^{S_{12}^{NP}}(t) + y_1^{S_{13}^{NP}}(t) \\ y_1^{S_5}(t) = y_1^{S_5^{NP}}(t) \\ y_1^{S_6}(t) = y_1^{S_7^{NP}}(t) + y_1^{S_8^{NP}}(t) \end{cases} \tag{3}$$

For CD №2, a similar vector of the RSES output coordinates is not determined by all  $S_m$  groups and can be written as follows:

$$\overline{y_2(t)} = y_2^{S_1}(t) + y_2^{S_2}(t) + y_2^{S_3}(t) + y_2^{S_6}(t) \tag{4}$$

The signals arriving at CD №2 from each group  $S_m$  represent the following system:

$$y_2^{S_m}(t) = \begin{cases} y_2^{S_1}(t) = y_2^{S_1^{f2P}}(t) \\ y_2^{S_2}(t) = y_2^{S_2^{f2P}}(t) + y_2^{S_7^{f2P}}(t) + y_2^{S_{10}^{f2P}}(t) + y_2^{S_{11}^{f2P}}(t) + y_2^{S_{14}^{f2P}}(t) \\ y_2^{S_3}(t) = y_2^{S_3^{f2P}}(t) + y_2^{S_4^{f2P}}(t) + y_2^{S_5^{f2P}}(t) + y_2^{S_6^{f2P}}(t) + y_2^{S_8^{f2P}}(t) + y_2^{S_{12}^{f2P}}(t) + y_2^{S_{13}^{f2P}}(t) \\ y_2^{S_6}(t) = y_2^{S_9^{f2P}}(t) \end{cases} \tag{5}$$

Thus, having passed the control devices, the output coordinates  $\overline{y_1(t)}$  and  $\overline{y_2(t)}$  arrive at the corresponding adder and affect the external environment in the form of a single vector  $Y(t)$ .

### 4.3 The Task of Managing MS RSES

The components of the state matrix  $S = \|S_{mn}\|_{m=1, n=1}^{14 \times 69}$  satisfy certain restrictions, i.e., the matrix  $S(t)$  in the state space should not go beyond the domain  $Q$  which is the region of admissible states, i.e.,  $S \in Q$ .

In the  $Q$  region, a certain subdomain of the  $Qc$  states is distinguished, which is desirable. We define  $Qc$  as a subdomain of  $Q$  (i.e.,  $Q_c \in Q$ ) bounded by the

materiality threshold previously described in [12]:

$$Q_c = \left\{ S_{mn}(t) : \frac{S_{mn}(t_{z+1}) - S_{mn}(t_z)}{S_{mn}(t_z)} \leq \frac{\sum_{i=1}^k \frac{S_{mn}(t_{z+1}) - S_{mn}(t_z)}{S_{mn}(t_z)}}{k}, S_{mn}(t) \in R \right\} \quad (6)$$

where  $S_{mn}(t_z)$  is some element of the matrix  $S$  in the  $z$ th year;  $k$  is the number of pairs of chain relative deviations of the values  $S_{mn}(t)$ ;  $z$  is the current time period.

Thus, the goal of the RSES control is to transfer the RSES from the initial state  $S(t_0)$  to the final state  $S(t_k)$ , where  $S(t_k) \in Q_c$ .

Let us determine that in order to achieve the RSES control goal, it is necessary to apply the corresponding controlling action to the input of the control object (RSES). Thus, we define the task of the control system of the RSES as follows: From the range of admissible values of  $Q$ , it is necessary to select such a vector of the controlling action  $\bar{u}^*(t)$  that the object control (RSES) for a given initial state and the known vector  $\bar{f}(t)$  would have a solution  $Y(t)$  satisfying the constraint  $Y(t) \in Q(Y)$  for all  $t \in [t_0, t_k]$  and the finite condition  $S(t_k) \in Q_c$ .

#### 4.4 Dynamic Processes Characterizing RSES

The control object (RSES) is characterized by transients, i.e., the RSES is characterized by the reaction of a dynamic system to an external action from the start of this effect to a certain state. Therefore, to describe the mathematical model of the control object (RSES), which allows testing the control object itself for dynamics, it is proposed to use the following expression:

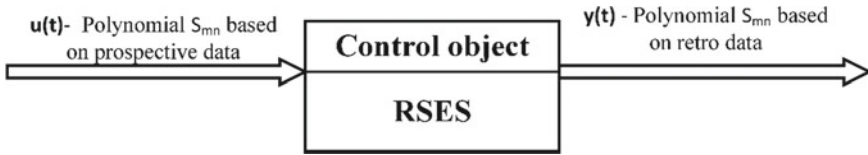
$$F(y', y'', \dots, y^{(\varphi)}, u', u'', \dots, u^{(\zeta)}) = 0, \quad (7)$$

where  $y', y''$  и  $y^{(\varphi)}$  are accordingly, the first, second and  $\varphi$  th derivatives of the output coordinates of the control object MS RSES;  $u', u''$  и  $u^{(\zeta)}$  are accordingly the first, second and  $\zeta$ th derivatives of the control action vector on the control object MS RSES.

Further, it is proposed to consider the features of dynamic processes inherent to the output coordinates  $\bar{y}(t)$  of the control object, RSES.

So, output coordinates  $y(t) = (y_1(t), \dots, y_\varphi(t), \delta_2(t))$  are polynomials formed on the basis of retrospective values  $S_{mn}$  which are elements of the state matrix of the control object. Accordingly, controlling influence  $u(t) = (u_1(t), \dots, u_\zeta(t), \delta_1(t))$  are polynomials based on prospective values  $S_{mn}$ , which are at the same time specific target settings obtained from the setting action  $\bar{g}(t)$  of the national projects of the Russian Federation (Fig. 3).

Dynamics of the output coordinates  $\bar{y}(t)$  may be represented with the help of polynomial, characterizing their development in the temporary process. Values  $S_{mn}(t)$



**Fig. 3** Illustration of using retro and perspective data in forming controlling action  $u(t)$  and output coordinates  $y(t)$

change in increments of 1 year, or  $t = 1$ , where  $t \in N$ . So, dynamics of output coordinates  $y(t)$  can be described through increments of each component  $S_{mn}$  in time, and, therefore can be differentiated

$$y_{S_{mn}}(t) = y_{S_{mn}}(t_z) + \Delta y_{S_{mn}}, \text{ где } \Delta y_{S_{mn}} = y_{S_{mn}}(t_z) - y_{S_{mn}}(t_{z-1}), \quad (8)$$

where  $y_{S_{mn}}(t_z)$  is the output coordinate of  $m$ th matrix element  $S$  in  $z$ th year;  $y_{S_{mn}}(t_{z-1})$  is the output coordinate of  $m$ th matrix element  $S$  in the year previous to  $z$ th year.

Let the output coordinate  $y_{223}(t)$  for the state matrix element PSES  $S_{223}$  “gross regional product” is described by the polynomial in the 4th degree and its derivatives:

$$F = \begin{cases} y_{16}(t) = -262t^4 + 46536t^3 - 23554t^2 + 63994t + 368926; \\ y'_{16}(t) = -262*4t^3 + 46536 * 3t^2 - 23554 * 2t + 63994; \\ y''_{16}(t) = -262*4 * 3t^2 + 46536 * 3 * 2t - 23554 * 2 \end{cases} \quad (9)$$

where  $t$  is the sequence number of the observation period.

Analyzing the increments dynamics of the first and second derivatives for the output coordinate  $y_{223}(t)$  of the RSES state matrix element  $S_{223}$  “gross regional product” shows that changing their indicator  $y_{223}(t)$  occurs with an increase in speed, but, practically, with constant acceleration (Table 1). Therefore, we can talk about the stability of the increment indicator  $S_{223}$  in time, therefore, to describe the indicator  $y_{223}(t)$  as an element of the control object of the MS RSES, the application of the first and second derivatives will suffice.

Controlling action  $u(t)$  can be represented in the form of the tuple (6):

$$u(t) = S(t) \rightarrow S^*(t), \text{ Ra} | S(t) = ||S_{mn}||_{m=1, n=1}^{14 \times 69} \quad (10)$$

where  $S^*(t_k)$  is the desired matrix state  $S(t_z)$ ; Ra is a set of production rules for transition  $S(t_z)$  to the state  $S^*(t_k)$ .

Such an interpretation  $u(t)$  allows taking into account the change in the control object (RSES) from the current state  $S(t_z)$  to the state  $S^*(t_k)$  under the influence  $g(t)$ , and this will consider the external influences due to the system of production rules.

To describe  $u(t)$  you need to have predicted target values  $S^*_{mn}(t_k)$ , which can be approximated into new polynomials, which will describe the desired state  $S^*(t_k)$ .

**Table 1** Increment of the first and second derivative of the output coordinate  $y_{223}(t)$  of the state matrix element RSES  $S_{223}$  “gross regional product”

Period	Value $y_{223}(t)$	$\Delta y_{223}(t)$	Value $y'_{223}(t)$	$\Delta y'_{223}(t)$	Value $y''_{223}(t)$	$\Delta y''_{223}(t)$
$t = 1$	413,757	–	155,446	–	228,964	–
$t = 2$	435,730	21,973	519,826	364,380	498,748	269,784
$t = 3$	453,331	17,601	1,150,846	631,020	762,244	263,496
$t = 4$	478,758	25,427	2,042,218	891,372	1,019,452	257,208
$t = 5$	517,921	39,163	3,187,654	1,145,436	1,270,372	250,920
$t = 6$	506,448	– 11,473	4,580,866	1,393,212	1,515,004	244,632
$t = 7$	501,667	– 4781	6,215,566	1,634,700	1,753,348	238,344
$t = 8$	490,624	– 11,043	8,085,466	1,869,900	1,985,404	232,056
$t = 9$	454,077	– 36,547	10,184,278	2,098,812	2,211,172	225,768
$t = 10$	366,496	– 87,581	12,505,714	2,321,436	2,430,652	219,480

These values represent, on the one hand, the forecast for developing the indicator, and on the other hand, its target value is in accordance with the national projects. The resulting polynomials  $u(t)$  can successfully be differentiated.

For example, the national project “health care” within the framework of the federal project “combating cardiovascular diseases” sets a target indicator called “decreasing mortality from diseases of the circulatory system (per 100 thousand people)”, and its target values for 2019–2024 are given. The polynomial approximation of the indicator allowed us to form the following dependence to describe the controlling action  $u_{12}(t)$ , which was later differentiated:

$$F = \begin{cases} u_{12}(t) = -2.9333t^3 + 18.9t^2 - 78.767t + 650.4 \\ u'_{12}(t) = -2.9333 * 3t^2 + 18.9 * 2t - 78.767 \\ u''_{12}(t) = -2.9333 * 3 * 2t + 18.9 * 2 \end{cases} \quad (11)$$

where  $t$  is the sequence number of the observation period.

Analyzing the increment dynamics of the first and second derivatives for the controlling action  $u_{12}(t)$  “decreasing mortality from diseases of the circulatory system (per 100 thousand people)” shows, that changing the indicator  $u_{12}(t)$  is characterized by quickly reducing the absolute values of the indicator and by having constant negative acceleration (Table 2). Therefore, we can talk about the stability of the reduction indicator  $u_{12}(t)$  in time, therefore, the application of the first and second derivatives will suffice to describe it.

To describe the dynamics  $\overline{u(t)}$  from expression (6), only transforming the matrix of the real state  $S(t_z)$  into the matrix of the desired state can be used  $S^*(t_k)$ , as  $S_{mn}$ , which are matrix elements lend themselves well to differentiation. It is not possible to evaluate the changes in time of production rules due to their fuzzy linguistic nature. Their presence in the description  $\overline{u(t)}$  is necessary to prioritize the RSES management tasks.



**Table 2** Increment of the first and second derivatives of the controlling action  $u_{12}(t)$  “decreasing mortality from diseases of the circulatory system (per 100 thousand people)”

Period	Value $u_{12}(t)$	$\Delta u_{12}(t)$	Value $u'_{12}(t)$	$\Delta u'_{12}(t)$	Value $u''_{12}(t)$	$\Delta u''_{12}(t)$
$t = 1$	587.6	–	–49.77	–	20.21	–
$t = 2$	545	–42.6	–38.37	11.4	2.61	–17.6
$t = 3$	505	–40	–44.57	–6.2	–15	–17.61
$t = 4$	450.01	–54.99	–68.37	–23.8	–32.6	–17.6
$t = 5$	362.41	–87.6	–109.77	–41.4	–50.2	–17.6

## 5 Conclusion

Using the control theory to describe the dynamic nature of the RSES allows us to show the causal nature of the phenomena of managerial influence and forming output coordinates. The proposed dependencies are the basis for the subsequent formation of a general mathematical model MS RSES, which describes the impact of the national projects of the Russian Federation on the subject of the Russian Federation of the “region” type.

## References

1. Burkov, V.N., Korgin, N.A., Novikov, D.A.: Introduction to the theory of management of organizational systems
2. Novikov D.A.: Theory of Management of Organizational Systems. Fizmatlit, Moscow, 584p (2007)
3. Novikov, D.A.: Introduction to the Theory of Educational Systems Management. Egves, Moscow (2009)
4. Voronin, A.A., Gubko, M.V., Mishin, S.P., Novikov, D.A.: Mathematical Models of Organizations. Lenand, Moscow, 360p. (2008)
5. Burkov, V.N., Danev, B., Enaleev, A.K. et al.: Large Systems: Modeling of Organizational Mechanisms. Nauka, Moscow (1989)
6. Chkhartishvili, A.G.: Game-theoretic Models of Information Management. PMSOFT, Moscow (2004)
7. Ajzerman, M.A., Aleskerov, F.: Choice of Options: Basic Theory, Nauka, Moscow GL. ed. Fiz.-Mat. lit. 240p. (1990)
8. Kostikova, A.V., Trelaske, P.V., Shuvaev, A.V., Parahina, V.N., Timoshenko, P.N.: Expert fuzzy modeling of dynamic properties of complex system. ARPN J. Eng. Appl. Sci. **11**(17), 10601–10608 (2016)
9. Tirelessly, P.V., Koroteev, M.V., Vasil'ev, O.I., Baktygulov, K.B., Ordabaev, B.S.: The variability of fuzzy aggregation methods for partial indicators of quality and the optimal method choice. ARPN J. Eng. Appl. Sci. **11**(13), 8312–8319 (2016)
10. Taha, H.A.: Chapter 14. Game theory and decision making. Introduction to operations research = Operations Research: An Introduction. 7-e Izd. – Moscow, “Williams”, S. 549–594 (2007) ISBN 0-13-032374-8
11. Hansson, S.O.: «Decision theory: a brief introduction», <https://web.archive.org/web/20060705052730/https://www.infra.kth.se/~soh/decisiontheory.pdf> (an excellent non-technical and fairly comprehensive primer)

12. Goodwin, P., Wright, G.: *Decision Analysis for Management Judgment*, 3rd edn. Wiley, Chichester (2004). ISBN 0-470-86108-8 (covers both normative and descriptive theory)
13. Clemen, R.: *Making Hard Decisions: An Introduction to Decision Analysis*, 2nd ed. Duxbury Press, Belmont CA (1996) (covers normative decision theory)
14. Brams, S.J., Taylor, A.: *Fair Division*. Cambridge University Press, New York (1996)
15. Raiffa, H.: *Decision Analysis: Introductory Readings on Choices Under Uncertainty*. McGraw Hill (1997)
16. Roth, A., Sotomayor, M.O.: *Two-Sided Matching*. Cambridge University Press, Cambridge (1990)
17. Smith, J.Q.: *Decision Analysis: A Bayesian Approach*. Chapman and Hall (1988)
18. Litvak, B.G.: *Development of Management Decisions*, Case, Moscow, 392p. (2004)
19. Orlov A.I.: Nonparametric method of least squares: accounting for seasonality. *J. Math. Sci.*
20. Sukharev, O.S.: Regional economic policy: structured approach and tools (theoretical formulation). *Reg. Econom.* **2**, 9–22 (2015)
21. Sukharev O.S.: Elements of the theory of self-development of regional economy: structure and management. *Vestnik AKSOR*, N 1 (2017)
22. Tatarkin, A.I.: The Theoretical Basis of Institution of Self-development of Socio-economic Systems. *Institutes of the Modern Economy*. T4, Aleteya, St. Petersburg, pp. 87–160 (2015)
23. Averchenkova, E.E., Averchenkov, A.V., Kulagina, N.A.: Designing of the information advising system to assess the potential of creation and development of cluster agglomeration in the industrial complex of the region. *International Conference on Information Technologies in Business and Industry 2016*. IOP Conf. Series: Journal of Physics: Conference Series **803** (2017) 012011. <https://doi.org/10.1088/1742-6596/803/1/012011>(Article ID in SCOPUS: 2-s2.0-85018367471)
24. Averchenkov, A.V., Averchenkova, E.E., Gorlenko, O.A., Miroshnikov, V.V.: Machine-building enterprise fuzzy model as the interrelated factor complex system. In: *International Conference on Information Technologies in Business and Industry 2016*. IOP Conference Series: Journal of Physics: Conference Series, vol. 803, 012009 (2017). <https://doi.org/10.1088/1742-6596/803/1/012009> (article ID in SCOPUS: 2-s2.0-85016636779).

# Transformation, Visualization and Analysis Different Kind of Study Information Contained in the Students' Electronic Portfolio



Elena Ilina , Yuliya Kocherzhinskaya , Nikita Dyakonov ,  
Daria Arefeva , Tat'yana Antonova , and Il'ya Levandovskii 

**Abstract** Software for accounting and systematization in educational activities is developed for the effective work of teachers and students. In Magnitogorsk State Technical University, every student who has achievements in educational, research, public, cultural, creative or sports activities has the opportunity to be assigned to an increased state academic scholarship (Order of the Ministry of Education and Science of the Russian Federation of December 27, 2016, No. 1663). For this purpose, on the educational portal of the Magnitogorsk State Technical University a student's portfolio is filled in for each of the activities, which is the link between the student and the teacher. For accounting and systematization of educational activities, a web module was created, integrated into the educational portal. Throughout the course of study, the student has the opportunity to view statistics and achievements in his/her academic activities, add, evaluate and analyze information for further effective study, obtaining increased scholarships and successfully defending graduate qualifying work. Reducing the time and labor costs required to collect and systematize the achievements in student learning activities will make the work of teachers more efficient and productive when interacting with students.

**Keywords** Data transformation · Visualization of information · Learning activities · Electronic portfolio systems · Software design

## 1 Introduction

In the Russian Federation, students enrolled in educational programs of higher education (bachelor's degree, specialty, master's degree programs), including foreign citizens (full-time students, budget) who have achievements in educational, research, public, cultural, creative or sports activities, provide set of their individual achievements in various fields of activity, i.e., fill out the portfolio.

---

E. Ilina · Y. Kocherzhinskaya (✉) · N. Dyakonov · D. Arefeva · T. Antonova · I. Levandovskii  
Nosov Magnitogorsk State Technical University, Magnitogorsk, Russia  
e-mail: [y.kocherzhinskaya@mail.ru](mailto:y.kocherzhinskaya@mail.ru)

The development of technology and the growth of the information with which to work lead to the automation of time-consuming activities. The actual problem is the collection and systematization of documents for subsequent verification, evaluation and analysis of students' achievements for later enrollment in the increased scholarship.

Automating these criteria allows you to simplify this process by reducing time and effort. Students need to monitor the effectiveness of their work throughout the entire period of study. To this end, software has been created for recording and systematizing the achievements, which consists of several modules responsible for different types of activity. So that students and teachers can keep records, analyze, systematize and add information on educational activities, and it was decided to create a web module based on the LMS Moodle distance learning system. Educational activities include such categories as the quality of training in academic disciplines; practice; research work; term papers and projects; candidate exams; absolute and high performance; state final certification; Olympiads and contests; project activities; studying massive open online courses; online exam; mastering foreign languages; mastering additional competencies; internships and academic mobility.

Currently, personality development in the process of education is gaining momentum in our country. The transition to a market economy has set somewhat different priorities in our society. The development of research activities in all spheres of life also affects the need for active, independently minded specialists, who, along with the ability to adapt to emerging conditions, could change them taking into account the new situation of professional activity, would be able to adequately assess the changes that occurred, have experience in self-fulfilling research competence. This leads to a complication of the mechanisms of education and, accordingly, the activities of people engaged in this field [1–3].

The system of electronic portfolio is a program of individual-oriented professional development of a student, which includes the collection, systematization, processing, accumulation and analysis of the results of real changes and individual achievements in the process of studying.

The e-portfolio abroad is not something new and has long been used in the field of education, in the USA such an idea arose in the mid-1985. The foreign market, in contrast to the Russian one, offers a wide range of ready-to-use IT solutions. All of them have different functionalities, and educational institutions can choose the most suitable system from numerous options, so they are more likely to buy such products than to develop their own. Such systems are often closed, and you can try them either by using the demo mode or by paying for the product. To use the demo mode, you must directly contact the company. And free solutions have poor functionality, and more are aimed at personal use [4–6].

In the Russian Federation, electronic portfolio systems have appeared relatively recently, but are already a requirement of educational programs of higher education of modern Federal State Educational Standards (FSES). Portfolio is filled for such activities as research, educational, cultural and creative, public and sports. Based on the achievements presented in the portfolio and the absence of academic debts and grades “satisfactorily”, an increased state academic scholarship is appointed

(Order of the Ministry of Education and Science of the Russian Federation dated December 27, 2016, No. 1663). There are many resources on the Internet that allow you to develop a portfolio of students. The most famous of these resources has one or another functionality of the social network: building relationships with other users of the portfolio system, messaging, rating, etc. This situation is justified by the fact that the portfolio is created to present to other people, to receive their assessment and to recognize the achievements of the author of the portfolio in various situations. This allows you to talk about a different model of developing a portfolio system when such a service is created on the basis of an existing software product [7–10].

Such a model of a social network can be the foundation for creating an educational portal. An educational portal is a Web site where, first of all, students and teachers are presented, and there are flexible possibilities for their interaction and joint educational activities in a virtual online environment. The educational portal of Nosov Magnitogorsk State Technical University has realized most of the functions of the electronic portfolio, and it includes the program complex “improving student scholarships”, which was created to facilitate decision making on the appointment of “increased” scholarships [11, 12]. The educational portal is based on the LMS Moodle distance learning system. In order to allow students to monitor the effectiveness of their work throughout the entire training period, a web module was created based on the LMS Moodle distance learning system [13].

## 2 Purpose and Methods of Research

### 2.1 Purpose of the Research

Purpose: improving the efficiency of the educational process and reducing the time spent by teachers and students. The subject of the research is the analysis of the effectiveness of the work of students and teachers. The object of the research is the electronic portfolio system and educational portal of the Nosov Magnitogorsk State Technical University.

To achieve this goal is to solved the following tasks:

1. analysis of the electronic portfolio formation systems;
2. study of the principles of work of LMS Moodle for the development of an electronic portfolio;
3. development of a module for analyzing the performance of students;
4. implementation of the module in the learning process and on the educational portal of the Nosov Magnitogorsk State Technical University.

## 2.2 Methods

When designing a web module for recording and systematizing the achievements in educational activities, the following requirements were taken into account.

The ability is to view statistics on educational activities. Fulfillment of this requirement will allow the student to see his grades in academic disciplines, in terms of papers, as well as educational achievements, which the user will add manually.

Adding information about the achievement. The achievement is added to the database table and after that the change is displayed on the page. If the user did not fill in all the fields or an error occurred while adding to the table, the system displays a corresponding message on the page.

Changing achievement information. This feature will fix an incorrectly completed achievement. If the user did not fill in all the fields or an error occurred while changing the achievement in the table in the database, the system displays a corresponding message on the page.

Removal of information about achievement. If the user has not selected the achievement to be deleted or an error has occurred during the deletion, the system will display a corresponding message on the page.

Loading of information on achievements on the page.

Work with files confirming the reality of achievement. This feature includes loading, storing, modifying and deleting files owned by the user.

The following parts are highlighted in the web module:

1. client part (frontend);
2. server part (backend).

**Development of the server part of the web module.** In the server part, many of the built-in functions of the PHP programming language are used in the development of the web module. Table 1 presents a description of some of them [14].

Since the data sharing scheme of the model–view–controller web application, smarty, a compiling template handler for PHP, was used. The smarty functions used were described in Table 2.

For the development of the server side, the language PHP is chosen—a general-purpose scripting language that is widely used for developing web applications is used in the LMS Moodle environment.

**Table 1** Description of some of the built-in language functions used when writing the program

Function	Description
<code>mysql_gettable</code>	Performs a table query to the database
<code>mysqli_query</code>	Fulfills database query
<code>echo</code>	Allows you to display lines on the screen
<code>isset</code>	Determines whether a variable is set
<code>json_encode</code>	Returns JSON encoded string (on success) or FALSE if an error occurs

**Table 2** Description of the smarty template handler functions used to write the web module

Function	Description
display	Displays a template
assign	Assigns a value to a pattern

The web server assembly usually includes at least three components: HTTP server, site development tool (programming language library, interpreter, etc.) and database management system. For such assemblies, free software is used, so the most common components are Apache web server, MySQL DBMS, programming languages PHP and Perl.

As a web server, XAMPP version 7.1.11 is used, which contains Apache, MySQL, PHP script interpreter and a large number of additional libraries that allow you to run a full-fledged web server.

XAMPP is one of the popular builds of a ready-made local server. The popularity of the server is due to the quality of the product, a sufficient number of tools, information support [15].

MySQL was chosen as the database management system. MySQL will have good speed, reliability and flexibility. Working with her, as a rule, does not cause great difficulties [16]. MySQL server support is automatically included in the PHP package.

The following technologies were used to implement the server part:

1. PHP is a general-purpose scripting language used for developing web applications.
2. MySQL is an open-source relational database management system.
3. Smarty is a compiling template handler for PHP, one of the tools for separating application logic and data from representation in the model–view–controller concept. Smarty is intended to simplify compartmentalization, allowing the front end of a web page to change separately from its back end. Ideally, this lowers costs and minimizes the efforts associated with software maintenance [17].

**Development of the client part of the web module.** In the client part, several built-in functions of the JavaScript programming language were used to develop the web module. Table 3 provides a description of the functions used.

**Table 3** Description of JavaScript functions used in the development of the web module

Function	Description
alert	Displays a modal message box
create_error_alert	Error message popup
confirm	Displays a message in a window with two buttons: OK and CANCEL. Returns true/false depending on where the user clicks

**Table 4** Description of some jQuery methods used to write a web module

Methods	Description
val	Gets the current attribute value from the first element in the set of matched elements
wrap	Places selected items inside the specified item
attr	Gets the attribute value for the first element in the set of matched elements
find	Searches for items within already selected items
html	Gets or changes the HTML content of the selected items
ajax	Performs a request to the server without reloading the page
load	Allows you to download data from the server and place the returned HTML code inside the necessary elements

jQuery was used to interact JavaScript and html. jQuery is a JavaScript library that helps you easily access any DOM element, access and manipulate the attributes and contents of DOM elements. The DOM is a platform and language-independent programming interface that allows programs and scripts to access the contents of HTML, XHTML and XML documents, as well as change the content, structure and design of such documents. The library also allows the client and server parts to interact through AJAX requests. Some of the jQuery library methods used in the implementation of the web module are presented in Table 4.

As the development environment was used LMS Moodle, which is focused on the interaction of the teacher and the student. The choice of development environment is due to the fact that it is based on educational portal of the Nosov Magnitogorsk State Technical University, as well as an open-source distance learning environment.

The following technologies were used to implement the client side:

1. HTML is a markup language for web pages;
2. CSS is a language for describing styles of elements of HTML pages;
3. Bootstrap—a free set of tools for creating Web sites and web applications, which includes HTML and CSS page design templates [18].
4. JavaScript—HTML scripting language;
5. AJAX—a set of techniques for building interactive user interfaces of web applications, which consists in the “background” data exchange browser with a web server;
6. jQuery is a JavaScript library to simplify working with HTML elements on the client side [19, 20].

### 3 Results

#### 3.1 *The Structure of the Web Module*

The web module implements five modules:



1. module for loading data on the page;
2. module for opening a form for adding or changing information about the achievement;
3. module for adding or changing achievement information;
4. module for deleting information about the achievement;
5. module for working with files confirming the validity of the achievement.

Module for loading data on the page. When the page loads, as well as adding, changing or deleting achievements, an AJAX request to the server occurs, where achievement data is requested from tables in the database. The server returns the generated data that is displayed in a table on the web module page. This module allows you to update information without a full page reload.

The module for opening the form of adding or changing information about the achievement. After clicking the “Add” or “Change” button, a value is transmitted that determines which button was activated. Clicking on the “Add” button will open a form with empty fields. Clicking on the “Change” button will check for the selected achievement. If an achievement has been selected, a form will open and an AJAX request will be sent to the server, where achievement data from the table in the database will be requested. The server returns data that will be displayed in the fields on the open form. In case of success, data from the form and the transferred value is transferred to the module for adding or changing information on the achievement by clicking on a certain button.

Module for adding or changing achievement information. After clicking on the “Save” button on the form, data from fields on the form will be checked for emptiness, if one of the fields has not been filled out on the page, a notification will be displayed and the module will stop working. If successful, an AJAX request will be sent to the server, where, depending on the transferred value, the data will either be added or changed in a table in the database. The server returns a response about the successful addition or change. If the answer is positive, the module loads the data to the page. Otherwise, the page displays an error message.

The module for deleting information about the achievement. After clicking on the “Delete” button, a check will be made on the selected achievements. If an achievement has been selected, the system will ask you to confirm the deletion, otherwise the page will display a notification that you need to select an achievement. When you confirm the deletion, an AJAX request will be sent to the server, where the achievement will be deleted from the table in the database. The server returns an answer about the deletion success. If the answer is positive, the module loads the data to the page. Otherwise, the page displays an error message.

Module work with files confirming the reality of achievement. This feature includes loading, storing, modifying and deleting files owned by the user.

The flowchart of the web module is shown in Fig. 1.

Flowchart of the `senddata_click()` function for sending a request to add information to a table in the database is shown in Fig. 2.

Flowchart of the function of updating the page content on the example of the function `intship_init()` is presented in Fig. 3.

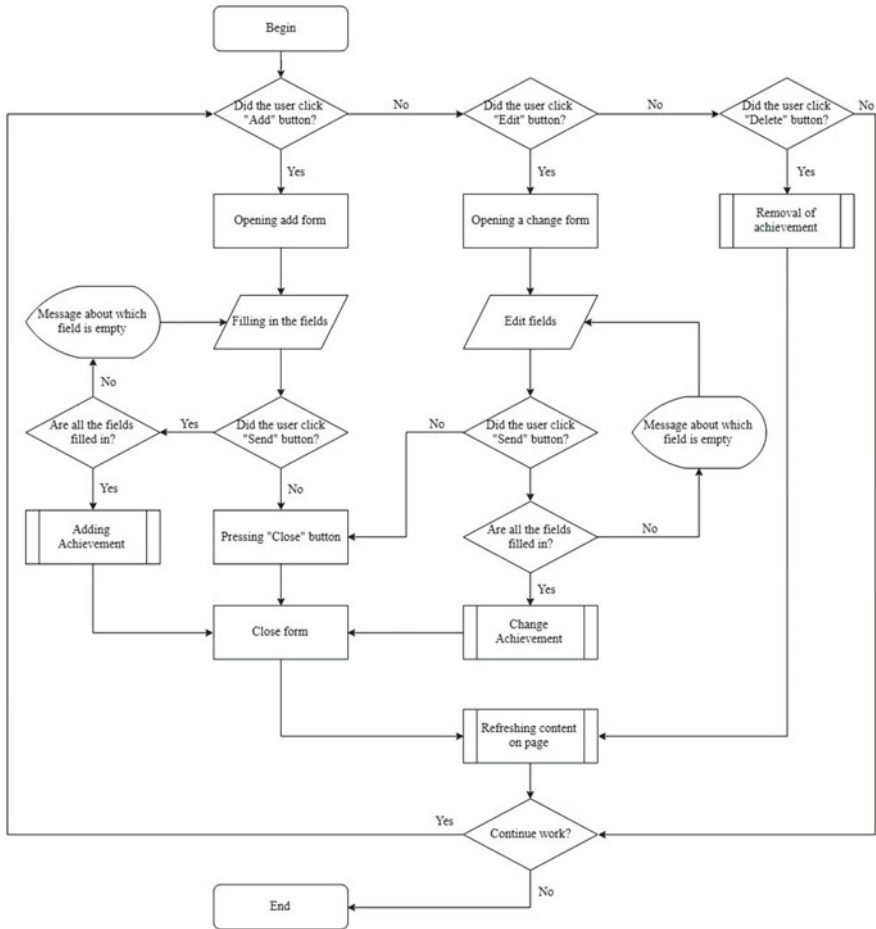


Fig. 1 Flowchart of the web module

The work of the web module is divided into five stages:

1. At this stage, in the index.php file, the user’s access rights are checked to view the contents of the web module, and the display function of the portf\_input.tpl template is called.
2. The portf\_input.tpl template is displayed.
3. After the display, data is loaded into the module from the database, the system waits for further user actions, during which it is possible to add, change or delete information for some tables.
4. At this stage, the forms are displayed, which are caused by clicking on the “Change” or “Add” button, the user fills out the form and presses the “Save” button.

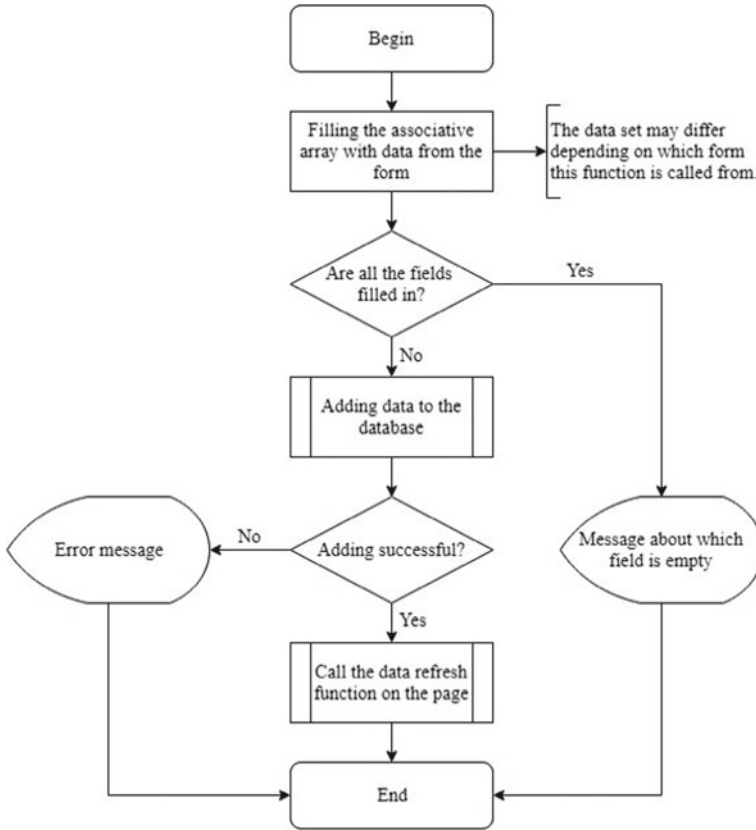


Fig. 2 Flowchart of the function senddata\_click()

- 5. After clicking on the “Save” button, the information is updated or added to the database, and the data in the table on the displayed page is updated.

The structure of the web module is presented in Fig. 4.

File index.php is responsible for checking the user’s access rights to view the contents of the web module, as well as for displaying the template portf.tpl.

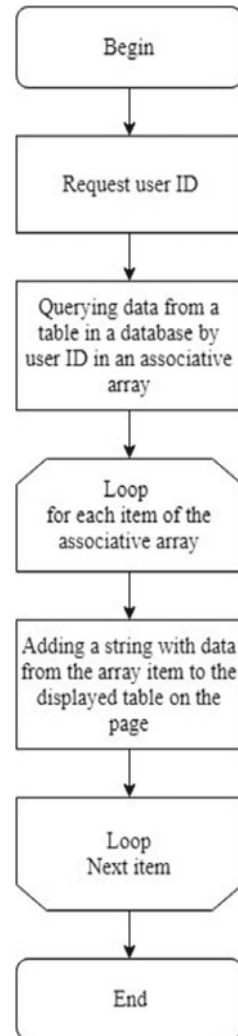
File portf.tpl is main page of the module, which implements the module interface and AJAX request functions.

Files recbook.php, candexam.php, acadperf.php, stfinatt.php, olymp.php, proj.php, mook.php, intexam.php, forlang.php, advcomp.php, intship.php—necessary for uploading data to tables on the main page.

File aj\_del is responsible for removing achievements from tables in the database.

Files form\_olymp.php, form\_mook.php, form\_proj.php, form\_intship.php, form\_forlang.php, form\_intexam.php, form\_advcom.php are responsible for displaying form templates such as form\_olymp.tpl, form\_mook.tpl, form\_proj.tpl,

**Fig. 3** Flowchart of the function `intship_init()`



`form_intship.tpl`, `form_forlang.tpl`, `form_intexam.tpl`, `form_advcom.tpl`, respectively, and also for transferring data from the main page to forms using queries.

Files `form_olymp.tpl`, `form_mook.tpl`, `form_proj.tpl`, `form_intship.tpl`, `form_forlang.tpl`, `form_intexam.tpl`, `form_advcom.tpl`—form templates that contain the fields required to complete the student’s achievements. With the help of requests, the data is transferred to `aj_olymp.php`, `aj_mook.php`, `aj_proj.php`, `aj_intship.php`, `aj_forlang.php`, `aj_intexam.php`, `aj_advcom.php` files, respectively.

Files `aj_olymp.php`, `aj_mook.php`, `aj_proj.php`, `aj_intship.php`, `aj_forlang.php`, `aj_intexam.php`, `aj_advcom.php` are responsible for adding or changing achievements in tables in the database.

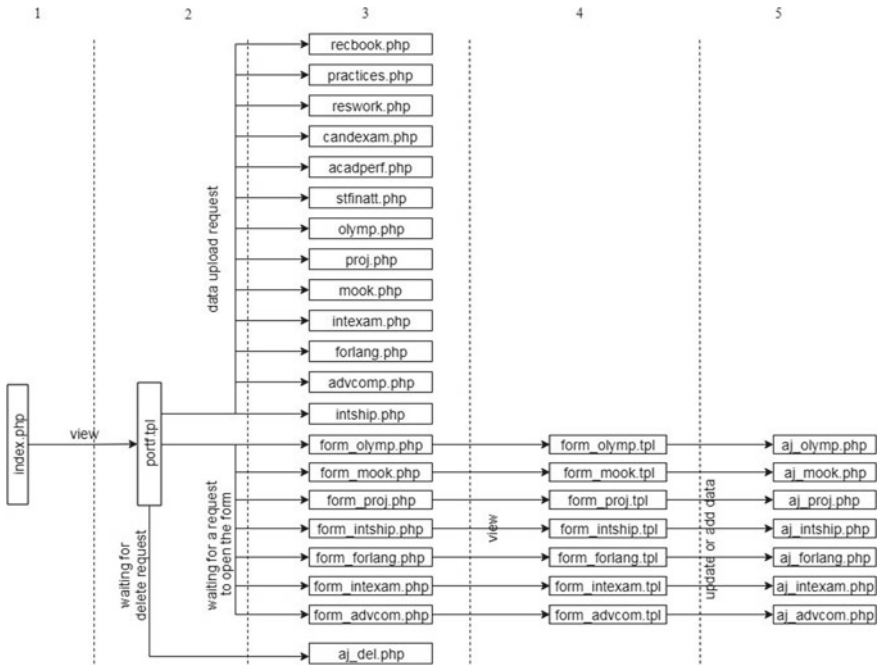


Fig. 4 Structure of the web module

### 4 Discussion and Conclusions

As a result, a web module was designed to add achievements in training activities and record achievements (analysis, comparison and systematization), implemented on the basis of the LMS Moodle learning environment. In the program module, two important parts were singled out: the client part and the server part.

When designing the web module, the following requirements were imposed: the ability to view statistics on training activities; adding achievement information; change of information on achievement; deletion of achievement information; uploading information about achievements to the page; work with files confirming the reality of achievement. A scheme of the web module as a whole and its individual functions is described, as well as a flowchart showing the operation of the algorithms.

### References

1. Poole, P., Brown, M., McNamara, G., O'Hara, J., O'Brien, S., Burns, D.: Challenges and supports towards the integration of ePortfolios in education. Lessons to be learned from Ireland. *Heliyon* 4(11) (2018)

2. Pullman, G.: Electronic portfolios revisited: the efolios project. *Comput. Compos.* **19**(2), 151–169 (2002)
3. Shepherd, C., Bolliger, D.: The effects of electronic portfolio tools on online student’s perceived support and cognitive load. *Internet High. Educ.* **14**(3), 142–149 (2011)
4. Mohammed, A., Mohssine, B., M’hammed, K., Mohammed, T., Abdelouahed, N.: Eportfolio as a tool of learning, presentation, orientation and evaluation skills. *Procedia Soc. Behav. Sci.* **197**, 328–333 (2015)
5. Huang, A., Wu, J., Yang, S., Hwang, W.: The success of ePortfolio-based programming learning style diagnosis: exploring the role of a heuristic fuzzy knowledge fusion. *Expert Syst. Appl.* **39**(10), 8698–8706 (2012)
6. Miller, A.: Professional learning ecosystem support for ePortfolio use in Australian higher education: an historical perspective. In: *ePortfolios in Australian Universities*, pp. 1–11 (2016)
7. Jones, S., Downs, E., Jenkins, S.: Transparency in the ePortfolio creation process. *TechTrends* **59**(3), 64–70 (2015)
8. Huang, C., Huang, Y., Yang, J., Wang, W.: A study of the wikipedia knowledge recommendation service for satisfaction of ePortfolio users. In: *Advanced Technologies, Embedded and Multimedia for Human-Centric Computing*, pp. 283–289 (2013)
9. McAllister, L.: An ePortfolio approach: supporting critical reflection for pedagogic innovation. In: *Teaching Reflective Learning in Higher Education*, pp. 173–187 (2014)
10. Giorgini, F.: An interoperable ePortfolio tool for all. In: *Sustaining TEL: From Innovation to Learning and Practice*, pp. 500–505 (2010)
11. The educational portal of Nosov Magnitogorsk State Technical University [Online]. Available: [https://newlms.magtu.ru/report/increased\\_grant](https://newlms.magtu.ru/report/increased_grant). Last accessed May 2019
12. Fizers, M.: *Working Effectively with Legacy Code*. Vilyam (2016)
13. LMS Moodle [Online]. Available: <https://moodle.org>. Last accessed May 2019
14. PHP Documentation [Online]. Available: <https://www.php.net/docs.php>. Last accessed May 2019
15. XAMPP [Online]. Available: <https://www.apachefriends.org/index.html>. Last accessed May 2019
16. MySQL Documentation [Online]. Available: <https://dev.mysql.com/doc/>. Last accessed May 2019
17. Smarty Documentation [Online]. Available: <https://www.smarty.net/documentation/>. Last accessed May 2019
18. Bootstrap 3.3.5 Documentation [Online]. Available: <https://bootstrapdocs.com/v3.3.5/docs/>. Last accessed May 2019
19. jQuery API Documentation [Online]. Available: <https://api.jquery.com/>. Last accessed May 2019
20. jQuery Russian Documentation [Online]. Available: <https://jquery-docs.ru/>. Last accessed May 2019

# An Automated Measuring Complex for Research Parameters of Unmanned Aerial Vehicle



Oleg Drozd , Pavel Avlasko , Semen Bordyugov , and Denis Kapulin 

**Abstract** In research, design and development of unmanned aerial vehicles (UAV), a key role belongs to automated measuring complexes for its simulation and prototyping electromechanical system with digital control. Design and debugging of such system are suggested and performed by using model-based approach with automated simulation tools. This approach means creating and using in further the simulation model of the measuring complex. In the paper, the simulation model for research parameters of unmanned aerial vehicles is discussed. This model is suggested to develop based on Gough–Stewart platform (six-axes platform manipulator) with UAV mounted on it. The researched model includes the UAV’s trajectory generator, automatic control device for UAV, kinematic manipulator model and decision subsystem to define the current attitude of the manipulator dynamic side. The model allows to evaluate the functioning parameters of proportional–integral–derivative controllers of spatial orientation angles, as well as to automatically obtain program code for implementing both automation of testing and the UAV control device. Also, the presented model can function in conjunction with the user interface of the measuring complex and can be used to evaluate the parameters of complex functioning.

**Keywords** Unmanned aerial vehicle · Simulation · Functional modeling · Gough–Stewart platform · Model-based approach · Electromechanical test system

## 1 Introduction

Currently, approaches to debugging and testing an unmanned aerial vehicle (UAV) are being actively developed using unit testing tools for UAV onboard systems based on simulation and hardware–software modeling [1, 2]. In this case, the key role is played by complex measuring systems (a set of test equipment) with full or partial imitation of onboard control system functioning. In a consolidated manner, it is a

---

O. Drozd (✉) · P. Avlasko · S. Bordyugov · D. Kapulin  
Siberian Federal University, Krasnoyarsk, Russia  
e-mail: [odrozd@sfu-kras.ru](mailto:odrozd@sfu-kras.ru)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021  
N. Voinov et al. (eds.), *Proceedings of International Scientific Conference on Telecommunications, Computing and Control, Smart Innovation, Systems and Technologies* 220, [https://doi.org/10.1007/978-981-33-6632-9\\_37](https://doi.org/10.1007/978-981-33-6632-9_37)

419

complex of hardware and software tools that ensures conducting the specified tests of UAV control system in order to assess the impact of disturbing influences on the characteristics of the test object and its operational parameters [3, 4].

To investigate traction characteristics, kinematic features of the structure and other key performance indicators of autonomous objects, it can be used with the well-known Gough–Stewart platform (platform manipulator) with an octahedral arrangement of prismatic actuators, or legs all of them connected simultaneously to a fixed base and a moving platform through spherical joints or attachments [5–7]. Such manipulators are used for research and design both in space engineering, and in production of new appliance, medical equipment, etc. A similar platform can be successfully used for researching of UAV operation modes. For this purpose, the UAV movement should be measured in six degrees of freedom for the manipulator with installation of corresponding sensitive elements on its legs. The development and debugging of such software and hardware measuring structure are a difficult engineering activity that can be solved by using the model-based design approach, which is considered in this paper [8–10].

## 2 Mathematical Description of the Platform Manipulator

The central element of the model-based design approach is a simulation model of the device under research and development—the Gough–Stewart hexapod manipulator with a servomotor driver system, motion control and measurement of the tested product physical parameters [11]. The model representation of the platform manipulator consists of a fixed base, a moving platform modeled by a disk of given mass and six extensible supports (legs), mass of which is taken equal to zero (Fig. 1).

**Fig. 1** Gough–Stewart platform manipulator H-850 (Physik Instrumente, Germany)





### 2.1 Kinematics of the Platform Manipulator

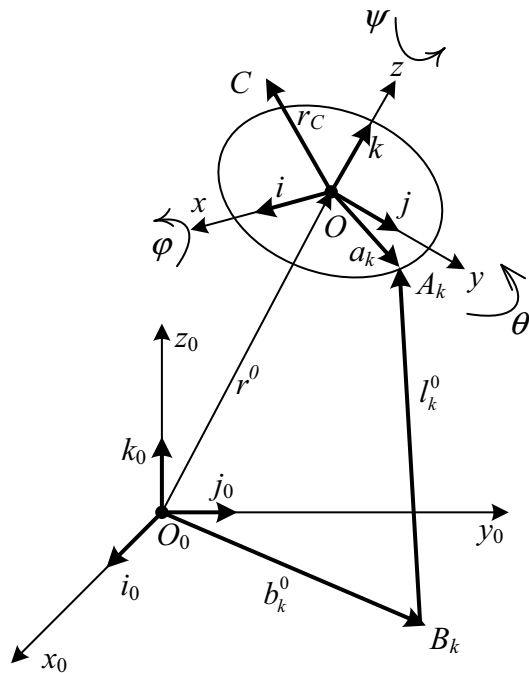
The classical method of obtaining a model of a closed kinematic system consists in examining the equivalent tree structure and the restrictions imposed on it by means of the Lagrange multipliers or the D'Alembert's principle [12, 13]. Other possible approaches are based on the use of the Newton–Euler equations [14, 15], as well as special forms of the motion equations of a rigid body [16].

It will discuss the mathematical model of the platform manipulator used in this research below. For this purpose, it must put a fixed coordinate system  $O_0x_0y_0z_0$  with unit vectors  $i_0, j_0, k_0$  and a moving coordinate system  $Oxyz$  with unit vectors  $i, j, k$ , rigidly connected to the moving platform (Fig. 2). Six extensible legs  $B_kA_k$  ( $k = \overline{1, 6}$ ) are secured by spherical hinges at the one end to the fixed points  $B_k$  of the coordinate space  $O_0x_0y_0z_0$ , and at the other points to the platform at the points  $A_k$  ( $k = \overline{1, 6}$ ). It is required to obtain the specified parameters the trajectory of platform movement by changing the length of the supports.

The orientation of the moving platform is determined by the position of the point  $O$  (pole):

$$\overrightarrow{O_0O} = r_0(t) = x_0r(t)i_0 + y_0(t)j_0 + z_0(t)k_0,$$

**Fig. 2** Kinematics of a six-axis platform manipulator



and three successive Eulerian angles of platform rotation around the pole at the yaw ( $\psi$ ), pitch ( $\theta$ ) and roll ( $\phi$ ) angles. The platform rotation tensor  $P(\psi, \theta, \phi)$  is determined by the equality:

$$P = \begin{pmatrix} p_{11} & p_{12} & p_{13} \\ p_{21} & p_{22} & p_{23} \\ p_{31} & p_{32} & p_{33} \end{pmatrix} = \begin{pmatrix} C_\psi C_\theta - S_\psi C_\phi + C_\psi S_\theta S_\phi & S_\psi C_\phi + C_\psi S_\theta C_\phi \\ S_\psi C_\theta & C_\psi C_\phi + S_\psi S_\theta S_\phi & -C_\psi S_\phi + S_\psi S_\theta C_\phi \\ -S_\theta & C_\theta S_\phi & C_\theta C_\phi \end{pmatrix},$$

where  $C_\phi = \cos\phi$ ,  $S_\theta = \sin\theta$ , etc.

It is shown Poisson's equation for the time derivative of the rotation tensor  $P$ :

$$\begin{aligned} \dot{P} &= \omega^0 \times P, \\ \omega^0 &= \omega_x^0 i_0 + \omega_y^0 j_0 + \omega_z^0 k_0 = \omega_x i + \omega_y j + \omega_z k, \\ \omega_x^0 &= \dot{\psi} \cos \theta \cos \psi - \dot{\theta} \sin \psi, \omega_y^0 = \dot{\psi} \cos \theta \sin \psi - \dot{\theta} \cos \psi, \omega_z^0 = \dot{\psi} - \dot{\phi} \sin \psi, \end{aligned}$$

where  $\omega_0$ —an angular velocity of rotation the platform.

Below, it puts a vector of generalized coordinates that determine the position of the moving platform:

$$q = \{q_i\} = \{x_0, y_0, z_0, \varphi, \theta, \psi\}. \quad (1)$$

If values (1) are given, then the lengths of the supports  $l_k$  and their directions  $e_{kt}^0$  are determined by explicit formulas:

$$\overrightarrow{B_k A_k} = l_k^0 = l_k e_{kt}^0 = r^0 + P \cdot a_k - b_k^0, \quad k = \overline{1, 6},$$

where the constant vectors  $a_k = \overrightarrow{O A_k}$  and  $b_k^0 = \overrightarrow{O_0 B_k}$  specify coordinates of the points  $A_k$  and  $B_k$  of the support attachment in the moving and fixed coordinate systems, respectively.

When specifying the extensible support length  $l_k$  to determine the coordinates (1), it is necessary to solve a system of six nonlinear equations

$$(r^0 + P \cdot a_k - b_k^0)^2 = l_k^2, \quad k = \overline{1, 6}, \quad (2)$$

in relation to quantities (1) entering into  $r_0$  and  $P$ .

After differentiating Eq. (2) by time, it will obtain a system of linear equations relatively to  $\omega^0$ , which can also be represented in a matrix form, while the matrix  $A$  will be composed of row vectors  $L_k$ :

$$\begin{aligned} r_0 \cdot e_{kt}^0 + (\omega^0 \times P \cdot a_k) \cdot e_{kt}^0 &= L_k \cdot V^0 = i_k, \quad k = \overline{1, 6}, \\ V_0 = \{\dot{r}^0, \omega^0\} &= (\dot{x}_0, \dot{y}_0, \dot{z}_0, \omega_x^0, \omega_y^0, \omega_z^0)^T, L_k = \{e_{kt}^0, a_k^0 \times e_{kt}^0\}, a_k^0 = P \cdot a_k, \quad (3) \end{aligned}$$

$$A \cdot V^0 = i, \quad i = (i_1, \dots, i_6)^T. \tag{4}$$

Thus, if the support length  $l_k$  is given as function of time, then, solving system (3) or (4), it can find the coordinates of the vector  $V_0$ . In derivatives  $\dot{\varphi}$ ,  $\dot{\theta}$  and  $\dot{\psi}$ , we find by formulas:

$$\dot{\varphi} = \frac{\omega_y^0 \sin \psi + \omega_x^0 \cos \psi}{\cos \theta}, \quad \dot{\theta} = \omega_y^0 \cos \psi - \omega_x^0 \sin \psi, \quad \dot{\psi} = \omega_z^0 + \dot{\varphi} \sin \theta.$$

Now we obtain the quantities  $q_i(t)$  by integration, assuming that at the initial time  $t = 0$  the values  $q_i(0) = q_i^0$  are known. Movement is possible as long as the determinant of the matrix  $A$  is different from zero. It is vanishing to zero which indicates an exit to the boundary of the controllability domain.

## 2.2 Differential Equations of Platform Manipulator Motion

The differential equation of center of mass motion  $C$  for the moving platform with a static load applied to it in the fixed coordinate system can be written as follows:

$$m(\ddot{r}^0 + \dot{\omega}^0 \times r_c^0 + \omega^0 \times (\omega^0 \times r_c^0)) + mgk_0 = F^0 = \sum_{k=1}^6 F_k e_{kt}^0, \quad r_c^0 = P \cdot r_c,$$

where  $m$  is the mass of the loaded moving platform,  $g$  is the gravitational acceleration,  $r_c^0$  is the radius vector of the center of mass of the platform in the moving coordinate system,  $r^0$  is the acceleration of the point  $O$  and  $F_k$  are forces acting on the moving platform from the side of the extensible supports.

The equation of moments relative to the center of mass in the moving coordinate system has the form:

$$J_c \cdot \dot{\omega} + \omega \times (J_c \cdot \omega) = M = \sum_{k=1}^6 F_k (\alpha_k - r_c) e_{kt}, \quad e_{kt} = P^T \cdot e_{kt}^0, \quad \omega = P^T \cdot \omega^0,$$

where  $J_c$  is the inertia tensor of the static load relative to point  $C$ .

In the case under consideration, the UAV with four engine propeller units was adopted as a static load, and it can be represented as a ball with radius  $R_S$  and mass  $M_S$ , at a distance  $l$  from the center of which there are balls with mass  $M_M$  and radius  $R_M$ , fixed by means of a cylindrical rod, mass  $M_K$ . In this case, the inertia tensor of the static load can also be calculated as follows:

$$J_c = \begin{bmatrix} I_x & 0 & 0 \\ 0 & I_y & 0 \\ 0 & 0 & I_z \end{bmatrix},$$

$$I_x = I_y = \frac{2 \cdot M_S \cdot R_S^2}{5} + 4 \cdot \frac{\left(\frac{l}{\sqrt{2}}\right)^2 \cdot M_K}{12} + \left[ 4 \cdot \frac{2 \cdot M_M \cdot R_M^2}{5} + 4 \cdot M_M \left(\frac{l}{\sqrt{2}}\right)^2 \right],$$

$$I_z = \frac{2 \cdot M_S \cdot R_S^2}{5} + 4 \cdot \frac{l^2 \cdot M_K}{12} + \left[ 4 \cdot \frac{2 \cdot M_M \cdot R_M^2}{5} + 4 \cdot M_M \cdot l^2 \right].$$

where  $I_x$ ,  $I_y$  and  $I_z$  are the axial moments of inertia of the UAV, kg m<sup>2</sup>.

Consider the direct problem of dynamics—finding the forces developed by the drives of extensible supports, providing a given movement of the platform [17, 18]. The direct problem of dynamics is considered under the condition that the generalized coordinates  $q_i(t)$  are given as functions of time, as a result of which the quantities  $F^0$  and  $M$  become known. In this case, the system of equations of motion of the center of gravity  $C$  and the loaded moving platform in the fixed coordinate system can be written in the following form, while the matrix representation of the system of equations is similar to (4):

$$\sum_{k=1}^6 F_k e_{kt}^0 = F^0, \quad \sum_{k=1}^6 F_k (a_k^0 \times e_{kt}^0) = M^0 + r_c^0 \times F^0 = \hat{M}^0, \quad M^0 = P \cdot M,$$

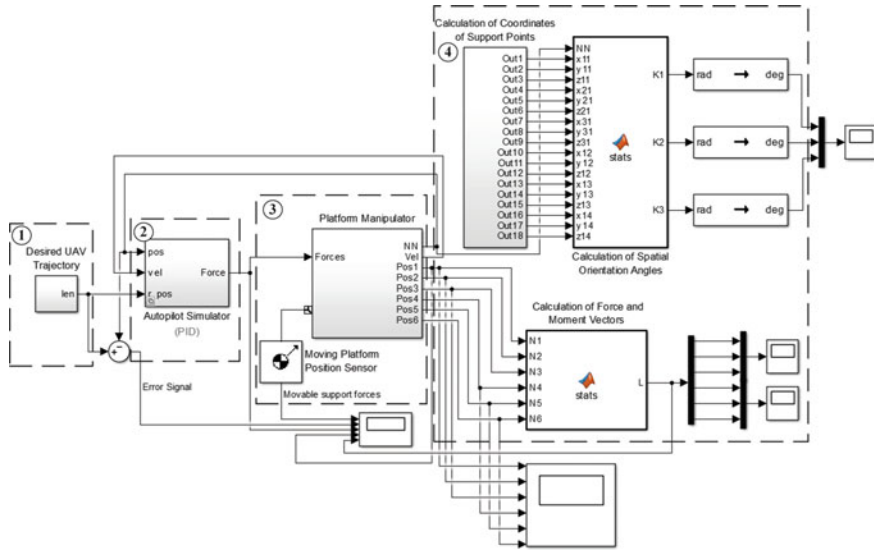
$$A^T \cdot F = \Upsilon^0, \quad F = (F_1, \dots, F_6)^T, \quad \Upsilon^0 = (F_x^0, F_y^0, F_z^0, \hat{M}_x^0, \hat{M}_y^0, \hat{M}_z^0)^T,$$

where  $r_c$  is the radius vector of the center of gravity of the platform in the moving coordinate system and  $F_k$  are the forces acting on the platform from the supports.  $F^0$  and  $M$  denote the main vector and the main moment of forces acting on the moving platform from the side of the supports of variable length.

### 3 Modeling the Measuring System

A functional model of the measuring systems was developed in the *MATLAB/Simulink* environment and represents the structure of macroblocks of the Simulink graphical environment modeling language (modules) with complete functionality. The model includes components as follows (Fig. 3).

- formation module of the desired UAV trajectory (1);
- simulator of an automatic control system (autopilot), which ensures stabilization of UAV at given angles of spatial orientation, presented in the form of combination three proportional–integral–derivative controllers [4, 19] (2);



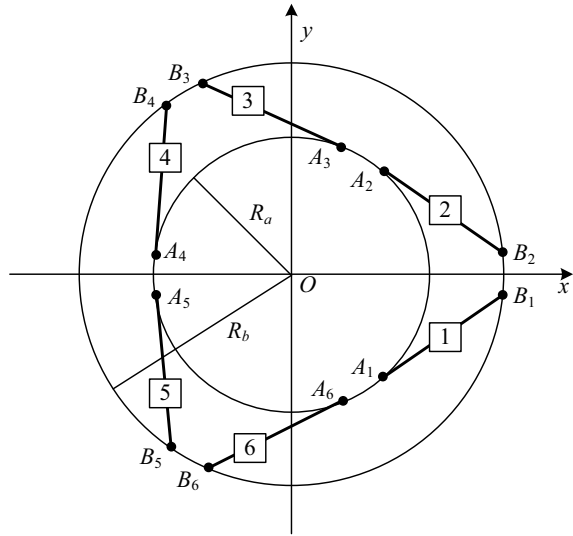
**Fig. 3** Structural organization of the UAV parameters measuring system functional model

- module that implements the mathematical model of kinematics and dynamics of the platform manipulator with six variable length supports and a set of drives and piezoelectric detectors (3);
- subsystem for calculating the components of the displacement vector of the moving platform of the manipulator and the spatial orientation angles of the platform (4).

The calculation of the mobile platform current position of the manipulator is carried out by two subsystems. The first subsystem allows determining the components of the displacement vector applied to the center of gravity of the moving platform and the moments of forces along the  $x$ ,  $y$  and  $z$  axes. The second subsystem is designed to calculate the values of the spatial orientation angles of the movable platform. The coordinate calculation procedure of the platform position includes the following steps:

1. Formation of a matrix of force vectors applied to common points of attachment for extensible supports pairs 1–2, 3–4, and 5–6 (Fig. 4).
2. Calculation of vectors and moments of forces along the  $x$ ,  $y$  and  $z$  axes applied to common points of attachment of support pairs 1–2, 3–4 and 5–6.
3. Formation of a matrix of force vectors applied to the center of gravity of the moving platform.
4. Calculation of the components of the displacement vector applied to the center of gravity of the moving platform and the moments of forces along the  $x$ ,  $y$  and  $z$  axes.

**Fig. 4** Arrangement of supports 1–6 of the platform manipulator in the projection onto the horizontal plane,  $R_a$  and  $R_b$  are radii of the moving platform and fixed base, respectively



The calculation procedure of the spatial orientation angles of the platform includes the following operations:

1. Coordinate transformation of attachment points of supports to the fixed base and the moving platform from a cylindrical coordinate system to Cartesian.
2. Compilation of a coordinate matrix of attachment points.
3. Calculation of the moving platform relative displacement respect to the fixed base for the common point of support pairs 1–2, 3–4 and 5–6 attachment.
4. Derived the equation of the moving platform plane by three attachment points of support pairs.
5. Calculation of a directional cosines matrix for the spatial orientation angles of the moving platform.
6. Calculation of the spatial orientation angles of the moving platform.

The nature of support lengths changing of the Gough–Stewart platform manipulator during the simulation process is determined by the UAV autopilot.

In the simulation, the following initial conditions were accepted the presented parameters correspond to the technical characteristics of the H-850 manipulator (manufacturer—Physik Instrumente, Germany) [20]:

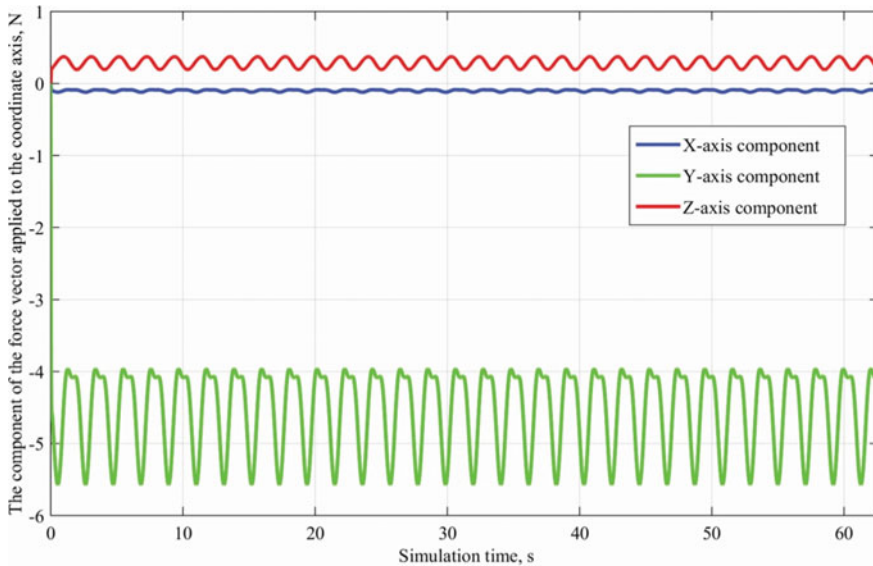
- angle of support inclination relative to the moving platform plane:  $20^\circ$ ;
- angle between the support attachment points to the moving platform and the coordinate axes:  $60^\circ$ ;
- diameter of the moving platform: 250 mm;
- height of the platform manipulator: 328 mm;
- distance between the application points of forces to piezoelectric detectors built into support pairs 1–2, 3–4 and 5–6: 180 mm.

The desired trajectory of the moving platform with a static load installed on it (UAV model) along the axes of the stationary coordinate system is adopted as follows:

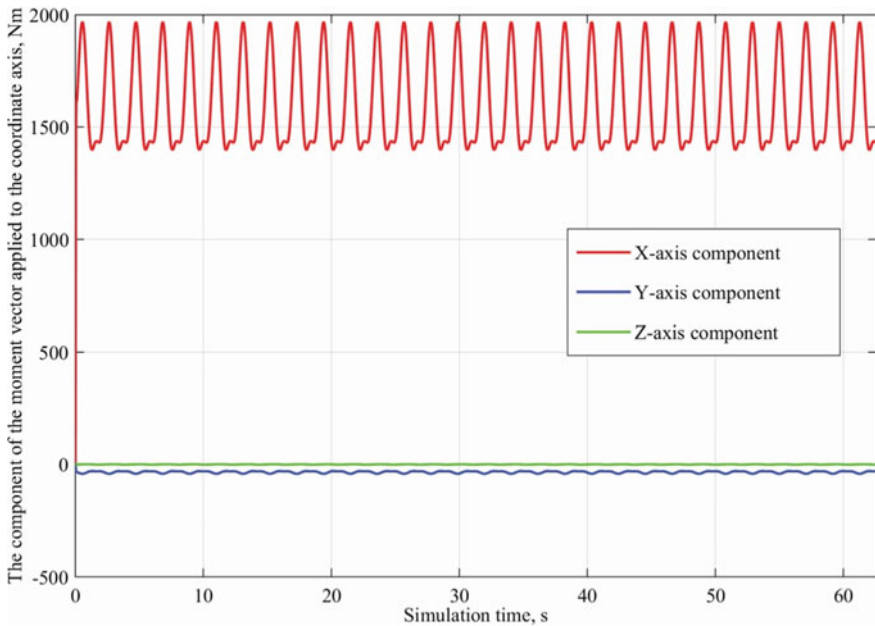
- $x, y$ -axis: displacement amplitude:  $\pm 50$  mm, cyclic frequency: 3 rad/s;
- $z$ -axis: displacement amplitude:  $\pm 25$  mm, cyclic frequency: 3 rad/s;
- spatial orientation angles:  $\pm 15^\circ$ , cyclic frequency: 3 rad/s.

During the functional modeling of the measuring system based on the platform manipulator, the obtained timing characteristics of the change in the values of the spatial orientation angles, support lengths and the forces applied to them were obtained. Figures 5 and 6 show examples of the obtained timing characteristics of changes in the values of the vector components applied to the center of mass of the moving platform. An analysis of the results of the model's functioning shows its sufficient accuracy for use in the process of designing new and research existing control algorithms for UAV or other autonomous vehicles that require testing using the hexapod platform manipulator.

In addition to performing basic tasks, the developed model allows us to evaluate the functioning parameters of PID controllers of spatial orientation angles, as well as to automatically receive program code for the implementation of both automation of parameter measuring and UAV propulsion control. Also, the presented functional model can function in conjunction with the user interface of the measuring system and be used to evaluate the parameters of its functioning.



**Fig. 5** Timing characteristics of the change in the values of the force vector components



**Fig. 6** Timing characteristics of the change in the values of the force moment vector components

## 4 Conclusion

For the rapid development of software and hardware for modeling unmanned aerial vehicle control systems, it is advisable to design a test measuring system (test stand) using the principles of model-driven design at early stages of work. The functional model of the test stand proposed in the paper performed in the *MATLAB/Simulink* environment is intended for the study of UAV control algorithms. The model includes the kinematics and dynamics of the hexapod platform manipulator, the computational modules of the vector of generalized coordinates for the moving platform and the UAV autopilot simulator. Changing the support lengths of the platform manipulator when simulating the movement process is determined by the UAV autopilot.

For the further development of the functional model, it is planned to detail the description of the kinematics and dynamics of the platform manipulator, in particular, taking into account the inertia and weight of the extensible supports, the forces and friction torques in the kinematic pairs and the influence of the payload. It is also planned to introduce algorithms for preliminary calculation of the UAV flight routes and the formation of the desired trajectory for the moving platform of the manipulator in accordance with a flight route.

**Acknowledgements** The work was supported by the Ministry of Education and Science of the Russian Federation (No. 02.G25.31.0313).



## References

1. Valavanis, K.P., Vachtsevanos, G.J. (eds.): Handbook of Unmanned Aerial Vehicles. Springer, Netherlands, Dordrecht (2015). <https://doi.org/10.1007/978-90-481-9707-1>
2. Budiyo, A., Riyanto, B., Joelianto, E. (eds.): Intelligent Unmanned Systems: Theory and Applications. Springer, Berlin (2009). <https://doi.org/10.1007/978-3-642-00264-9>
3. Paw, Y.C., Balas, G.J.: Development and application of an integrated framework for small UAV flight control development. *Mechatronics* **21**, 789–802 (2011). <https://doi.org/10.1016/j.mechatronics.2010.09.009>
4. Liu, M., Egan, G.K., Santoso, F.: Modeling, autopilot design, and field tuning of a UAV with minimum control surfaces. *IEEE Trans. Control Syst. Technol.* **23**, 2353–2360 (2015). <https://doi.org/10.1109/TCST.2015.2398316>
5. Stewart, D.: A platform with six degrees of freedom. *Proc. Inst. Mech. Eng.* **180**, 371–386 (1965). [https://doi.org/10.1243/PIME\\_PROC\\_1965\\_180\\_029\\_02](https://doi.org/10.1243/PIME_PROC_1965_180_029_02)
6. Jang, T.K., Lim, B.S., Kim, M.K.: The canonical Stewart platform as a six DOF pose sensor for automotive applications. *J. Mech. Sci. Technol.* **32**, 5553–5561 (2018). <https://doi.org/10.1007/s12206-018-1101-0>
7. Furqan, M., Suhaib, M., Ahmad, N.: Studies on Stewart platform manipulator: a review. *J. Mech. Sci. Technol.* **31**, 4459–4470 (2017). <https://doi.org/10.1007/s12206-017-0846-1>
8. Drozd, O.V., Russkikh, P.A., Chentsov, S.V., Kapulin, D.V.: Integration of hardware and software tools for VLSI SoC class design support. *IOP Conf. Ser. Mater. Sci. Eng.* **450**, 1–6 (2018). <https://doi.org/10.1088/1757-899X/450/5/052011>
9. Drozd, O.V., Russkikh, P.A., Kapulin, D.V.: Special software for automated measuring complex based on Rohde & Schwarz equipment. In: 2019 International Siberian Conference on Control and Communications (SIBCON), pp. 1–4. IEEE, Tomsk (2019). <https://doi.org/10.1109/SIBCON.2019.8729644>
10. Köhler, C.: Enhancing Embedded Systems Simulation. Vieweg+Teubner, Wiesbaden (2011). <https://doi.org/10.1007/978-3-8348-9916-3>
11. Lebret, G., Liu, K., Lewis, F.L.: Dynamic analysis and control of a Stewart platform manipulator. *J. Robot. Syst.* **10**, 629–655 (1993). <https://doi.org/10.1002/rob.4620100506>
12. Liu, M.-J., Li, C.-X., Li, C.-N.: Dynamics analysis of the Gough-Stewart platform manipulator. *IEEE Trans. Robot. Autom.* **16**, 94–98 (2000). <https://doi.org/10.1109/70.833196>
13. Abdellatif, H., Heimann, B.: Computational efficient inverse dynamics of 6-DOF fully parallel manipulators by using the Lagrangian formalism. *Mech. Mach. Theory* **44**, 192–207 (2009). <https://doi.org/10.1016/j.mechmachtheory.2008.02.003>
14. Dasgupta, B., Mruthyunjaya, T.S.: A Newton-Euler formulation for the inverse dynamics of the Stewart platform manipulator. *Mech. Mach. Theory* **33**, 1135–1152 (1998). [https://doi.org/10.1016/S0094-114X\(97\)00118-3](https://doi.org/10.1016/S0094-114X(97)00118-3)
15. Gallardo, J., Alcaraz, L.A.: Kinematics of the Gough-Stewart platform by means of the Newton-homotopy method. *IEEE Lat. Am. Trans.* **16**, 2850–2856 (2018). <https://doi.org/10.1109/TLA.2018.8804248>
16. Ruiz-Tolosa, J.R., Castillo, E.: From Vectors to Tensors. Springer, Berlin (2005). <https://doi.org/10.1007/b138560>
17. Schulz, S., Seibel, A., Schreiber, D., Schlattmann, J.: Sensor concept for solving the direct kinematics problem of the Stewart-Gough platform. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1959–1964. IEEE, Vancouver (2017). <https://doi.org/10.1109/IROS.2017.8206015>
18. Zhu, Q., Zhang, Z.: An efficient numerical method for forward kinematics of parallel robots. *IEEE Access* **7**, 128758–128766 (2019). <https://doi.org/10.1109/ACCESS.2019.2940064>

19. Mardiyanto, R., Hidayat, R., Aprilian, E., Suryoatmojo, H.: Development of autopilot system of unmanned aerial vehicle for aerial mapping application. In: 2018 International Seminar on Intelligent Technology and Its Applications (ISITIA), pp. 357–361. IEEE, Bali (2018). <https://doi.org/10.1109/ISITIA.2018.8710966>
20. H-850 6-Axis Hexapod For Loads up to 250 kg. <https://www.pi-usa.us/en/products/6-axis-hexapods-parallel-positioners/h-850-6-axis-hexapod-700800/>. Last accessed 2019/10/29

# Research and Evaluation of the Most Significant Quantitative Characteristics of MPLS Equipment



Andrey Krasov , Pavel Karelsky , Igor Zuyev , Max Kovzur ,  
and Aleksander Tasyuk 

**Abstract** The article presents an analysis of the characteristics of operator's equipment for the construction of MPLS networks, which will be important in assessing the proposed procurement units for network modernization. The main characteristics of the devices are considered in the work and the most significant of them are analyzed. Method: An analytical dependence is proposed for the characteristics of L2 VPN services. Core results: A practical experiment was carried out to confirm theoretical conclusions on such characteristic as the number of the LDP neighbors. Practical relevance: The results of the experiment are presented and conclusions are made about the significant characteristics for MPLS equipment for L2 VPN services.

**Keywords** MPLS · LDP · Multi-protocol label switching · Experiment

## 1 Introduction

The requirements for a modern backbone network are: high transmission speed, high bandwidth, good scalability, and reliability.

But the current situation in the telecommunications services market places higher demands on service providers. Now it is not enough for the provider to give an easy access to their backbone network, as recently clients have preferred connecting to the provider's Virtual Private Networks.

One of the steps toward the modernization of a multiservice network is the use of multilevel switching methods, which allows to logically structure the network without sacrificing its performance.

Currently, the switching mechanism for Internet trunks is the technology of multi-protocol label switching (MPLS). The progenitors of MPLS were ATM and FrameRelay technologies. Their important feature was the determination of the path in advance. So the whole direction of the package was predictable and manageable.

---

A. Krasov · P. Karelsky (✉) · I. Zuyev · M. Kovzur · A. Tasyuk  
The Bonch-Bruевич St. Petersburg State University of Telecommunications, St. Petersburg,  
Russia  
e-mail: [pasha.karelsky@yandex.ru](mailto:pasha.karelsky@yandex.ru)

However, the lack of these technologies was inertness—their reaction to rebuilding the network was really poor.

The IP-MPLS network is flexible and scalable as it is built on top of the IP protocol. It provides a predetermined path, as in ATM. But at the same time it allows to quickly respond to rebuilding the network. This allows (under certain conditions) to provide the required channel parameters, in particular bandwidth, delay, jitter, etc.

Pure MPLS is not used in modern networks. Nowadays it is used in combination with MPLS-IP or MPLS-TP. This means that the network itself is built on top of IP and Ethernet or in a normal video signal, but at the same time it can transfer data from many other protocols.

All of these is ideal for organizing a transparent traffic transfer service between local networks of companies in different cities with guaranteed quality, speed and other SLA parameters.

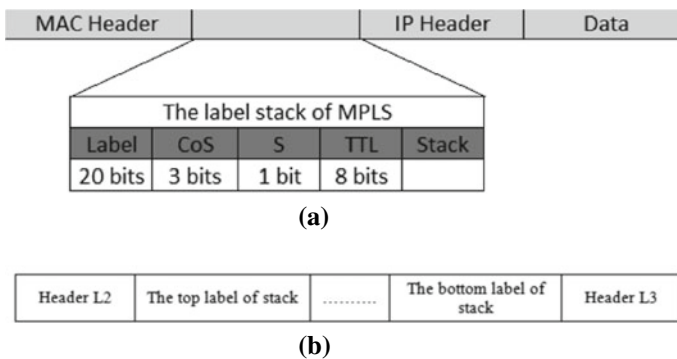
MPLS provides the ability to procuring a QoS value which guarantees higher security. Moreover, for the same set of nodes, can create several different virtual networks (using different labels), for example, for different types of QoS.

To provide structured flows, a label stack is created in the package (see Fig. 1a,b). In a normal video signal, it is placed between the headers of the network and channel levels (L2 and L3, respectively). Each entry on the stack takes 4 octets.

The Label field is the label itself. Its length is 20 bits. The CoS field corresponds to a subfield of the ToS field priority. It has three bits, which is enough for the IP header priority field or Ethernet CoS priority field. Sometimes this field is also called TC—TrafficClass, which carries the priority of the packet, like the DSCP field in IP.

Field S (Bottom of Stack)—the indicator of the bottom of the label stack, which is 1 bit long. There may be several MPLS headers on a packet, for example, external—for switching in an MPLS network, and internal—indicates belonging to a particular VPN.

TTL field (TimeToLive)—similar to the field in the IP TTL header. It has a length of 8 bits. The main task is to prevent the packet from constantly being in the network in the event loop.



**Fig. 1 a** The label stack format. **b** Placing tags on the stack

All in all, MPLS is a technology based on the label transfer method. Labels define both routes and service attributes. During the development of the operator, new customers and other operators appear who need to organize a service of transparent traffic transfer between cities with guaranteed quality, speed and other parameters.

In order to solve this problem among service providers, MPLS-IP technology has been widely used, which serves building L2/L3 VPN tunnels. It corresponds to the required functionality, performance, flexibility, fault tolerance. It also provides scalability, security, and quality of network service, as well as the most efficient use of network resources.

Before introducing certain services into an existing network, it is necessary to carefully study and evaluate this technology.

In case the equipment of the service provider does not support the MPLS technology, there is a need to update the fleet of devices. However, when updating the equipment, several questions arise:

- What should be looked at while choosing the equipment?
- What are the most significant characteristics?

Determination of the equipment’s most significant characteristics is an important task and its importance is showed in this article.

## 2 Materials and Methods

An analysis of the solutions presented on the market showed that mainly manufacturers pay attention to price and description of characteristics, rather than quantitative parameters of equipment. In the course of the work, Table 1 was compiled with the characteristics of the equipment of several manufacturers [1–3].

**Table 1** An example of the characteristics of the equipment of various vendors

Characteristic	JuniperEX4550	RaisecomiTN8800	Juniper MX-104	Juniper QFX5100	Cisco ASR 9K
The number of MAC addresses	32K	32K	512K	288K	512 K
Fib IPv4 table capacity	14K	16K	4M	100K	128 K
RIB table capacity	10K	–	21M	–	–
Max pseudo wires	–	4K	16K	–	4 K
VRF	254	1K	2K	1K	4K
MPLS labels	125	–	–	16K	16K

**Table 2** Interconnection of equipment and services characteristics

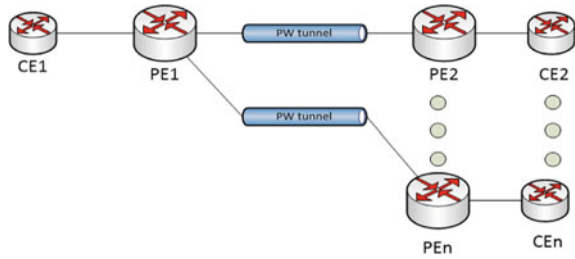
L2VPN service	L3VPN service	Common to services
The number of LDP neighbors	The number of FIB/RIB	The number of MAC Addresses
The number of VSI	The number of sessions BGP	The number of LSP
The number of PW	The number of IGP neighbors	MPLS labels
	The number of VRF	

It is worth noting that the prices for the equipment presented above may differ by several orders of magnitude. Therefore, it is necessary to understand the resources of which equipment will be sufficient to achieve the objectives. As a result of studying the mechanisms of operation of MPLS technology, as well as L2/L3 VPN services based on it, the following, most important, equipment characteristics were highlighted, presented in Table 2 [4].

Let's explain the chosen characteristics:

- The number of MAC addresses supported by the device. Media Access Control is the physical address of a device or interface on Ethernet networks. It allows to determine how many MAC addresses a single device can remember.
- The number of entries in the Forwarding Information Base (FIB). This table affects packet forwarding.
- The number of entries in the routing table Routing Information Base (RIB)—affects the number of routes.
- Number of bidirectional static Label Switch Path (LSPs). This is a bidirectional path through the MPLS network.
- Number of Label Distribution Protocol (LDP) LSP. Defines the number of virtual paths/channels for the LDP protocol.
- The number of supported Pseudo Wire (PWs).
- The number of Virtual Routing and Forwarding instance (VRF) supported on the device. In fact, this is the number of supported L3 VPNs on the device.
- The number of Virtual Switching Instance (VSI). It is a virtual switch within a single node and affects the number of VPLS L2 VPN services terminated on equipment.
- The number of Interior Gateway Protocol (IGP) neighbors, which affects the exchange of routing information between shared routers on a network.
- The number of LDP sessions—affects the number of devices on which L2 VPN services will be terminated.
- The number of BGP sessions. This parameter affects network scalability in terms of L3 VPN services.

**Fig. 2** Point-to-point topology



To modernize the network, it is necessary to choose the most optimal equipment, with sufficient resources to meet the operator’s needs and within the established budget [5, 6].

One of the methods for determining the most significant characteristics is calculation of network scalability for L2 VPN services [7].

The first calculation will be derived from the point-to-point topology (see Fig. 2).

Let’s introduce the assumption that the chosen equipment for the network will have the following characteristics:

- $M = 30,900$ —the number of maximum supported MAC addresses;
- $MLN = 32$ —the number of maximum supported LDP neighbors;
- $MPL = 8000$ —the number of maximum supported MPLS tags.

The calculation was made taking into consideration the following characteristics:

- $D = 25, 100$ —the number of devices on the operator’s network;
- $NoC = 1... 1000$ —the number of clients connected to one device;
- $MpC = 30$ —the number of MAC addresses per client.

It is necessary to evaluate the percentage of equipment resource usage based on the technical characteristics of MPLS devices, the number of possible clients, as well as client parameters.

Let’s determine the consumption of MAC addresses per number of clients— $MpC$ . This characteristic shows how many MAC addresses will fall on one client with a uniform distribution of addresses. We use the ratio:

$$MpC = \frac{M}{NoC} \tag{1}$$

where  $M$  is the number of MAC addresses supported by the equipment,  $NoC$  is the number of connected clients. Figure 3 shows that with an increase in the number of clients, a smaller number of addresses appear per client with an equal distribution of addresses.

Next, we will evaluate such characteristics as the percentage of use of MAC addresses (MAC Usage), LDP neighbors (LDPN Usage), labels (Label Usage) for a different number of devices.

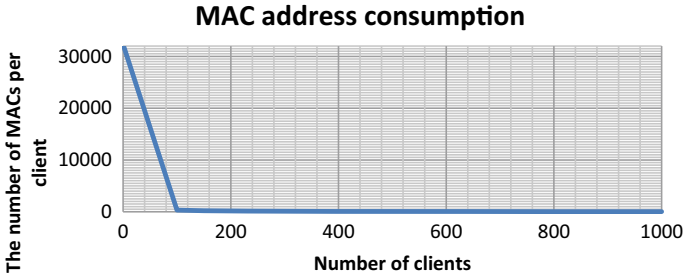


Fig. 3 MAC address consumption

$$\text{MAC Usage} = \frac{\text{MpC} * \text{NoC} + D * \text{MpD}}{M} \tag{2}$$

$$\text{LDPN Usage} = \begin{cases} \frac{D}{\text{MLN}}, & \text{at } \text{MLN} < \text{NoC} \text{ and } D \leq \text{MLN} \\ 1, & \text{at } \text{MLN} < \text{NoC} \text{ and } \text{MLN} < D \leq \text{NoC} \\ 1, & \text{at } \text{MLN} < \text{NoC} \text{ and } \text{NoC} < D \\ \frac{D}{\text{MLN}}, & \text{at } \text{NoC} \leq \text{MLN} \text{ and } D \leq \text{NoC} \\ \frac{\text{NoC}}{\text{MLN}}, & \text{at } \text{NoC} \leq \text{MLN} \text{ and } \text{NoC} < D \leq \text{MLN} \\ \frac{\text{NoC}}{\text{MLN}}, & \text{at } \text{NoC} \leq \text{MLN} \text{ and } \text{MLN} < D \end{cases} \tag{3}$$

$$\text{Label Usage} = \frac{\text{NoC} + D}{\text{MPL}} \tag{4}$$

where MpC is the number of MAC addresses per client, NoC is the number of clients connected to one device, *D* is the number of devices on the operator’s network, MpD is the number of MAC addresses per device, *M* is the number of MAC addresses, MLN is the number of maximum supported LDP neighbors, MPL—the number of maximum supported MPLS labels. Formulas 2, 3, and 4 are presented as graphical dependencies (see Figs. 4 and 5) [8].

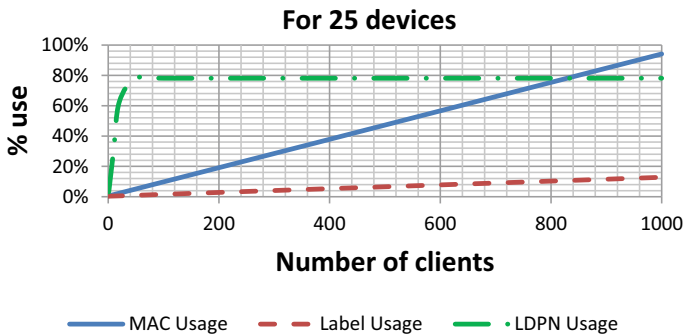


Fig. 4 Performance graphs for 25 devices



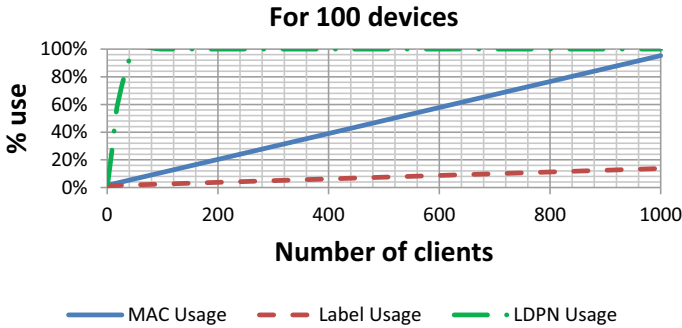


Fig. 5 Performance graphs for 100 devices

The second method for determining the most significant characteristics is estimation of equipment resource reserves.

When choosing equipment, it is necessary to take into account the quantitative resources that will be required from the equipment to perform the required task. ‘Equipment resources’ refer to the quantitative characteristics which are necessary to provide certain services. Since the possibilities of ASIC and CPU are not unlimited, it is advisable to talk about such concept as maximum equipment resources [9]. The definition of such term can be presented as quantitative limited characteristics, which can be both a limitation of the vendor and a limitation of the specialized chip or processor used by equipment [10, 11].

To accurately understand the capabilities of the equipment after a certain setup, need to know the so-called resource reserve, which is understood as the ratio of unused resources to the maximum equipment resources, expressed as a percentage [12, 13].

Initial data for assessing the stock of equipment resources are following:

- Network and customer characteristics:
  - 100 nodes in the network (D); 2000 clients (20 clients per device), (NoC);
  - 100 MAC addresses per client (MpC);
  - 10,000 MAC for multicast and other services (MpD).
- Equipment specifications:
  - 32,000 MAC (M);
  - 32 LDP neighbors (MLN);
  - 1000 VRF; 8000 labels (MPL);
  - 32 single-hop + 32 multi-hop sessions BGP.

Calculation for L2 VPN service.

- MAC usage:

**Table 3** Final calculations for L2 VPN service

Label usage	LDPN usage	MAC usage
26.3%	62.5%	37.5%

$$\text{MAC Usage} = \frac{\text{MpC} * \text{NoC} + D * \text{MpD}}{M} = \frac{100 * 2000 + 100 * 100 * 100}{32,000} = 37.5\% \quad (5)$$

- Label usage:

$$\text{Label Usage} = \frac{\text{NoC} + D}{\text{MPL}} * 100 = \frac{2000 + 100}{8000} * 100 = 26.3\% \quad (6)$$

- Using LDP Neighbors:

$$\text{LDPN Usage} = \frac{20 \text{ clients per device}}{32\text{MLN}} * 100 = 62.5\% \quad (7)$$

### 3 Results

After carrying out calculations of network scalability for L2 VPN, the following conclusions can be made.

Figures 4 and 5 show that with an increase in the number of devices, the use of characteristics increases. As a result, for the chosen topology and equipment the most significant characteristic will be the number of LDP neighbors. If the number of clients is more than the MLN parameter, connected to one MPLS device and terminated in other mismatched devices, it is required to use the H-VPLS functionality [7, 14, 15].

The results of the assessment of equipment resource reserves are presented in Table 3.

The table shows that the LDPN parameter plays the most decisive role.

To confirm the calculations, it is necessary to conduct an experiment.

### 4 Discussion

To verify the results of the calculations, an experiment was conducted. The objective of the experiment is to configure the required number of PW tunnels in order to show the consumption of LDPN. A network topology was assembled, consisting of ten routers connected by a ring (see Fig. 6) [16, 17].

An experiment was conducted for L2 VPN to characterize LDPN (number of LDP neighbors). First of all, the basic configuration of the routers was provided:

- adding addresses on interfaces;

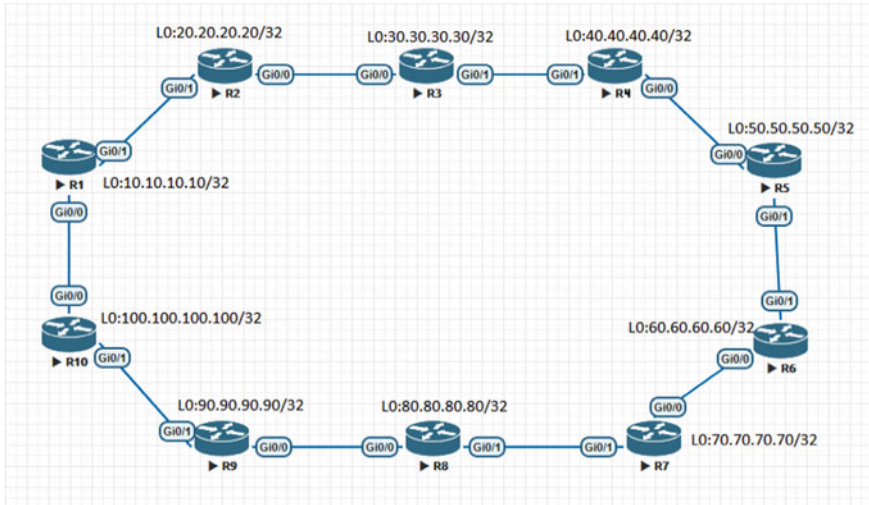


Fig. 6 Network topology

- OSPF Dynamic Routing Protocol configuration;
- loopback addresses creation.

The next step was the configuration of PW tunnels from R1 to each next router. After the configuration was completed, the number of LDP neighbors for R1 can be displayed using the “show mpls ldp neighbor” command (see Fig. 7).

According to results, in addition to routers directly connected to R1, there are also other LDP neighbors. Tunnels from R1 to R5 were also configured to increase the number of possible clients (see Fig. 8).

After this step, let’s check the number of LDP neighbors on R1 (see Fig. 9).

According to Fig. 9, the number of neighbors has not been changed. In order to visually show the dependence, the graph of the usage of LDP neighbors for a different number of clients was built.

Such dependence presented in Table 4 can be described due to the fact that the first few clients are evenly distributed between devices on the network. And subsequent clients are already connected to existing devices. As a result of the experiment, a graph was obtained (see Fig. 10).

Based on the obtained graph, we can conclude that the calculations that were performed earlier are confirmed by the results of the experiment.

## 5 Conclusion

Of the considered characteristics, the most significant for the L2 VPN service is the number of LDP neighbors. As a result, the calculations carried out earlier were

**Fig. 7** LDP neighbor information before configuring additional tunnels

```

Peer LDP Ident: 20.20.20.20:0; Local LDP Ident 10.10.10.10:0
TCP connection: 20.20.20.20.60293 - 10.10.10.10.646
State: Oper; Msgs sent/rcvd: 77/76; Downstream
Up time: 00:46:47
LDP discovery sources:
  GigabitEthernet0/1, Src IP addr: 2.2.2.2
Addresses bound to peer LDP Ident:
  3.3.3.2  20.20.20.20  2.2.2.2
Peer LDP Ident: 30.30.30.30:0; Local LDP Ident 10.10.10.10:0
TCP connection: 30.30.30.30.11722 - 10.10.10.10.646
State: Oper; Msgs sent/rcvd: 59/59; Downstream
Up time: 00:30:21
LDP discovery sources:
  Targeted Hello 10.10.10.10 -> 30.30.30.30, active, passive
Addresses bound to peer LDP Ident:
  3.3.3.1  30.30.30.30  4.4.4.1
Peer LDP Ident: 40.40.40.40:0; Local LDP Ident 10.10.10.10:0
TCP connection: 40.40.40.40.24494 - 10.10.10.10.646
State: Oper; Msgs sent/rcvd: 57/54; Downstream
Up time: 00:26:27
LDP discovery sources:
  Targeted Hello 10.10.10.10 -> 40.40.40.40, active, passive
Addresses bound to peer LDP Ident:
  40.40.40.40  4.4.4.2  5.5.5.1
Peer LDP Ident: 50.50.50.50:0; Local LDP Ident 10.10.10.10:0
TCP connection: 50.50.50.50.28478 - 10.10.10.10.646
State: Oper; Msgs sent/rcvd: 41/40; Downstream
Up time: 00:14:25
LDP discovery sources:
  Targeted Hello 10.10.10.10 -> 50.50.50.50, active, passive
Addresses bound to peer LDP Ident:
  50.50.50.50  5.5.5.2  6.6.6.1

```

```

R1#show mpls l2transport summary
Destination address: 30.30.30.30, total number of vc: 1
  0 unknown, 1 up, 0 down, 0 admin down, 0 recovering, 0 standby, 0 hotstandby
  1 active vc on MPLS interface Gi0/1
Destination address: 40.40.40.40, total number of vc: 1
  0 unknown, 1 up, 0 down, 0 admin down, 0 recovering, 0 standby, 0 hotstandby
  1 active vc on MPLS interface Gi0/1
Destination address: 50.50.50.50, total number of vc: 10
  0 unknown, 10 up, 0 down, 0 admin down, 0 recovering, 0 standby, 0 hotstandby
  10 active vc on MPLS interface Gi0/1

```

**Fig. 8** Tunnel configuration information

confirmed by the experimental part. It follows that the obtained formulas carry practical value in evaluating equipment parameters, in particular in determining the necessary values of such characteristic as LDPN.

In terms of network security, MPLS L2/L3 VPN technologies offer a new level of protection for network traffic. Despite the fact that packets are transmitted on the same core network, the traffic of each client is isolated inside the VPN [18, 19].

```

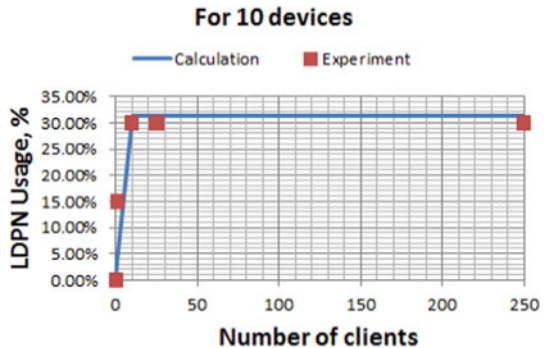
R1#show mpls ldp neighbor
Peer LDP Ident: 20.20.20.20:0; Local LDP Ident 10.10.10.10:0
TCP connection: 20.20.20.20.40088 - 10.10.10.10.646
State: Oper; Msgs sent/rcvd: 74/74; Downstream
Up time: 00:45:09
LDP discovery sources:
  GigabitEthernet0/1, Src IP addr: 2.2.2.2
Addresses bound to peer LDP Ident:
  3.3.3.2      20.20.20.20      2.2.2.2
Peer LDP Ident: 30.30.30.30:0; Local LDP Ident 10.10.10.10:0
TCP connection: 30.30.30.30.21460 - 10.10.10.10.646
State: Oper; Msgs sent/rcvd: 78/74; Downstream
Up time: 00:44:34
LDP discovery sources:
  Targeted Hello 10.10.10.10 -> 30.30.30.30, active, passive
Addresses bound to peer LDP Ident:
  3.3.3.1      30.30.30.30      4.4.4.1
Peer LDP Ident: 40.40.40.40:0; Local LDP Ident 10.10.10.10:0
TCP connection: 40.40.40.40.31695 - 10.10.10.10.646
State: Oper; Msgs sent/rcvd: 75/77; Downstream
Up time: 00:44:24
LDP discovery sources:
  Targeted Hello 10.10.10.10 -> 40.40.40.40, active, passive
Addresses bound to peer LDP Ident:
  5.5.5.1      40.40.40.40      4.4.4.2
Peer LDP Ident: 50.50.50.50:0; Local LDP Ident 10.10.10.10:0
TCP connection: 50.50.50.50.23476 - 10.10.10.10.646
State: Oper; Msgs sent/rcvd: 86/85; Downstream
Up time: 00:44:23
LDP discovery sources:
  Targeted Hello 10.10.10.10 -> 50.50.50.50, active, passive
Addresses bound to peer LDP Ident:
  5.5.5.2      50.50.50.50      6.6.6.1
    
```

Fig. 9 LDP neighbor information after configuring tunnels

Table 4 Results of experiment

LDPN, %	NoC, qty
0	0
15	10
30	20
30	250

Fig. 10 Comparison of theoretical evaluation and experimental results



Security mechanisms used on the operator's network should prevent denial of service attacks and unauthorized access to the network and should also include the ability to protect both the control plane and the data plane.

Service providers and enterprises can use these capabilities to implement reliable and secure MPLS networks, maximize network reliability and minimize the potentially adverse effects of a network attack [20].

## References

1. IP-MPLS PE and Pre-Aggregation System. <https://www.raisecom.com/>
2. MPLS: Layer 2 VPNs, Configuration Guide. <https://cisco.com/>
3. TechLibrary. <https://www.juniper.net/>
4. Karel'skij, P.V., Kovcur, M.M., Rjazancev, K.S.: Development of the Project for the Upgrading of the Provider's Network with the Implementation of MPLS Based Services. Aktual'nye problemy infotelekkommunikacij v nauke i obrazovanii (APINO 2018). VII Mezhdunarodnaja nauchno-tehnicheskaja i nauchno-metodicheskaja konferencija: sbornik nauchnyh statej, vol. 1, pp. 446–450 (2018) (in Russian)
5. Kovcur, M.M.: Issledovanie neperesekajushhijsja marshrutov global'noj seti. Nauka vchera, segodnja, zavtra. №6 (6). sbornik statej po materialam VI mezhdunarodnoj nauchno-prakticheskoy konferencii. Izd. "SibAK", Novosibirsk, pp. 19–24 (2013) (in Russian)
6. Alejnikov, A.A., Biljatinov, K.Z., Krasov, A.V., Levin, M.V.: Kontrol', izmerenie i intellektual'noe upravlenie trafikom: monografija, Sankt-Peterburg (2016) (in Russian)
7. Service Provider Security. <https://www.cisco.com/c/en/us/about/security-center/service-provider-infrastructure-security.html>
8. Karel'skij, P.V., Kovcur, M.M., Bondarenko, A.I.: Ocenka kolichestvennyh charakteristik oborudovanijsja MPLS, igradjushhih reshajushhuju rol' pri masshtabirovanii seti operatora. Regional'naja informatika 2018. SPb. Sbornik trudov RIIB, vol. 5, pp. 146–151 (2018) (in Russian)
9. Deshevyyh, E.A., Konjuhov, V.M., Krylov, K.Ju., Ushakov, I.A.: Studying Methods of Protection Against Insider Attacks. V sbornike: Aktual'nye problemy infotelekkommunikacij v nauke i obrazovanii IV Mezhdunarodnaja nauchno-tehnicheskaja i nauchno-metodicheskaja konferencija: sbornik nauchnyh statej, vol. 2, pp. 310–313 (2015) (in Russian)
10. Toktogonov Samat Almazbekovich, Abdybek kyzy Ajkerim – Mnogoprotokol'naja kommutacija po metkam kak sposob rasshirenija spektra uslug u operatorov svjazi Vestnik Kyrgyzsko-Rossijskogo slavjanskogo universiteta, №1 (2016) (in Russian)
11. Shelkovyj, D.V., Mironov, O.Ju., Basov, O.O.: Modelirovanie potokov dannyh real'nogo vremeni v zashhishhennyh korporativnyh mul'tiservisnyh setjah svjazi na osnove determinirovannogo setevogo ischislenija Nauchnye vedomosti Belgorodskogo gosudarstvennogo universiteta. Serija: Jekonomika. Informatika, №3 (2018) (in Russian)
12. Kundimana, Zh., Dzhalalov, I.K.: Modernizacija IP/MPLS setej. V sbornike: Tehnologii informacionnogo obshhestva Materialy XIII Mezhdunarodnoj otraslevoj nauchno-tehnicheskoy konferencii, pp. 52–53 (2019) (in Russian)
13. Vasil'ev, D.S.: Osobennosti organizacii i ispol'zovanija MPLS VPN. V sbornike: Fundamental'nye nauchnye issledovanija: teoreticheskie i prakticheskie aspekty sbornik materialov X Mezhdunarodnoj nauchno-prakticheskoy konferencii. Kemerovo, pp. 10–13 (2019) (in Russian)
14. Zhil'cov, V.A., Jaremko, I.N.: Study of tunneling and optimization of traffic in the networks technology MPLS TE. Vestnik Doneckogo nacional'nogo universiteta. Serija G: Tehnicheskie nauki, № 2, pp. 16–24 (2019) (in Russian)
15. Krasov, A.V., Levin, M.V., Shterenberg, S.I., Isachenkov, P.A.: Model' upravlenija potokami trafika v programmno-opredeljaemoj seti s izmenjajushhejsja nagruzkoj. Naukoemkie tehnologii v kosmicheskikh issledovanijah Zemli, vol. 8, № 4, pp. 70–74 (2016) (in Russian)

16. Krasov, A.V., Levin, M.V., Cvetkov, A.Ju.: Upravlenie setjami peredachi dannyh s izmenjajushhejsja nagruzkoj. Vserossijskaja nauchnaja konferencija po problemam upravlenija v tehničkih sistemah, № 1, pp. 141–146 (2015) (in Russian)
17. Karaeva, A.M.S.: Postroenie virtual'noj chastnoj seti na baze MPLS. V sbornike: Tradicii i innovacii v sisteme obrazovanija Lepshokova S.M. (otvetstvennyj redaktor), pp. 141–144 (2017) (in Russian)
18. Doniev, Je.T., Nigmatov, Z.Z.: Osobnosti MPLS dlja upravlenija trafikom v IP-setjah. Molodoj uchenyj, № 35 (169), pp. 1–3 (2017) (in Russian)
19. Danilov, A.N., Derevjagin, P.B., Korolev, O.A.: Ispol'zovanie mehanizma prioritetov dlja shem s razdeljaemoj zashhitoy v seti IP/MPLS. V sbornike: Telekommunikacionnye i vychislitel'nye sistemy—2017 Trudy mezhdunarodnoj nauchno-tehničkoj konferencii, pp. 92–93 (2017) (in Russian)
20. Kartashevskij, V.G., Buranova, M.A.: Research efficiency tunneling in the MPLS network. V sbornike: II Nauchnyj forum telekommunikacii: teorija i tehnologii TTT-2017. problemy tehniki i tehnologij telekommunikacij PTITT-2017 materialy XVIII Mezhdunarodnoj nauchno-tehničkoj konferencii, pp. 155–157 (2017) (in Russian)

# Study of the Microstrip Waveguide Prototype Model for Use as a Retunable Diffraction Grating



Dmitry Nikulin , Valery Reichert , Sergey Shergin , Igor Karmanov , Vladimir Korneyev , and Polina Zvyagintseva 

**Abstract** The prototype model (hereinafter—the prototype) with strip micromechanical waveguides in a free state with piezoelectric oscillation excitation was studied. The study methodology for the basic functional characteristics of the prototype with strip micromechanical waveguides in a free state was developed. The test bench was made. Visual inspection of the prototype model was performed and its quality was assessed using microscope. Tests on the prototype model were carried out. Diffraction patterns of light reflected from a flexural wave in waveguides at different frequencies of mechanical excitation were obtained. Diffraction patterns were analyzed.

**Keywords** Strip waveguide · Elastic acoustic flexural waves · Acousto-optics · Electrically-controlled diffraction gratings

## 1 Introduction

One of the units in optical spectral systems is retunable diffraction gratings operating on the principle of interaction of light waves with elastic or flexural waves in thin-film membranes or micromechanical waveguides [1]. Elastic surface wavelength in these units is from ten to several hundred  $\mu\text{m}$  which allows making of retunable gratings within the respective period range (wavelengths).

Thin-film metal structures in free state have been widely used for a long time. They are used as: optical filters [2], conductive reflective layers of fluorescent screens [3], a vacuum-tight window for transmitting soft X-rays [4], a screen mask for vacuum spraying [5, 6], etc. The possibility of using thin-film structures in free state as diffraction optical elements, for example, strip microwave diodes of a tunable diffraction grating [7], working both on reflection and on transmission, is promising. Research is also known in the field of microoptoelectromechanical (MOEMS) deflectors and

---

D. Nikulin (✉) · V. Reichert · S. Shergin · I. Karmanov · V. Korneyev · P. Zvyagintseva  
Siberian State University of Geosystems and Technologies, Novosibirsk, Russia  
e-mail: [polinasugit@mail.ru](mailto:polinasugit@mail.ru)



**Fig. 1** Prototype model with micromechanical strip waveguides in a free state



light modulators using torsional oscillations in film membrane waveguides [6–8] and bending waves in thin-film strip membrane waveguides [9–19].

In order to achieve optimal performance of micromechanical strip waveguides based on elastic wave acousto-optics principles, a set of measures is required including a study of a prototype model of these units [1].

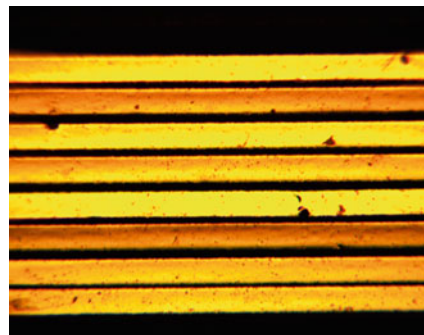
This article includes experimental results of performance characteristics study of the prototype with micromechanical strip waveguides in a free state as shown in Fig. 1.

## 2 Experiments

In the prototype under study, waveguides are polyimide based with metal coating. Metal coating thickness is  $100 \text{ nm}$  and total thickness of the waveguide is  $1.9 \text{ }\mu\text{m}$ .

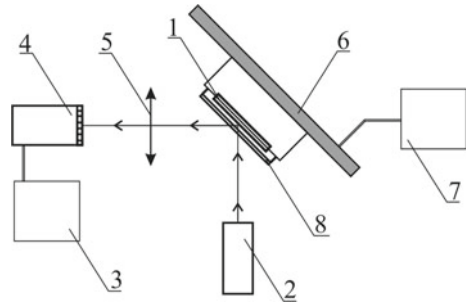
Defect-free zone of waveguides consisting of eight strips with even borders and without overlaps was found using microscope, see Fig. 2. The width of one waveguide equals to  $100 \text{ }\mu\text{m}$ . Distance between waveguides is  $10\text{--}20 \text{ }\mu\text{m}$ .

**Fig. 2** Scaled-up size of defect-free zone of strip waveguides



**Fig. 3** Test bench diagram:

1—strip waveguide prototype, 2—laser, 3—PC;  
 4—TV camera matrix,  
 5—lens, 6—piezoexciter,  
 7—excitation generator,  
 8—slit for identification of illuminated region of waveguides



During visual inspection of the waveguides using microscope, in the process of microscope field depth adjustment, it was found that all the strips are single plane-oriented and have fairly flat form in section.

In order to evaluate performance characteristics of the prototype with micromechanical strip waveguides in a free state, the test bench was made; its diagram is shown in Fig. 3.

The prototype was glued to piezoelectric transducer. The piezoelectric transducer was excited by a tunable sine wave oscillator. Converse piezoelectric effect became a reason for deformation in piezoelectric transducer which caused vibration of the prototype together with microstrip waveguides. Standing flexural waves produced inside the microwave guide are a diffraction grating with a period equal to half of standing wavelength. Diffraction pattern appearing in reflected light during laser interaction with periodic texture of microwave guides was directed to TV camera matrix through the lens.

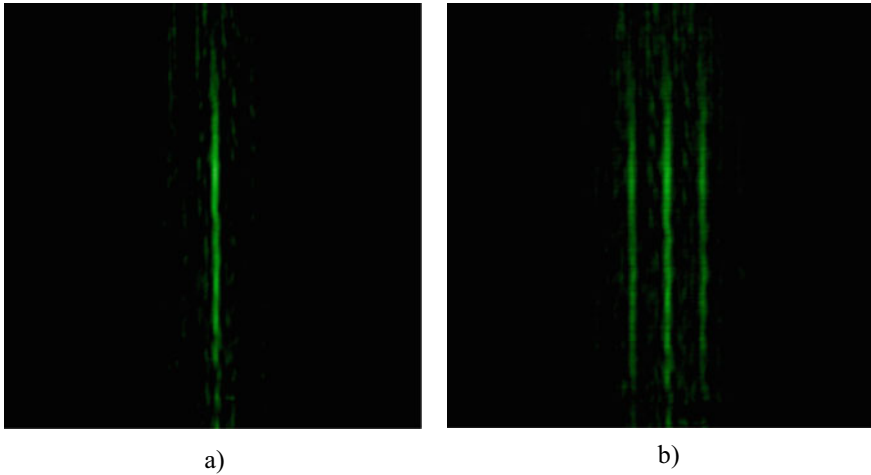
A slit with the width of 1 mm was placed above the defect-free zone of the waveguides in parallel to waveguides and horizontally. Radiation was directed to the defect-free zone of waveguides where diffraction in reflected light had been detected. Then, it entered the lens in the focal plane of which TV camera matrix was located. TV camera was connected to a personal computer via USB port.

Results in the form of pictures of diffraction patterns were recorded in PC. Laser power was 5 mW and remained constant; wavelength was  $0.532 \mu\text{m}$ . TV camera pixel dimension was  $2.8 \mu\text{m}$ .

The camera was set as follows: At first, the camera was set automatically (auto mode). In this mode, the pattern was light-struck and dynamic range was limited. In order to eliminate this problem, auto mode was subsequently turned off and laser power was being reduced until a good-quality pattern was obtained.

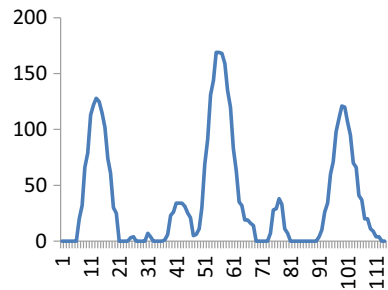
Following patterns are shown in Fig. 4a in a free state (without piezoelectric excitation) and in Fig. 4b when flexible waves appeared (with piezoelectric excitation) with oscillator frequency of 336 kHz.

Relative intensity distribution graph was obtained for (central) horizontal row 491 of the diffraction pattern. Vertical line shows relative radiation intensity and horizontal-geometric position of incident radiation in pixels (Fig. 5).



**Fig. 4** Example of pictures obtained by TV camera matrix located in the focal plane of the lens

**Fig. 5** Relative intensity distribution of radiation affecting middle row of TV camera matrix



Position of central maximum on the figure corresponds to 56 pixels since the lens with focal distance of 50 mm and the laser with a wavelength of  $0.532 \mu\text{m}$  are used in this experiment. Minus first maximum is 43-pixel lower than the central and plus first maximum is 42-pixel lower, which is equivalent to  $120.4 \mu\text{m}$  to each maximum. Calculations show that obtained diffraction pattern is formed by diffraction grating with a period of  $220 \mu\text{m}$  [20] as per calculations.

### 3 Results

The studies have confirmed smooth operation of the prototype model and allowed us to define its following parameters:

- Useful excitation frequency range of the prototype model used to observe the diffraction pattern is from 196 to 336 kHz;

- Flexural wave range obtained under piezoexcitation is 220–525  $\mu\text{m}$ , which corresponds to retunable range of diffraction grating period;
- Maximum excitation frequency with diffraction was 336 kHz;
- Minimum flexural wavelength of waveguides was 220  $\mu\text{m}$  under excitation frequency of 336 kHz;
- Assessment of flexural wave amplitude, when it is still possible to observe a diffraction pattern under reflection of visible light, is 100–200  $\text{m}\mu$ .

## 4 Conclusion

Experimental studies carried out by the authors have defined basic functional characteristics of the prototype model with strip micromechanical waveguides in a free state with piezoelectric oscillation excitation. The principal potential for use of strip micromechanical waveguides as retunable diffraction gratings was confirmed.

In the future, technique improvement is assumed as follows: quantitative evaluation of excited flexural wave amplitude of waveguides and evaluation of intensity distribution of the obtained diffraction pattern.

## References

1. Chesnokov, V., Chesnokov, D., Reichert, V.: Piezoelectric excitation of elastic bending waves in free thin-film structures. In: GEO-Siberia-2011, vol. 5, pp. 55–63, SSGA, Novosibirsk (2011)
2. Hass, J.: Physics of Thin Films, vol. 7. Mir, Moscow (1977)
3. Gugel, B.: Phosphors for Electrovacuum Industry. Energia, Moscow (1967)
4. Chesnokov, V., Fedchenko, V., Naz'mov V.: Study of SR beam line windows within the range of 17–80 nm. Nucl. Instr. Meth. Phys. Res. **308**, 333–335 (1991)
5. Chesnokov, D., Shergin, S., Nikulin, D.: Laser perforation of thin-film membranes. In: Congress GEO-Siberia-2007, vol. 1, pp. 220–224, SSGA, Novosibirsk (2007)
6. Nikulin, D., Shergin, S., Chesnokov, A.: Thin-film membrane microtrapharets for vacuum sputtering of drawings on a substrate. In: Collection of Scientific Works of Postgraduates and Young Scientists, vol. 4, pp. 21–24 (2007)
7. Nikulin, D., Reichert, V., Shergin, S.: Study of the layout of a strip micro-waveguide for its use as a tunable diffraction grating, vol. 1, pp. 237–243. SSUGT, Novosibirsk (2018)
8. Korneev, V.: Development and research of optical magnetically controlled micromechanical devices: thesis of candidate of technical sciences: 01.04.05. SSGA, Novosibirsk (2010)
9. Chesnokov, D.: Micromechanical deflector of light streams. Opt. J. **4**, 51–54 (2007)
10. Knyazev, I.: Modeling of dynamic characteristics of switching elements of micro-optoelectromechanical tunable diffraction grating. Vestn. Sgugit **22**(1), 235–251 (2017)
11. Chesnokov, D., Chesnokov, V., Nikulin, D.: Electrostatic and laser excitation of elastic waves in thin-ribbon strip waveguides. In: GEO-Siberia-2007, pp. 203–209, SSGA, Novosibirsk (2007)
12. Danilin, B.: Vacuum Application of Thin Films. Energia, Moscow (1967)
13. Okatov, M., Antonov, E., Baigozhin A.: Handbook of Technologist-Optics. Polytechnic, SPb (2009)
14. Korneev, V.: Features of spectral characteristics of micromechanical controlled diffraction grating. In: GEO-Siberia-2009, vol. 4, pp. 24–28, SSGA, Novosibirsk (2009)

15. Korneyev, V.: Micro-opto-mechanical scanner for terahertz spectrum diapason. In: 9-th International Symposium on Measurement Technique and Intelligent Instrument, vol. 2, pp. 361–365, SPb (2009)
16. Korneev, S.: Experimental investigation of torsional vibrations of the strips of the MEMS controllable diffraction gratings. *Bull. SSGA* **1**(12), 177–181 (2010)
17. Korneev, V.: Calculation of the amplitudes of oscillation for rectangular and circular membranes form. *Bull. SSGA* **4**(22), 213–220 (2017)
18. Mysel, L., Glang, R.: *Thin Film*. Sov. Radio, Moscow (1977)
19. Grigoriev, I., Melikhov, E.: *Physical Quantities*. Energoatomizdat, Moscow (1991)
20. Chesnokov, D., Chesnokov, V., Nikulin, D.: Diffraction of light on elastic waves in thin-film membrane structures. In: *GEO-Siberia-2007*, vol. 1, pp. 201–203. SSGA, Novosibirsk (2007)

# The Use of Digital Cameras for Multispectral Registration with an Unmanned Aircraft



Evgenij Gritskevich , Sergei Novikov , Polina Zvyagintseva ,  
Diana Makarova , Marina Egorenko , and Aelita Shaburova 

**Abstract** Monitoring of the environment with the help of unmanned aerial vehicles is currently one of the most developing branches of optoelectronic instrument-making, since the digital cameras installed on these devices make it possible to survey the underlying surface in order to further highlight the signs of this surface that carry information about its state. The use of unmanned aerial vehicles for the control of agricultural lands is a particular and very perspective case of such monitoring. The technique of measuring the spectral reflection coefficients of surfaces used to identify the state of vegetation observed in the field of view of multispectral digital camera, which monitors the environment from the board of unmanned aerial vehicles. The method allows to determine the spectral reflectance of the calibration surfaces against the reference surfaces. The results of the work are applied in the analysis and processing of images obtained in the course of the unmanned aviation system that monitors agricultural lands.

**Keywords** Spectral reflectance coefficient · Reference surface · Calibration surface · Working surface · Spectrophotometer · Unmanned aerial vehicle · Measurement technique

## 1 Introduction

For the purposes of remote sensing, unmanned aircraft systems are currently actively used. At the same time, an unmanned aircraft, which is part of such a system, can be equipped with various shooting cameras. Widely used are digital cameras that shoot in the visible range of the electromagnetic spectrum and specialized multispectral cameras that shoot in several fairly narrow spectral ranges. Such cameras are used to determine various qualitative and quantitative characteristics of the underlying surface [1, 2].

---

E. Gritskevich · S. Novikov · P. Zvyagintseva (✉) · D. Makarova · M. Egorenko · A. Shaburova  
Siberian State University of Geosystems and Technologies, Novosibirsk, Russia  
e-mail: [p.a.zvjaginceva@sgugit.ru](mailto:p.a.zvjaginceva@sgugit.ru)

In the agricultural sector seems quite topical solution to the problem of operational handling situations related to the discovery, identification and subsequent timely response to various signs of vegetation condition that requires immediate attention.

The most reliable and simple way to detect signs indicative of a particular state of plants is to change the reflective properties of their surfaces according to the spectral composition of reflected solar radiation [1, 3–7]. Analysis of such changes allows:

- identify areas with oppressed vegetation;
- determine the ripening period of plants;
- establish wetland segments;
- find vegetation affected by diseases or pests, etc.

Thus, the development of a methodology that allows one to determine the reflective properties of the underlying surface seems to be an actual practical problem, for the solution of which multispectral cameras based on matrix photodetectors are used. Therefore, the object of the study in the proposed work will be similar systems installed on board an unmanned aircraft and designed to identify the underlying surface by the spectral composition of the reflected radiation.

Accordingly, the subject of research is a technique for identifying the state of a surface by the spectral composition of reflected radiation. The aim of the study is to develop such a technique that involves modeling the processes of formation of optical images on a photodetector array. The main attention will be paid to the preparation of initial data for such a simulation, namely: obtaining the spectral characteristics of scanned sections of the underlying surface, as one of the most important identifying signs of the current state of vegetation located on this surface.

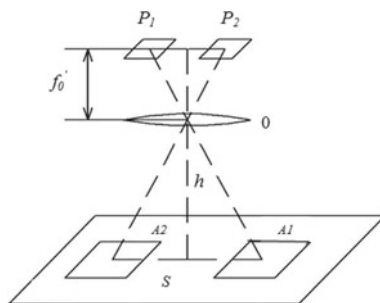
## 2 The Mathematical Model of Energy Transformations Implemented by the Optical Channel of a Digital Camera

Figure 1 shows a simplified diagram of the shooting of the underlying surface from the board of an unmanned aircraft using one digital camera.

An unmanned aircraft is located at a height  $h$  above the underlying surface and incorporates a digital camera. Figure 1 shows the lens  $O$  of digital camera with rear focal length  $f'_0$ , moreover  $h \gg f'_0$ . Elementary photodetectors  $P_1$  and  $P_2$  are the pixels of the photodetector matrix of a digital camera. These pixels are optically coupled to elementary regions of the underlying surface  $A_1$  and  $A_2$ . Accordingly, the illuminances of the photosensitive areas  $P_1E'_1$  and  $P_2E'_2$  are determined by the area average (spatial-integral) illuminances of the areas  $A_1E_1$  and  $A_2E_2$ , as well as the reflective properties of the latter.

In the model under consideration, the following assumptions and limitations are introduced.

Firstly, the illumination of the underlying surface is formed due to direct and (or) scattered solar radiation.



**Fig. 1** Simplified shooting scheme with a single camera.  $S$ —the underlying surface;  $O$ —lens of a digital camera;  $A_1$  and  $A_2$ —elementary sections of the underlying surface, optical conjugate with elementary photodetectors  $P_1$  and  $P_2$ ;  $h$ —the flight altitude of the unmanned aircraft system above the underlying surface;  $f'_0$ —back focal length of a digital camera

Secondly, during the observation time this illumination does not change.

Thirdly, the underlying surface is a diffuse (Lambert) reflector.

Fourth, the photodetectors  $P_1$  and  $P_2$  are absolutely identical.

Fifthly, the flight altitude of an unmanned aircraft above the underlying surface  $h$  (length of the observation path) does not exceed 200 m, and the meteorological range of atmospheric visibility exceeds 20 km. The last assumption (along the path length and the meteorological visibility range) allows us to consider the effect of radiation losses in the atmospheric channel to be negligible and not to take into account the change in the height of the unmanned aircraft system above the underlying surface when determining the conditions and illumination values of this surface and the photodetector during video recording (except for changes sizes of sites  $P_1$  and  $P_2$ ).

Sixth, the material of which the lens is made is selected for operation in the spectral region of the optical radiation, determined by the spectral sensitivity of the photodetector, and its characteristics (primarily spectral transmittance) do not significantly affect the optical signal detected by the photodetector.

The assumptions and limitations listed above allow the mathematical formalization of the model at a level sufficient for the engineering use of the technique in question. At the same time, they are not artificial or compulsory in nature, but are conditions usually used in engineering calculations of optoelectronic systems. Restrictions on flight altitude are determined by the natural conditions of operation of an unmanned aircraft.

Let the natural spectral illumination of the underlying surface be  $E_n(\lambda)$ , and sections  $P_1$  and  $P_2$  are covered with vegetation with spectral reflection coefficients  $\rho_1(\lambda)$  and  $\rho_2(\lambda)$ , respectively. It is assumed that these areas are covered with different types of vegetation. The spectral density of the brightness of the sections will be respectively [8, 9]:

$$L_1(\lambda) = \frac{E_n(\lambda)}{\pi} \rho_1(\lambda), \quad (1)$$



$$L_2(\lambda) = \frac{E_n(\lambda)}{\pi} \rho_2(\lambda) \quad (2)$$

Let sections  $P_1$  and  $P_2$  be next to each other near the line of sight of the lens (in the paraxial region). If we take the brightness of the sections as the initial useful signals, then the difference between the signals from each section will be determined only by the difference between the spectral reflection coefficients of the coatings of these sections. Since expressions (1) and (2) are of the same type, to reduce the mathematical records in the following presentation, we will use the notation of the generalized spectral quantities  $\rho(\lambda)$  and  $L(\lambda)$ . Since the spectral reflection coefficients of different types of underlying surfaces differ from each other in magnitude at different wavelengths, it is advisable to use the spectral filtering method [10] to solve the problem of detecting these differences, which involves introducing a spectral filter with a spectral transmittance into the optical channel of a digital camera  $\tau(\lambda)$ .

Taking into account the above assumptions for the model under consideration, the generalized signal at the output of any elementary photodetector can be calculated by the formula [9, 11]:

$$u = \frac{1}{4} \left( \frac{D_{\text{ep}}}{f'_0} \right)^2 \int_{\Delta\lambda_{\text{sf}}} E_n(\lambda) \rho(\lambda) \tau(\lambda) S(\lambda) d\lambda, \quad (3)$$

where

- $D_{\text{ep}}$  diameter of the entrance pupil of the digital camera lens;
- $\Delta\lambda_{\text{sf}}$  spectral bandwidth of the spectral filter (passband);
- $S(\lambda)$  absolute spectral sensitivity of the photodetector matrix in terms of illumination.

The essence of the method for detecting spectral reflection coefficients of underlying surfaces is as follows. Immediately before the start of the flight of an unmanned aircraft, a preliminary spectral calibration of the digital camera is performed on the reference surface. The calibration technique is described in detail in [12]. At the same time, in order to take into account possible changes in the lighting conditions, it is necessary to provide the possibility of spectral calibration for different points of the shooting area at different times (achieved by placing several spectral standards over the area of the shooting area). As a result of calibration, the output signal of the photodetector array for the used spectral filter is fixed. It is obvious that  $u_{\text{st}} \sim \rho_{\text{st}}(\Delta\lambda_{\text{sf}})$ , where  $\rho_{\text{st}}(\Delta\lambda_{\text{sf}})$ —the average value of the spectral reflection coefficient of the reference surface in the passband  $\Delta\lambda_{\text{sf}}$ . During the flight of an unmanned aircraft, the output signal of the photodetector matrix  $u_w$  from the working (test) surface is recorded. Then  $u_w \sim \rho_w(\Delta\lambda_{\text{sf}})$  in bandwidth  $\Delta\lambda_{\text{sf}}$ . The proportionality coefficients are the same in both cases, therefore:

$$\rho_w(\Delta\lambda_{\text{sf}}) = \frac{u_w}{u_{\text{st}}} \rho_{\text{st}}(\Delta\lambda_{\text{sf}}) \quad (4)$$

Of course, a single reflection coefficient at a single wavelength is not enough to identify the type of underlying surface. Therefore, an unmanned aircraft is equipped with several identical digital cameras with different spectral filters installed in them. Accordingly, there should be several reference surfaces when calibrating the system.

Reference surfaces used for calibration are selected on the basis of the need to ensure the required dynamic range of measurements (from light to dark tones). In addition, they should have a uniform reflection coefficient over the entire spectral range of the sensitivity of the photodetector. The functions of the spectral reflection coefficients of the reference surfaces are pre-measured in the laboratory using a spectrophotometer.

Discrete reflection coefficients of a limited set of a priori assumed working underlying surfaces for discrete bandwidths of applied spectral filters can be taken from specialized reference books, for example, [3–5, 13], or can also be obtained by laboratory spectrophotometric measurements.

When performing an analysis of the object composition of the underlying surface on multispectral images, various algorithms of automated decryption can be used, which significantly increases the efficiency of obtaining information about the area being shot.

### **3 Multispectral Method for Identifying the State of a Reflecting Surface Using a Calibration Procedure**

So, the main way to obtain information about the reflective properties of underlying surfaces is multispectral shooting from an unmanned aerial vehicle, which implies the installation of several identical digital cameras on such a vessel, each of which is equipped with an individual absorbing spectral filter that transmits radiation in a narrow wavelength range (quasi-monochromatic radiation) [2, 10, 14]. Thus, the photodetector arrays of each camera record an image only for a specific spectral interval. Accordingly, the signals at the output of a separate camera will correspond to the quasi-monochromatic radiation for which the spectral filter used in it is intended. If you place a surface with a known function of distributing spectral reflection coefficients over wavelengths in the field of view of the cameras, then the output signals taken from the pixels of the photodetector arrays can be used as reference signals for calibration of other signals that do not get the image of the calibration surface.

Before or during the flight, the signals from the outputs of the pixels on which the image of the calibration surfaces is located are recorded in the memory of the digital camera for each spectral range used. The signals taken from the outputs of the pixels on which the image of the working surfaces under study is located are quantitatively compared with calibration signals.

Thus, for the identification of working surfaces, it is necessary to know the values of the reflection coefficients of radiation in the distinguished spectral ranges, both for the calibration surface and for the working surfaces under study. These values

must be obtained in advance (known a priori) either from the results of experimental laboratory studies, or from the corresponding reference literature, or determined by a combined method, i.e., using both experimental and reference data.

An analysis of the literature on the identification of the underlying surface by the spectral composition of the reflected radiation [15, 16] showed that the calibration method is the main method used when using unmanned aircraft in agriculture.

In the literature devoted to the identification of various types of vegetation state [17–19], the so-called index is most often used as an identifying quantitative trait *NDVI* (normalized difference vegetation index), calculated by the formula

$$NDVI = \frac{\rho_{IR} - \rho_R}{\rho_{IR} + \rho_R}, \quad (5)$$

where

$\rho_{IR}$  reflection coefficient of the investigated surface in the infrared region of the spectrum;

$\rho_R$  reflection coefficient of the same surface in the red region of the spectrum.

This index is an indicator of plant health. Chlorophyll absorbs red waves, and the cellular structure reflects near infrared waves. Therefore, a healthy plant, in which there is a lot of chlorophyll and a good cellular structure, actively absorbs red light and reflects near infrared. A diseased plant is the other way around.

## 4 Experimental Research

White teflon plates were used as calibration surfaces. The spectral reflection coefficients of the calibration surfaces were measured by the standard method [20] using an automated complex that included a spectrophotometer and a computer with the appropriate software, that allowed to receive the processed results in a graphic form in real time. The appearance of the complex is shown in Fig. 2.

In the role of reference surfaces, two barite plates were used. This is due to the fact that in the region of interest of the spectrum, barite has a uniform reflection coefficient over all wavelengths, as shown in Fig. 3 obtained from [13] by digitizing the graph given in this work.

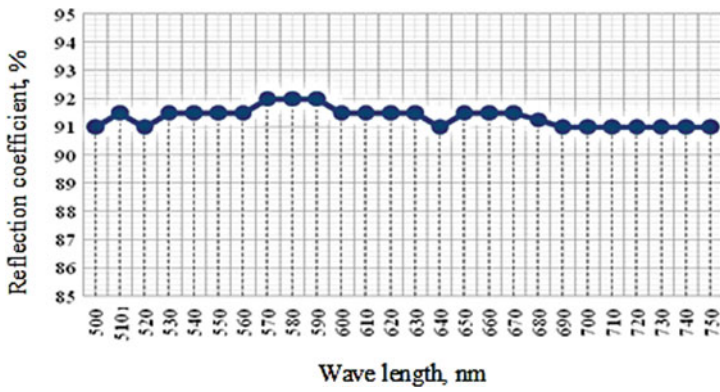
All points of the spectral coefficients were averaged over the points highlighted in the figure, as a result of which a weighted average of 0.914 was obtained.

Figure 4 shows a graph of the Teflon spectral reflection coefficients obtained experimentally and averaged over two barite reference samples. The obtained dependence was used to determine the spectral reflection coefficients of the studied working plant surfaces.

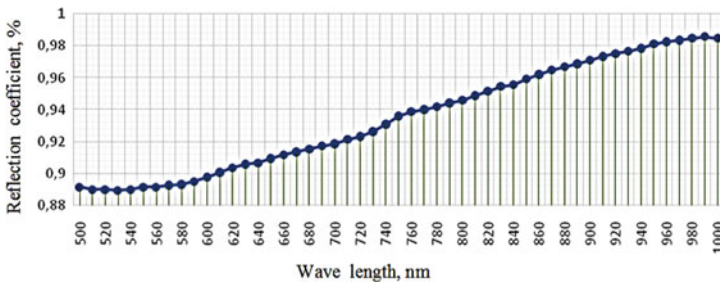
As the working surfaces, green and yellow sheets of a houseplant, shown in Fig. 5.

Figures 6 and 7 are shown graphs of spectral reflection coefficients of yellow and green sheets on a Teflon plate, experimentally obtained on a spectrophotometer.

**Fig. 2** Appearance of an automated measuring complex

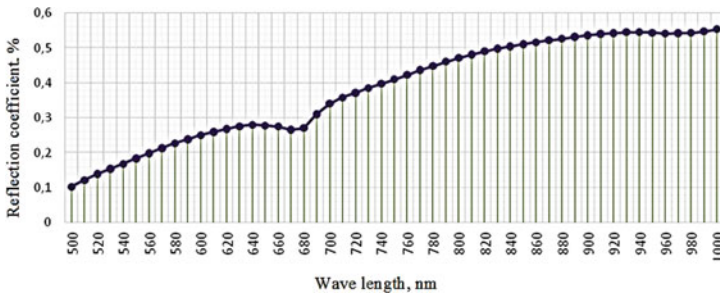


**Fig. 3** Barite spectral reflection coefficient graph

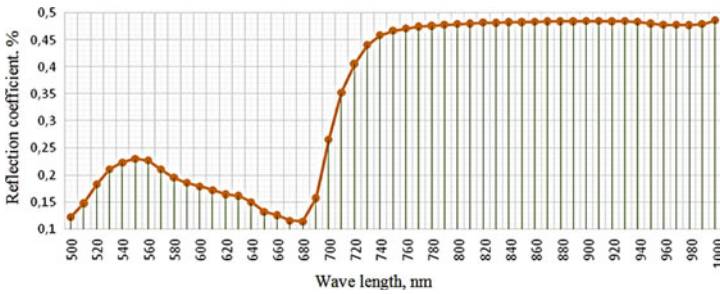


**Fig. 4** Averaged graph of the barite

**Fig. 5** Work reflective surfaces



**Fig. 6** Graph of the spectral reflection coefficient of the yellow sheet on a Teflon plate, experimentally obtained on a spectrophotometer



**Fig. 7** Graph of the spectral reflection coefficient of a green sheet on a Teflon plate, experimentally obtained on a spectrophotometer

## 5 Recommendations on the Use of the Model and Conclusions on the Work

For the effective application of the results obtained in the practical use of multispectral digital cameras placed on an unmanned aircraft for the needs of agricultural production, it is necessary to fulfill a number of recommendations discussed below.

It is desirable to have several Teflon plates, for each of which, it is necessary to measure the spectral reflection coefficients. This will allow us to average the results of observations, which will provide an increase in the accuracy of measurements.

The flight time of an unmanned aircraft must be limited in order to avoid changes in lighting conditions for one shooting period.

When scanning small areas from the air, it is advisable to place Teflon plates evenly along the intended flight path. If a large area is studied or hard-to-reach areas are located on the studied territory, then calibration of chambers using several Teflon samples should be done before the start of the flight, placing these samples in a local area. At the same time, shooting should be done several times from different heights. The results of such surveys must be averaged.

The considered model of energy transformations carried out by the optical channel of a digital camera, together with the proposed method for measuring the spectral reflection coefficients of various diffuse surfaces, as a way to obtain arrays of initial data for modeling, allows developing and technically implementing automated methods for identifying underlying surfaces during remote sensing of them from an unmanned aerial vehicle vessel.

## References

1. Multispectral analysis for determining plant health. <https://publiclab.org/wiki/multispectral-analysis-for-determining-plant-health>. Last accessed 2019/10/26
2. Drone data management system. <https://event38.com/drone-data-management-system/>. Last accessed 2019/10/28
3. Dzhabiev, A., Ishanin, G., Pankov, E.: Optical Radiation of Natural Objects and Backgrounds and Its Imitation. GITMO, Saint-Petersburg (2001)
4. Schauvenerdt, R.: Remote Sensing. Models and Methods of Image Processing. Technosphere, Moscow (2010)
5. Danilin, I., Medvedev, E., Melnikov, S.: Laser Location of the Earth and Forest. Forest Institute named after V.N. Sukacheva SB RAS, Krasnoyarsk (2005)
6. Richardson, A., Wiegand, C.: Distinguishing vegetation from soil background information. *Photogram. Eng. Remote Sens.* **43**, 1541–1552 (1977)
7. Crist, E., Cicone, C.: Application of the tasseled cap concept to simulated thematic mapper data. *Photogram. Eng. Remote Sens.* **50**, 343–352 (1984)
8. Schroeder, G., Traiber, H.: Technical Optics. Technosphere, Moscow (2006)
9. Yakushenkov, Yu.: Fundamentals of Optical-Electronic Instrumentation. Logos, Moscow (2013)
10. Tarasov, V., Yakushenkov, Yu.: Dual and Multi-band Optoelectronic Systems with Matrix Radiation Detectors. University Book, Logos, Moscow (2007)

11. Ishanin, G., Chelibanov, V.: *Optical Radiation Receivers*. Publishing House "Lan", Saint-Petersburg (2014)
12. Vasin, B., Malkova, S., Osipov, M.V., Puzyrev, V.N., Saakyan, A.T., Starodub, A.N., Fedotov, S.I., Fronya, A.A.: Method for measuring the spectral sensitivity of a CCD matrix. Preprint FIAN 18,18.-20 (2007)
13. Krinov, E.: *Spectral Reflectivity of Natural Formations*. Publishing House of the Academy of Sciences of the USSR, Moscow-Leningrad (1947)
14. Parrot Sequoia. [https://pdf.directindustry.com/pdf/airinov/parrot-sequoia/177048-660644-\\_8.html](https://pdf.directindustry.com/pdf/airinov/parrot-sequoia/177048-660644-_8.html). Last accessed 2019/10/26
15. Clevers, J.G.P.W.: The derivation of a simplified reflectance model for the estimation of leaf area index. *Remote Sens. Environ.* **35**, 53–70 (1988)
16. Lillesand, T., Kiefer, R.: *Remote Sensing and Image Interpretation*, 2nd edn. Wiley, New York (1987)
17. Baret, F., Guyot, G.: Potentials and limits of vegetation indices for LAI and APAR assessment. *Remote Sens. Environ.* **2**(3), 161–173 (1991)
18. Crippen, R.: Calculating the vegetation index faster. *Remote Sens. Environ.* **34**, 71–73 (1990)
19. Leprieux, C., Verstraete, M., Pinty, B., Chehbouni, A.: NOAA/AVHRR vegetation indices: suitability for monitoring fractional vegetation cover of the terrestrial biosphere. In: *Proceedings of Physical Measurements and Signatures in Remote Sensing, ISPRS*, pp. 1103–1110 (1994)
20. Laricheva, E., Mercov, S., Sokolova, Yu.: *Optical Spectroscopic Methods of Analysis. Laboratory Workshop*. NRNU MEPhI, Moscow (2010)

# Computer Model for Analysis of the Process of Image Construction in Optical-Electronic Visualization Systems



Evgenij Gritskevich , Marina Egorenko , Diana Makarova , Sergei Novikov , Alexey Polikanin , and Aelita Shaburova 

**Abstract** A simulation computer model of an image visualization system is considered. The model allows to analyze the effectiveness of the use of such systems together with the observer's eye according to the probability criterion for solving the observational problem using Johnson's equivalent bar. This method of formalizing input actions provides the possibility of their analytical presentation in the form of a set of harmonic signals, which greatly simplifies the modeling task while maintaining the adequacy of the model. The proposed simulation model is intended for computer analysis of any visualization systems that include links of various physical nature. In this case, each link is displayed by its own model, taking into account the features of the link, but the composition of the information signal parameters at the input and output of each link does not change, which allows a uniform methodological approach to the development of software modules that implement link models. At the model output, the probability of solving the observational problem by the observer's visual apparatus on the device's screen is calculated at the required degree of decryption of the object. This allows for a given probability to determine the range of the device, and for a given range and probability—to carry out optimal coordination of link parameters. In addition, the model provides the calculation of the most important characteristics of visualization systems, for example, the modulation transfer function, which, combined with the calculation of the effective values of the noise fluctuations of the output image, makes it possible to generate these images for typical objects of observation on the screen of a computer monitor, as when operating a real visualization system.

**Keywords** Visualization system · The probability of solving an observational problem · Object-background situation · Optoelectronic path · Simulation model

---

E. Gritskevich · M. Egorenko (✉) · D. Makarova · S. Novikov · A. Polikanin · A. Shaburova  
Siberian State University of Geosystems and Technologies, Novosibirsk, Russia  
e-mail: [e\\_m\\_p@mail.ru](mailto:e_m_p@mail.ru)



## 1 Introduction

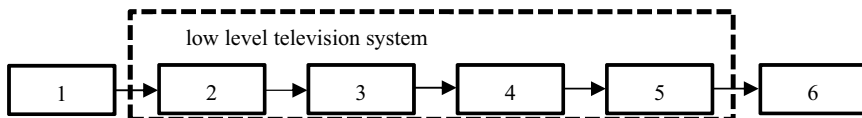
Modern image visualization systems are modular optoelectronic complexes consisting of links of various physical nature. Computer analysis of such systems involves an end-to-end simulation of the process of information signals passing through all modules of the optoelectronic path from input to output, that is, from an external module that implements the object-background situation in which the device operates, to a module that provides the construction of the output image under study the eye of the observer. An objective criterion for assessing the joint work of the visualization system and the operator is the probability of the eye detecting those details in the image whose size determines the required degree of decryption of the object when solving the observation problem (detection, recognition, identification).

Imaging systems include night vision devices. The basis of modern night vision devices is integrated modules, consisting of an electron-optical converter and a matrix detector, structurally integrated into one housing. The screen of the electron-optical converter can be optically paired with the photosensitive plane of the matrix. Another way to convert the input image into an electrical video signal is to transfer the electronic image obtained at the output of the microchannel plate into the plane of the electron-sensitive matrix without creating an intermediate optical image. Such modules are called the fifth-generation electron-optical converter [1]. The night vision devices themselves, implemented on the basis of these modules, are often called low-level television systems.

## 2 The Structure of the Simulation Computer Model of the Visualization System

The model is formed according to the block principle, where each block is a separate software module that simulates the operation of a specific functional link of the optoelectronic path. The block diagram of the model is shown in Fig. 1.

Modeling of the processes of the passage of signals through a link should take into account the distortion of the signals introduced by this link and the possible change in the physical nature of the signals (optical radiation—electrical analog video signal—digital code). The model of an individual link implements in the program code a



1 – object-background situation; 2 – lens; 3 – electron-optical converter of the 5th generation; 4 – image processor; 5 – information display unit; 6 – visual analyzer.

**Fig. 1** Block diagram of a simulation computer model of a low-level television system

generalized transfer function of this link, which takes into account the modulation of the useful signal (object—background) and its frequency transformations, the presence of a dark signal, and noise fluctuations. The basic computational procedure implemented by the simulation model is an end-to-end analysis procedure that simulates the passage of signals from the input to the output of the device, that is, from the “object-background situation” module to the “visual analyzer” module.

The first program module “object-background situation” is designed to generate a limited set of parameters and characteristics of the input signals converted by subsequent modules. The last module “visual analyzer” should calculate the criterion for distinguishing between the eye of the observer the image of the object of interest with a given degree of decryption of the latter (as a rule, detection or recognition).

### 3 Principles for the Implementation of the Mathematical Model of the Optical-Electronic Path of the Visualization System

Since the process of visual perception is random in nature, the criterion for solving the observation problem can be the probability that the observer’s eye detects details in the image whose size determines the required degree of decryption. To calculate the value of this probability, we use the technique obtained by converting the analytical expressions given in [2].

In this work, a method for determining the logarithmic threshold contrast of the brightness of the images of the object and the background on the television screen is given. This contrast is defined as:

$$\delta_{\text{thr}} = \delta_0 \prod_{i=1}^5 f_i, \quad (1)$$

where  $\delta_0$ —the magnitude of the logarithmic contrast sensitivity of the eye for normalized observation conditions (probability of detection is 50%, the equality of the brightness of the background and the field of adaptation of the eye, the absence of noise, large angular sizes of the image of the object, unlimited observation time);  $f_i$ —functions that take into account the deterioration of normalized conditions. These functions are approximations of the experimental data presented in [3–6].

Function  $f_1 = f(\alpha_o)$  expresses the dependence of the threshold contrast on the magnitude of the angular size of the image of the object  $\alpha_o$ . Obviously, in this case, an object must be understood as its part, when resolving with the eye on the screen, the observational task is considered to be performed at the required level of decryption. Function  $f_2 = f(L_a/L_d^b)$  takes into account the difference between the brightness of the field of adaptation of the eye  $L_a$  and background on the display screen  $L_d^b$ . Function  $f_3 = f(\sigma_n)$  takes into account the presence of noise having an effective value  $\sigma_n$ , equivalent to the logarithm of brightness. Function  $f_4 = f(t)$  expresses

the dependence of the threshold contrast on the time of presentation of the signal  $t$ . Finally function  $f_5 = f(P)$  takes into account the dependence of the threshold contrast on the probability of its detection by the eye:

$$f(P) = 1 + \alpha_P/2, \quad (2)$$

where  $\alpha_P$ —relative deviation of a random normal variable, i.e.,

$$P = F(\alpha_P),$$

where  $F$ —integral function of the normal distribution law.

Using a function of the form (2) in expression (1) makes it possible to calculate the required probability. Knowing the logarithmic contrast value obtained on the display screen  $\delta$  and considering its threshold (at maximum range), we can write

$$P = F \left\{ 2\delta / \left( \delta_0 \prod_{i=1}^4 f_i \right) - 2 \right\}. \quad (3)$$

The value of  $\delta$  is defined as

$$\delta = \lg(1 + C_d), \quad (4)$$

where  $C_d = |L_d^o - L_d^b|/L_d^b$ —contrast of the image on the display screen between the brightness of the object  $L_d^o$  and background  $L_d^b$ .

Having obtained specific values of  $\delta$  and functions  $f_i$  for given external conditions of observation, we can calculate the probability of solving the observational problem. This, in turn, makes it possible to determine the distance to the object using the iteration method, at which the observational problem is solved with the required probability  $P_{OP}$  (device range) [7].

Expression (2), used as the initial one for calculating the probability of solving the observation problem, was determined in [2] based on an analysis of the results of [3], in which the criterion for the threshold perception was the experimentally obtained contrast detection probability of 50%. The same probability value is given in [8–11], which considers the processes of threshold perception of images by the eye. Thus, the algorithm for determining the range of the observational device can be described in the form of the following steps:

1. An end-to-end analysis of the optoelectronic path is carried out, as a result of which the probability of solving the observational problem  $P_{OP}$  with the eye on the device's screen is calculated
2. If this probability is more than 50%, then the length of the observation path increases, otherwise it decreases, and the thorough analysis procedure is repeated.
3. When the probability value reaches 50% within the specified error, the iterative process stops, and the value of the last path length is taken as the range of action.

Increasing the reliability of solving the observational problem is provided by increasing the value of the required probability. So, in [12], this probability for thermal imaging images is set at the level of 80%. The same probability value was determined for night vision devices in [13]. Thus, the determination of the range should not be strictly tied to the probability value of 50%, but may vary when setting the initial data. That is, the value  $P_{OP}$  determined by the researcher based on expert judgment, taking into account all factors affecting the processes of visual perception.

It is necessary to determine the type of input useful signal, the detection of which on the screen with the eye would indicate a solution to the observational problem. For this, the well-known Johnson criterion [14, 15] is used, which allows replacing real objects and backgrounds with equivalent bar. The intensities of light and dark strokes of such a bar are proportional to the brightness of the object and background, and the period of the bar is determined depending on the required degree of decryption of the object and its minimum overall size. The linear spatial frequency of the equivalent bar, reduced to the conditional plane of the object, is defined as

$$\nu_{em} = N_{em}/h_{lens}, \quad (5)$$

where  $N_{em}$ —the number of periods (strokes) corresponding to the required degree of decryption;  $h_{lens}$ —the minimum overall size of the object (for ground equipment—height, for a growth figure—width).

Such a method of generating input actions in a computer model is convenient because it allows one to describe these input actions analytically, providing them with a mathematical formalization with respect to real prototypes. This makes it possible to consider the process of passing a useful signal through the optoelectronic path from both the energy and spatial-frequency points of view, and, in addition, it allows you to quickly evaluate the transmission function of the modulation of the device, as well as determine its maximum and working resolution.

The “object-background situation” module should contain mathematical (tabular and analytical) dependencies that take into account the location, configuration, reflecting the characteristics of objects and backgrounds, their illumination, the size of the observation object and its detailing, transmission and scattering of radiation by an atmospheric channel. It is impossible to foresee the totality of object-background situations in which the device will work. Therefore, it is necessary to limit oneself to the typical choice of simulated situations (in accordance with the functional purpose of the device). It is advisable to set the external observation conditions on the basis of the requirements for field tests of observational instruments.

Some of the functions included in expression (3) can be excluded from it. For example, it is permissible to neglect the difference between the brightness of the field of adaptation of the eye and the brightness of the background on the display screen, given that the observer has the ability to adjust the brightness of the screen, achieving optimal conditions for the perception of the image. In addition, if during the test the specific time interval for the presentation of the signal is not specified, then the corresponding function can also be equated 1.

At the output of the “object-background situation” module, a set of a limited number of signal parameters is formed, called the vector of phase variables [16]. These variables will characterize the information and physical state of the optoelectronic path. The interaction between the links is realized by transferring the vector of phase variables from one program module to another. Each module converts the vector of phase variables in accordance with the mathematical model of a particular link. The vector structure includes: frequency of the main (Johnson) harmonic component of the useful signal (spatial or temporal), signal levels from the object and background, modulation coefficient of the useful signal, ratio of the useful signal to noise. A useful signal is the difference between the levels of signals from the object and the background.

Thus, the sets of input and output parameters of all modules are the same. This allows you to apply a single methodological approach to building a mathematical model of each link and determine the set of required algorithms. The mathematical model should be an unnormalized transfer function of the realized link and include the following procedures: conversion of signal levels for the fundamental harmonic frequency from input to output, taking into account the dark signal and frequency response of the link; recalculation of the signal-to-noise ratio from input to output, taking into account the intrinsic noise of the link; frequency conversion of the fundamental harmonic component of the desired signal.

Let  $S_{\text{in(out)}}^{o(b)}$ —signal level from the object (background) at the input (output) of the link,  $\mu_{\text{in(out)}}$ —signal-to-noise ratio at the input (output) of the link,  $\nu_{\text{in(out)}}$ —the frequency of the main harmonic component of the useful signal at the input (output) of the link. The generalized equations for the coupling of the input and output signal parameters, implemented in a separate software module, have the following form:

$$S_{\text{out}}^{o(b)} = 0.5k_l^S \left\{ S_{\text{in}}^{o(b)} [1 + T_l(\nu_{\text{in}})] + S_{\text{in}}^{b(0)} [1 - T_l(\nu_{\text{in}})] \right\} + S_l^d, \quad (6)$$

$$\mu_{\text{out}} = \left[ (k_l^\mu \mu_{\text{in}})^{-2} + \mu_l^{-2} \right]^{-1/2}, \quad (7)$$

$$\nu_{\text{out}} = k_l^\nu \nu_{\text{in}}, \quad (8)$$

where  $k_l^S$ —energy conversion coefficient of the signal, which makes sense the ratio of the output signal to the input without taking into account the frequency response and the dark signal of the link (energy sensitivity);  $T_l(\nu_{\text{in}})$ —value of the frequency response of a link at a frequency  $\nu_{\text{in}}$ ;  $S_l^d$ —link output dark signal value;  $k_l^\mu$ —conversion coefficient of the input signal-to-noise ratio to the link output, taking into account the spectral composition of noise and the frequency response of the links;  $\mu_l$ —ratio of useful signal to link noise;  $k_{s_o}^\nu$ —conversion coefficient of the value of the fundamental harmonic frequency of the useful signal from input to output of the link, taking into account a possible change in the physical nature of the signal.

A generalized mathematical model of the process of transforming a vector of phase variables from input to output of a link can be represented as an expression:

$$\vec{V}_{\text{out}} = \varphi_l(\vec{V}_{\text{in}}), \quad (9)$$

where  $\vec{V}_{\text{in}}$ —input vector of phase variables,  $\vec{V}_{\text{out}}$ —output vector of phase variables,  $\varphi_l$ —statement defining an expression-based transformation process (6–8).

According to the vector of phase variables calculated on the display screen, the image parameters of the equivalent bar are determined, namely:  $K_m^d$ —output modulation factor,  $\nu_m^d$ —spatial frequency of such modulation,  $\mu_d$ —signal to noise ratio in the output image. Then, under the condition of an unlimited time for the presentation of the signal and the presence of feedback from the operator to the display, the probability of solving the observational problem is calculated as an integral function of the normal distribution law

$$P_{\text{OP}} = F(K_m^d, \nu_m^d, \mu_d), \quad (10)$$

where parameter values  $K_m^d, \nu_m^d, \mu_d$  converted to values  $C_d, \alpha_0, \sigma_n$ , used in determining function values  $f_i$  and quantities  $\delta$ , included in expressions (3) and (4).

The mathematical models of the links underlying the corresponding software modules describe the functioning processes of these links with varying degrees of completeness. The mathematical model of the electron-optical converter takes into account in sufficient detail the physics of the operation of real converters [17]. Models of the object-background situation, input optics, matrix photodetector also provide a high degree of detail of the physical and mathematical description. For some links, generalized formalized descriptions are used, based on the use of external converter parameters and characteristics. This is due, first of all, to the fact that not for each link the researcher has the required information. However, as noted above, the mathematical models of the links can be refined as such information arrives.

For example, using the signal-to-noise ratio for a photoelectronic link  $\mu_l$  in expression (7) can be based on the use of a generalized noise parameter specified by the value of the input threshold signal, or it can involve a detailed account of fluctuations with the separation of additive and multiplicative noises depending on the source of their occurrence (physical nature). Obviously, in the latter case, it is necessary to know the technological features of the manufacture of a specific conversion unit.

In addition, it is assumed that the researcher has the opportunity to select the required details in the description of the link. So, for an electron-optical converter, one can use experimentally obtained values of the modulation transfer function, and one can approximately calculate the modulation transfer function from the limit resolution [18]. The same applies to the modulation transmission function of the input optics, matrix photodetector, display.

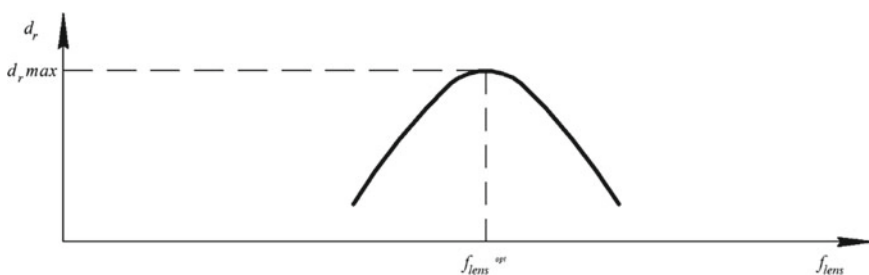
## 4 Information Support of Modeling Processes

Particular attention is paid to information support of the simulation process, which is understood as a set of parameters and characteristics of the typical parts of devices that are required during energy calculations. This primarily relates to the spectral characteristics of natural illumination, the atmospheric channel, objects and backgrounds, and photodetectors. All these characteristics are presented in the form of specialized software modules containing tables of relative values by wavelength. Access to these modules occurs automatically at the request of the main modeling programs. Information modules can be used autonomously as an automated reference book, providing the provision of characteristics in the form of numerical data and graphs at the user's request.

Currently, information modules have been developed that provide modeling of optoelectronic systems in the visible and near infrared (IR) spectral ranges from 0.31 to 1.3  $\mu\text{m}$ . Work is underway to expand the IR range to 1.8  $\mu\text{m}$ , since this spectral region appears to be very promising from the point of view of increasing the information efficiency of night-vision devices [19].

## 5 Discussion of Simulation Results

Using the simulation model, it is possible to conduct virtual tests of a low-level television system that reproduces real laboratory and field tests of systems. If in a real experiment the set of variable parameters of the device controlled by the researcher and the external conditions of his work is limited, then in virtual tests these limitations are determined only by considerations of the physical realizability of the studied factors. Figure 2 shows a graphical dependence of the range  $d_r$  of the observing device from the focal length of the input lens  $f_{\text{lens}}$ , obtained by computer simulation. The scales along the axes of abscissas and ordinates, as well as the dimensions of the controlled quantities are set on the graph in arbitrary units, since their exact values are not important in this case.

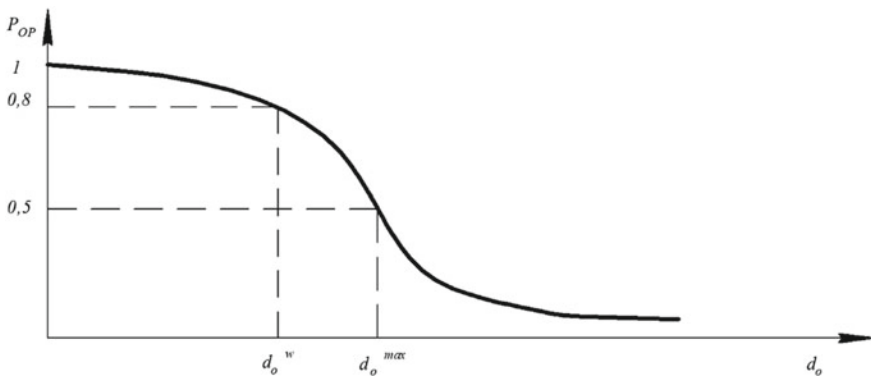


**Fig. 2** The dependence of the range of the observing device on the focal length of the input lens (units and scales along the coordinate axes are arbitrary)

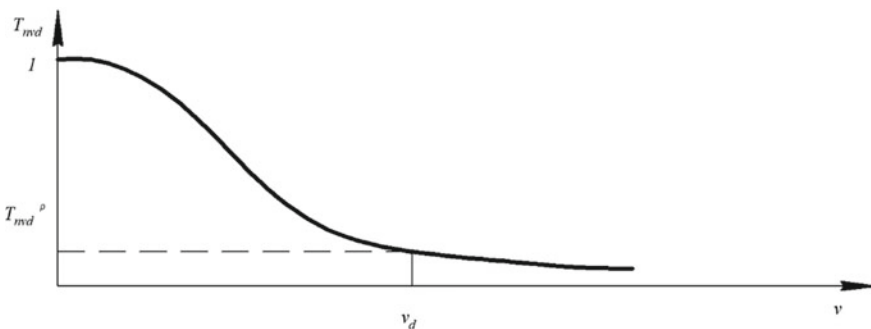
Presence of maximum function graph  $d_r(f_{\text{lens}})$  explained as follows. An increase in focal length leads to an increase in the image scale and, as a result, to an improvement in the overall spatial resolution of the system. At the same time, the energy sensitivity of the device deteriorates due to a decrease in the relative aperture of the input lens (with a constant diameter of the entrance pupil). These multidirectional trends lead to the appearance of the optimal value of the focal length, providing maximum range.

Figure 3 shows a graph of the probability of solving the observational problem  $P_{OP}$  from observation distance  $d_o$  (distance scale and unit—conditional). This graph is also obtained using computer simulation. From the graph you can determine the maximum distance  $d_o^{\text{max}}$ , corresponding to the probability value  $P_{OP} = 0.5$ , and working range  $d_o^w$ , at  $P_{OP} = 0.8$ , and also evaluate their difference.

Another example of the effective application of virtual tests of a night vision device is presented in Fig. 4, which shows the curve of the transmission function



**Fig. 3** Dependence of the probability of solving the observational problem on the observation distance (scale and unit of measurement of distance—conditional)



**Fig. 4** Transmission modulation function of the night vision device (unit of measurement and scale along the spatial frequency axis  $\nu$ —conditional)



of the modulation of the night vision device calculated using the model  $T_{\text{nvd}}(\nu)$  depending on input spatial frequency  $\nu$ .

Here, as before, the scale and unit of measurement of the spatial frequency are arbitrary. The modulation transfer function was obtained as follows. A multiple end-to-end analysis of night-vision devices was carried out at fixed values of the spatial frequency of the equivalent bar, the mathematical image of which was formed by the program module “object-background situation”. After a single pass-through analysis for the current frequency  $\nu_i$  the values of the modulation coefficients of the equivalent bar at the input of the night vision device were determined  $K_m^{\text{in}}(\nu_i)$  (at the output of the “object-background situation” module) and on the display screen  $K_m^{\text{out}}(\nu_i)$ . The modulation transfer function of the night vision device was calculated as

$$T_{\text{nvd}}(\nu_i) = K_m^{\text{out}}(\nu_i)/K_m^{\text{in}}(\nu_i) \quad (11)$$

followed by interpolation of the values.

The resulting transmission function of the modulation of the device can be used for its qualitative assessment and comparison with analogs. According to the set operating value of the modulation coefficient  $T_{\text{nvd}}^w$  can determine the maximum resolution of the device  $\nu_{\text{max}}$ . In addition, the researcher has the opportunity to visualize on the computer monitor screen virtual output images that will create a real device at a given observation distance. For this, a technique is applied that includes the following steps:

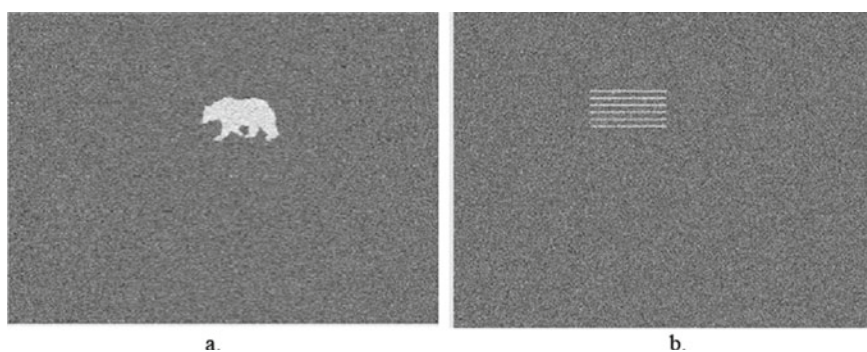
1. From the archive of typical objects of observation, the required one is selected (in accordance with the purpose of the device).
2. Using the model, geometric, modulation (contrast) and noise parameters are determined that characterize the optical signals from the object and the background at the input of the visualization system for a given observation distance.
3. By direct Fourier transform using the integral function of the transmission modulation of night vision devices  $T_{\text{nvd}}(\nu)$  the spatial frequency spectrum of the output signal is calculated.
4. Using the inverse Fourier transform, the output image of the visualization system is generated on the screen of the computer monitor, to which the noise is added. Taking into account that the screen sizes of real observation devices are usually smaller than the screens of computer monitors, the output image is displayed in a scale corresponding to the specified screen size of the video monitoring device of the visualization system.

Playing a virtual output image of a low-level television system allows you to subjectively evaluate the quality of this image. In addition, it becomes possible to analyze the effectiveness of the application of digital image processing methods performed by a specialized processor integrated into the optoelectronic path (block 4 in Fig. 1).

The following is a sequence of on-screen copies explaining the above. Figure 5 shows a view of the ideal image for a typical object (Fig. 5a), the corresponding



**Fig. 5** An ideal input signal and the output image without processing: **a** typical ideal object; **b** equivalent ideal bar; **c** the output image of typical object without processing; **d** the equivalent bar of typical object without processing



**Fig. 6** Output image after processing: **a** typical object; **b** equivalent bar

equivalent bar (Fig. 5b), shows the resulting virtual output images without processing (Fig. 5c, d).

In the latter case, the size of the object, its contrast with respect to the background, and the noise level correspond to those calculated for the output image using the pass-through analysis procedure at a given observation distance.

Figure 6 shows a view of the output image after digital processing. This processing included: low-pass filtering, contouring, contrast enhancement and zooming.

The virtual representation of output images is an effective tool in the analysis of multichannel optoelectronic systems, since it allows you to simultaneously observe the action of each channel in relation to a typical input signal, as well as simulate the process of combining multi-scale images, if provided for in the device.

## 6 Conclusion

The stated principles of model building make it possible to quickly expand the composition of modules, refine and modernize their mathematical models. The operation of individual links can be analyzed autonomously from the operation of the entire

device. The described simulation model meets the principle of building software systems with an open architecture that provides flexibility and efficiency of software product maintenance, as well as the ability to adapt them to the requirements of specific users.

A computer model of a low-level television system allows:

- analyze the operation of a low-level television system at the level of the functional diagram and, based on the results of the analysis, conclude that the device meets its intended purpose;
- compare different instrument variants with each other;
- assess the degree of influence of an individual link on the overall performance of the device;
- virtually generate any physically possible system structure that is physically acceptable from the point of view of a set of software modules.

Quite often, during the operation of real observational instruments, in parallel with the observation process, the spatial position of the object relative to the observer is determined (spatial coordinates are measured). The visualization channel is combined with the measuring channel, to which specific requirements are imposed. Simulation computer simulation allows you to analyze the coordinate measurement process and solve the problem of optimal coordination of the parameters of the links of the optoelectronic path according to the criterion of minimizing the measurement error [20]. Obviously, the combination of two heterogeneous processes in one simulation model—observation and measurement—will significantly expand the possibilities of the proposed test virtualization methods for various types of optoelectronic systems.

## References

1. Degtyarev, E.: Night vision devices: prospects for the development of the component base. *Electron. Sci. Technol. Bus.* **8**, 34–35 (2005)
2. Efimov, A.: Contrast visual sensitivity when watching TV images. *Tech. Film Telev.* **2**, 45–48 (1977)
3. Blackwell, H.: Contrast thresholds of the human eye. *JOSA* **36**(10), 624–632 (1946)
4. Siedentopf, H.: Kontrastschwelle Und Sehscharfe. *Das Licht* **1**, 35–41 (1941)
5. Glezer, V., Zuckerman, I.: *Information and Vision*. Academy of Sciences of the USSR, Moscow (1961)
6. Louise, A.: *Inertia of Vision*. Oborongiz, Moscow (1961)
7. Gritskevich, E.: An iterative method for determining the range of an optical-electronic observation device. In: *News of Universities Series. Instrument Making*, vol. 12, pp. 53–57 (1988)
8. Ratner, E., Matskovskaya, Yu.: On the threshold sensitivity of two-dimensional image receivers. *Opt. Mech. Ind.* **2**, 3–6 (1972)
9. Koshchavtsev, N., Romanov, S., Sokolov, D.: Determination of the Probability of Recognition of Objects of Observation, vol. 167, pp. 38–40. *Tr. Moscow Energy Institute* (1973)

10. Romanov, S., Sokolov, D., Yaroshevich, D.: Relation of the Probability of Recognition of Objects of Observation with the Transfer Characteristic of the System, vol. 192, pp. 139–141. Tr. Moscow Energy Institute (1974)
11. Travnikova, N.: The Effectiveness of Visual Search. Engineering, Moscow (1985)
12. Ovsyannikov, V., Ovsyannikov, Y., Filippov, V.: Improving the reliability of expert assessment of the probability of detection and recognition of objects from thermal imaging images. *Opt. J.* **3**, 1–6 (2012)
13. Sysoev, P., Boldyrev, M., Lopatkin K.: A method for assessing the feasibility of technical requirements for optoelectronic night vision systems based on fifth-generation electron-optical converters. In: *Electronic Equipment. Series 2—Semiconductor Devices*, vol. 1(224), pp. 64–66 (2010)
14. Rosell, F., Willson, R.: Basics of detection, recognition and identification in electro-optical formed imagery. In: *Proceedings of SPIE33*, pp. 107–119 (1972)
15. Lloyd, J.: *Thermal Imaging Systems*. Mir, Moscow (1978)
16. Norenkov, I.: *Computer-Aided Design Systems in Radio Electronics*. Radio and Communications, Moscow (1986)
17. Avdeev, S., Gritskevich, E.: Method for calculating the signal-to-noise ratio in the image on the image intensifier screen. In: *News of Universities Series. Instrument Making*, vol. 11, pp. 73–77 (1988)
18. Gritskevich, E.: Estimation of the optical transfer function of the image intensifier by the magnitude of the limiting resolution. *Opt. Mech. Ind.* **6**, 61–75 (1989)
19. Koshchavtsev, N., Koshchavtsev, A., Fedotova, S.: Analysis of the prospects for the development of night vision devices. *Appl. Phys.* **3**, 66–69 (1999)
20. Gritskevich, E.: Minimization of the measurement error of the optical-electronic coordinate sensor. *Sens. Syst.* **4**, 18–20 (2012)

# Solution of Partial Differential Equations on Radial Basis Functions Networks



Mohie Alqezweeni  and Vladimir Gorbachenko 

**Abstract** The solution of boundary value problems described by partial differential equations on networks of radial basis functions is considered. An analysis of gradient learning algorithms for radial basis functions networks showed that the widely used first-order method, the gradient descent method, does not provide a high learning speed and solution accuracy. The fastest method of the second order, the trust region method, is very complex. A learning algorithm based on the Levenberg–Marquardt method is proposed. The proposed algorithm, with a simpler implementation, showed comparable results in comparison with the trust region method.

**Keywords** Partial differential equations · Radial basis functions networks · Neural network learning · Levenberg–Marquardt method

## 1 Introduction

In the modern industry, digital twin is widely used [1, 2]. A digital twin is a dynamic virtual model of a system, process or service. A digital double is constantly learning and updating its parameters, receiving information from many sensors, correctly represents the state of a physical object. During learning, it uses current data from sensors, from control devices, from the external environment. Digital twins allows real-time monitoring of systems and processes and timely analysis of data to prevent problems before they occur, schedule preventative maintenance, reduce downtime, open up new business opportunities and plan future updates and new developments.

Digital doubles of objects with distributed parameters are mathematically boundary value problems for partial differential equations (PDE) [3]. In most cases, boundary value problems are solved by numerical methods, since analytical solutions exist only for a very limited range of problems. For the numerical solution of boundary value problems for PDE, the methods of finite differences and finite

---

M. Alqezweeni (✉) · V. Gorbachenko  
Penza State University, Penza, Russia  
e-mail: [mohieit@mail.ru](mailto:mohieit@mail.ru)

elements are widely used [4]. These methods require the construction of computational grids. Generating meshes for two- and three-dimensional areas of complex configuration is a complex and time-consuming task. The complexity of grid formation for real problems often exceeds the complexity of solving a system of difference equations [5]. Large computational costs lead to the use of low-order approximations, which provide continuous approximation of the solution on the network, but not its partial derivatives. Modeling of objects with distributed parameters by the methods of finite differences and finite elements is reduced to solving sparse systems of algebraic equations of very large dimension. These systems are characterized by poor conditioning, which requires high costs for their solution. Reconstructing a solution from its discrete approximation is a separate rather time-consuming task.

When modeling complex technical objects, software packages based on the finite element or finite difference method are usually used. However, modeling a real object with their help encounters a number of fundamental difficulties [6]. First, accurate information about differential equations describing the behavior of an object is usually absent due to the complexity of the description of the processes occurring in it. Secondly, to apply the methods of finite elements and finite differences, one needs to know the initial and boundary conditions, information about which is usually incomplete and inaccurate. Thirdly, during the operation of a real object, its properties and characteristics, parameters of the processes occurring in it can change. This requires appropriate adaptation of the model, which is difficult to carry out with models built on the basis of finite element methods and finite differences.

An alternative to finite difference methods and finite elements are meshless methods [7], most of which are projection methods. These methods give an approximate analytical solution in the form of a sum of basis functions multiplied by weights. As basis functions, radial basis functions (RBF) are popular [8, 9]. Methods using RBF allow one to obtain a differentiable solution at an arbitrary point in the solution domain in the form of a function satisfying the required smoothness conditions; they are universal, allow working with complex geometry of computational domains and are applicable for solving problems of any dimension. RBF-based methods require, for the selected parameters of the radial basis functions, to find the vector of weights, so that the resulting approximate solution ensures that the equation and boundary conditions are satisfied with an acceptable error on a certain set of sampling points. For example, the sum of the squared residuals at the sampling points should be small. The main disadvantage of using RBF is the need for unformalized selection of parameters of basis functions.

Promising is the implementation of meshless methods on neural networks. The solution of boundary value problems for PDE is possible on multilayer perceptrons [9, 10]. But, the most promising is the use of radial basis function networks (RBFN) [11], since RBFNs contain only two layers, one of which is linear, and the solution formation is local in nature, which simplifies the learning of such networks. The use of RBFN allows you to configure both weights and RBF parameters during learning networks. Applications of RBNF for solving boundary value problems are considered in the works of Jianyu L., Siwei L., Yingjian Q., Yaping H., Mai-Duy N.,

Tran-Cong T., Sarra S., Chen H., Kong L., Leng W., Kumar M., Yadav N., Vasilieva A. N., Tarkhova D. A., Gorbachenko V. I. [12–15].

To build digital models of twins, it is promising to use the ideas of machine learning and neural networks to build models of real objects. This approach allows you to build adaptive models that are refined and rebuilt in accordance with the observations of the object. Therefore, the urgent task is the development of neural network modeling technologies, a more complete account of historical and newly arriving data, improving methods for automatically adjusting architecture and model parameters, classification and prediction methods [6]. Using neural network models allows us to develop a unified approach to solving various modeling problems. For example, in [16] a unified approach to solving direct and inverse boundary value problems described by partial differential equations was proposed.

The solution to the problem is formed in the learning process RBFN. Therefore, it is important to reduce network learning time. But at present, for learning RBFN in solving boundary value problems, mainly the simplest gradient methods of the first order based on gradient descent are used [10]. Second-order fast methods are practically not used in solving boundary value problems on RBFN. An exception is the confidence area method proposed in [15]. But, the method is very complicated, since it requires at each iteration the solution of the minimization problem to solve the conditional minimization problem.

The aim of this work is to improve the algorithms for learning networks of radial basis functions in solving boundary value problems, which reduce the time of solving the problem.

## 2 Related Works

RBF [8] is the functions of the distance of a space point from a function parameter called the center of the function:  $\varphi(\|\mathbf{x} - \mathbf{c}\|, \mathbf{p})$ , where  $\mathbf{x}$ —the space point,  $\mathbf{p}$ —the vector of function parameters,  $\mathbf{c}$ —the center of the radial basis function,  $\|\mathbf{x} - \mathbf{c}\|$ —the Euclidean norm (distance) between the point and center. Various RBFs are applied. In this paper, we use the Gauss function (Gaussian)

$$\varphi(\|\mathbf{x} - \mathbf{c}\|, a) = \exp\left(-\frac{\|\mathbf{x} - \mathbf{c}\|^2}{2a^2}\right),$$

where  $\mathbf{c}$ —the position of the function center,  $a$ —the shape parameter, often called the width.

When using RBF for solving boundary value problems, the type and parameters of RBF are selected before solving the problem. This procedure is informal, requires experimental verification and does not have unambiguous recommendations. Only some recommendations on choosing RBF and their parameters are known [17].

The solution of boundary value problems using RBF is based on the approximation of functions. Since when solving boundary value problems, an approximation of an

unknown solution is performed, minimization of the residual at the sampling points is used. E. J. Kansa proposed a method for solving boundary value problems using RBF [18, 19], which became the basis for other methods using RBF. We consider the boundary value problem in operator form

$$Lu(\mathbf{x}) = f(\mathbf{x}), \quad \mathbf{x} \in \Omega, \quad Bu(\mathbf{x}) = p(\mathbf{x}), \quad \mathbf{x} \in \partial\Omega, \tag{1}$$

where  $u$ —the solution to the problem;  $L$ —the differential operator; the operator  $B$ —the boundary condition operator;  $\Omega$ —the solution domain;  $\partial\Omega$ —the boundary of the region;  $f$  and  $p$  are known functions.

Inside the solution domain and at the boundary, many sampling points are defined

$$\{\mathbf{x}_i |_{i=1,2,\dots,N} \subset \Omega\} \cup \{\mathbf{x}_i |_{i=N_1+1,N,\dots,N+K} \subset \partial\Omega\}, \tag{2}$$

where  $N$ —the number of sampling points in the inner region of  $\Omega$  and  $K$ —the number of sampling points on the border of  $\partial\Omega$ .

The solution to the problem is in the form of a weighted sum of basis functions.

$$u_{\text{RBF}}(\mathbf{x}) = \sum_{j=1}^M w_j \varphi_j(\mathbf{x}), \quad \mathbf{x} \in \bar{\Omega} = \Omega \cup \partial\Omega \tag{3}$$

where  $\varphi_j$ —RBF;  $w_j$ —weights,  $M$ —the number of RBF.

In (3), the number of RBFs is taken equal to the number of sampling points:  $M = N + K$ . RBF parameters are set. The unknown coefficients in (3) are found as a solution to a system of linear algebraic equations, which is obtained from the residuals of problem (1) at sampling points after substituting (3) in (1). For this, the RBF must be differentiable as many times as necessary. The result is a system of linear algebraic equations

$$\mathbf{A}\mathbf{w} = \mathbf{b}, \tag{4}$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{G}_L \\ \mathbf{G}_B \end{bmatrix}, \quad \mathbf{G}_L = \begin{bmatrix} L[\varphi_1(\mathbf{x}_1)] & L[\varphi_2(\mathbf{x}_1)] & L[\varphi_3(\mathbf{x}_1)] & \dots & L[\varphi_N(\mathbf{x}_1)] \\ L[\varphi_1(\mathbf{x}_2)] & L[\varphi_2(\mathbf{x}_2)] & L[\varphi_3(\mathbf{x}_2)] & \dots & L[\varphi_N(\mathbf{x}_2)] \\ \dots & \dots & \dots & \dots & \dots \\ L[\varphi_1(\mathbf{x}_N)] & L[\varphi_2(\mathbf{x}_N)] & L[\varphi_3(\mathbf{x}_N)] & \dots & L[\varphi_N(\mathbf{x}_N)] \end{bmatrix},$$

$$\mathbf{G}_B = \begin{bmatrix} B[\varphi_1(\mathbf{x}_{N+1})] & B[\varphi_2(\mathbf{x}_{N+1})] & B[\varphi_3(\mathbf{x}_{N+1})] & \dots & B[\varphi_N(\mathbf{x}_{N+1})] \\ B[\varphi_1(\mathbf{x}_{N+2})] & B[\varphi_2(\mathbf{x}_{N+2})] & B[\varphi_3(\mathbf{x}_{N+2})] & \dots & B[\varphi_N(\mathbf{x}_{N+2})] \\ \dots & \dots & \dots & \dots & \dots \\ B[\varphi_1(\mathbf{x}_M)] & B[\varphi_2(\mathbf{x}_M)] & B[\varphi_3(\mathbf{x}_M)] & \dots & B[\varphi_N(\mathbf{x}_M)] \end{bmatrix},$$

$$\mathbf{b} = [f(\mathbf{x}_1), f(\mathbf{x}_2), \dots, f(\mathbf{x}_{N_1}), g(\mathbf{x}_{N_1+1}), g(\mathbf{x}_{N_1+2}), \dots, g(\mathbf{x}_M)]^T,$$



$$\mathbf{a} = [w(\mathbf{x}_1), w(\mathbf{x}_2), \dots, w(\mathbf{x}_M)]^T.$$

System (4) has a square matrix, and its solution is a weight vector  $\mathbf{w}$ . The Kansa method generates an asymmetric matrix, which makes it difficult to solve the system with a large number of sampling points. With a large number of sampling points, the  $\mathbf{A}$  matrix is poorly conditioned. When using RBF with a global domain of definition, the matrix is dense, which also worsens conditioning. A serious drawback is the unformalized selection of the best RBF parameters.

Known works do not consider the relationship between the number of RBF and the number of sampling points. Usually take the number of sampling points equal to the number of RBF. However, the ratio between the number of RBF  $M$  and the number of sampling points  $N + K: M \propto (N + K)^{\frac{1}{3}}$ , where  $\propto$  means proportionality [20], is known for approximation problems. Since the number of sampling points in this case significantly exceeds the number of RBFs, system (4) is overridden. To solve such systems, the singular value decomposition method is convenient [21].

When solving non-stationary problems, one can use RBF to approximate the differential operator with respect to spatial variables, preserving the differential operators with respect to time (direct method). The result is an ordinary differential equation containing a differential operator approximable by RBF. A simpler method is when the time derivative is replaced by a bed difference and a stationary problem is solved on each time layer using RBF. For example, the equation  $\frac{\partial u}{\partial t} = LU$  after approximating the time derivative takes the form  $\frac{u^k - u^{k-1}}{\tau} = Lu^k$ , where  $\tau$ —the step of sampling time and  $k$ —the time layer number. Then, on the temporary  $k$  layer, the stationary  $\tau Lu^k - u^k = -u^{k-1}$  problem is solved.

Thus, the use of RBF allows you to implement meshless methods and obtain a solution in an approximate analytical form. The resulting solution makes it possible to calculate the solution and its derivatives at arbitrary points in the region. But methods using RBF require solving poorly conditioned systems of linear algebraic equations with dense rectangular matrices. There are no formalized methods for determining the position and parameters of the RFB form. Networks of radial basis functions are free from most of these shortcomings, all parameters of which are determined during the networks learning.

RBFN includes two layers [11]. The first layer consists of RBFs that perform nonlinear transformation of the input vector  $\mathbf{x} = [x_1, x_2, \dots, x_d]$ —the coordinates of the point at which the approximation to the solution is calculated ( $d$ —the dimension of space). The second RBFN layer is a linear weighted adder

$$u(\mathbf{x}) = \sum_{m=1}^M w_m \varphi_m(\mathbf{x}; \mathbf{p}_m), \tag{5}$$

where  $M$ —the number of RBF,  $w_m$ —RBF weight  $\varphi_m$ ,  $\mathbf{p}_m$ —parameter vector.

The process of solving boundary value problems using RBFN was considered using the example of problem (1) defined in the operator form. In the simplest case, it consists of three stages:

1. From the sets  $\Omega$  and  $\partial\Omega$ , choose  $N$  internal and  $K$  boundary sampling points (2) (points at which the error of the solution is controlled). When there is no a priori information about the solution, it is advisable to use random uniform distribution of sampling points in the region and on the boundary of the solution. If there is a priori information about the solution of the problem, you can increase the number of sampling points in those areas in which it is necessary to obtain increased accuracy of the solution. For example, it is advisable to increase the number of sampling points in areas in which a change in the characteristics of the solution is expected.

Since the properties of the solution to the problem are a priori difficult to evaluate, you can first find a rough solution to the problem using the minimum number of sampling points, and then, having determined the areas in which the error functional takes on the greatest value, decide on the number of sampling points and their location. As already noted, the ratio between the number of RBF  $M$  and the number of sampling points  $N + K$  is known. However, when approximating the solutions of boundary value problems using RBFN, this dependence gives an excessive number of sampling points; therefore, it is necessary to select the number of sampling points. An increase in the number of sampling points leads to an increase in the computational complexity of the problem. Periodic random regeneration of a limited number of sampling points, used to prevent network retraining, reduces the number of sampling points.

2. Define the RBFN structure: network type, number of RBF, type RBF, set initial values for the vector of weights and parameter vectors of RBF. There are no definite recommendations for choosing the type of RBF. When solving a second-order PDE, it is necessary to calculate the second derivatives of the network output. Therefore, it is advisable to use the Gaussian function, the domain of definition of which is comparable with the domain of definition of its derivatives, which cannot be said of multiquads, for which there is a large spread of values. Unlimited values of multiquadrics also complicate their use in the uneven distribution of RBF centers. When choosing preliminary values, it is necessary to set the RBF parameters and the weight vector. The methods for choosing the location of the RBF centers are very similar to the methods for selecting sampling points. Centers can be arranged in nodes of a uniform grid or randomly. You can increase the density of RBF in areas where a change in the nature of the solution is expected. You can start the solution with a minimum amount of RBF and add RBF in areas with large error values during learning [12]. When placing RBF centers in the nodes of a uniform grid, it is advisable to set the same preliminary width values for all RBFs. The width values in this case are selected depending on the step size. With a random distribution of the centers, the width can be chosen randomly from a certain interval. The boundaries of the interval can be the same for all RBFs or depend on the distance

between the center of the RBF and the centers of its neighbors. Weights are usually triggered by small random numbers.

3. Perform network learning, i.e., select such values of weights and RBF parameters so that the error functional at the sampling points takes a minimum value. The solution of the boundary value problem (1) on RBFN is an approximation of an unknown solution on the set of sampling points (2). Since the solution at the sampling points is unknown, only minimization of the residuals on the set of sampling points is possible. To construct the functional error, the least squares method is used. The functional error for searching for  $\mathbf{w}$  weights and  $\mathbf{p}$  RBF parameters minimizing discrepancies at sampling points has the form

$$\begin{aligned}
 J(\mathbf{w}, \mathbf{p}) = & \sum_{i=1}^N [Lu_{\text{RBF}}(\mathbf{x}_i; \mathbf{w}, \mathbf{p}) - f(\mathbf{x}_i)]^2 \\
 & + \lambda \sum_{i=N+1}^{N+K} [Bu_{\text{RBF}}(\mathbf{x}_i; \mathbf{w}, \mathbf{p}) - p(\mathbf{x}_i)]^2 \rightarrow \min, \quad (6)
 \end{aligned}$$

where  $\mathbf{x}_i$ —sampling points (2),  $\lambda$ —matched penalty factor,  $u_{\text{RBF}}$ —approximate solution obtained at RBFN (3).

The penalty factor  $\lambda$  ensures the fulfillment of boundary conditions, since in meshless methods the conditions at the boundary are not fixed. As can be seen from (6), the use of RBFN allows us to optimize not only the weights, but also the RBF parameters (in the case of the Gauss function, the coordinates of cents and the width). The functional error (6) may include terms with penalty factors that are also responsible for other conditions for the formulation of the problem, for example, relations at media interfaces.

Learning RBFN networks differs from solving the problem of unconditional optimization of the functional (6). Functional (6) is minimized on a limited set of sampling points. A trained network should have the generalization property, that is, provide a solution with a given accuracy indicator not only at sampling points, but also at arbitrary points in the solution domain. When learning the network, relearning is possible: At sampling points, the accuracy indicator can be small, and at other points, it can be large. The possibility of relearning is reduced by using a large number of sampling points. But, this approach increases the solution time. The way out is periodic random regeneration of a set of sampling points [14]. From the modern point of view on the learning of neural networks, this technique is the implementation of mini-batch (stochastic) learning [22]. When using sampling point regeneration, the RBFN learning process is organized as a process of minimizing a set of functionals error, each of which is obtained by a specific choice of sampling points. Each functional error is not minimized to the end. Between the regeneration of sampling points, only a few steps are taken of the selected method of minimizing the functional error. This approach circumvents the problem of getting into a local extremum, which is typical for most methods of global nonlinear optimization.

The vast majority of RBFN learning algorithms are based on gradient optimization methods [23]. All gradient methods are local optimization methods, which in general does not guarantee the achievement of a global minimum of the functional error. At the same time, the search for the global minimum of the functional error, generally speaking, is not necessary; it is enough to find the local minimum with some given accuracy. There are known applications of genetic algorithms for learning RBFN networks in solving classification problems [24], which are much simpler than PDE solutions. Three classes are distinguished among gradient methods: zero-order methods that use only the values of the optimized function and not the values of its derivatives during optimization, first-order methods that use the first derivatives of the optimized function (function gradient) and second-order methods that use the second derivatives (Hessian matrix).

Methods to minimize the functional error can be divided into two groups. The first group includes methods for sequentially adjusting weights and RBF parameters. The weights that have the greatest impact on the functionality error are tuned first, and then, the RBF parameters are tuned. Since the weights enter linearly into the formula for outputting the network (5), optimization methods other than those used for learning RBF parameters that are nonlinear in (5) can be used for their learning.

In the well-known works devoted to solving PDE on RBFN [9, 10, 12–14], the simplest first-order method is used—the gradient descent method. Let us consider the implementation of the fastest descent method using the example of the two-dimensional problem (1) and the use of Gaussian as RBF. Consider a single parameter vector RBFN

$$\boldsymbol{\theta} = [w_1, w_2, \dots, w_{n_{\text{RBF}}}, c_{11}, c_{21}, \dots, c_{n_{\text{RBF}}1}, c_{12}, c_{22}, \dots, c_{n_{\text{RBF}}2}, a_1, a_2, \dots, a_{n_{\text{RBF}}}]^T, \tag{7}$$

where  $w_j$ —RBF weights,  $j = 1, 2, 3, \dots, n_{\text{RBF}}$ ,  $n_{\text{RBF}}$ —number of RBF,  $c_{j1}$  and  $c_{j2}$ —coordinates of the centers,  $a_j$ —width.

Correction of vector (7) at the iteration  $k$  in the gradient descent method is carried out according to the formula

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} + \Delta\boldsymbol{\theta}^{(k+1)}, \tag{8}$$

where  $\Delta\boldsymbol{\theta}^{(k+1)} = -\eta\nabla J(\boldsymbol{\theta}^{(k)})$ —vector of parameter correction,  $\eta$ —learning speed, selected hyperparameter,  $\nabla J(\boldsymbol{\theta}^{(k)})$ —the gradient vector of functional (6) over the components of the vector  $\boldsymbol{\theta}^{(k)}$  (7) at the iteration  $k$ .

Calculations by (8) end with a small value of functional (6). The gradient descent method has a low convergence rate, which does not allow solving problems with high accuracy.

Second-order methods are based on a quadratic approximation of the functional error. In the vicinity of the next approximation of the parameter vector  $\boldsymbol{\theta}^{(k)}$  of the network, the functional error (6) is approximated by the Taylor formula

$$\begin{aligned}
 J(\boldsymbol{\theta}^{(k)} + \Delta\boldsymbol{\theta}^{(k+1)}) &\approx J(\boldsymbol{\theta}^{(k)}) + [\nabla J(\boldsymbol{\theta}^{(k)})]^T \Delta\boldsymbol{\theta}^{(k+1)} \\
 &\quad + \frac{1}{2} [\Delta\boldsymbol{\theta}^{(k+1)}]^T \mathbf{H}(J(\boldsymbol{\theta}^{(k)})) \Delta\boldsymbol{\theta}^{(k+1)}, \tag{9}
 \end{aligned}$$

where  $\nabla J(\boldsymbol{\theta}^{(k)})$ —functional gradient and  $\mathbf{H}(J(\boldsymbol{\theta}^{(k)}))$ —the Hessian matrix (the matrix of the second derivatives of the functional) calculated with  $\boldsymbol{\theta}^{(k)}$ .

From the minimum condition for functional (9), the network parameter correction vector  $\Delta\boldsymbol{\theta}^{(k+1)}$  can be obtained, which ensures a decrease in the functional error. Due to the complexity of calculating the Hessian matrix for multilayer perceptron, various approximations of the Hessian matrix are used. For example, the conjugate gradient method uses the Fletcher–Reeves formulas [25] and Polak–Ribier [26]. In quasi-Newtonian methods, the Hessian approximation matrix is calculated at each training step, for example, according to the Broyden–Fletcher–Goldfarb–Shanno (BFGS) formula [27]. In the Levenberg–Marquardt method [23], the Hessian matrix is approximated using the product of the Jacobian matrices of the network error vector.

Second-order methods are not widely used in RBFN learning. However, the presence of only one layer with nonlinear functions and the differentiability of most RBFs provide the possibility of applying second-order optimization methods for learning RBFN. In [28], when solving the approximation problem, the nonlinear layer was studied by the conjugate gradient method, and the weights were studied by the method of orthogonal least squares. In [29], an algorithm was proposed for the conjugate gradient adjustment method for RBFN weights, which differs from the known ones taking into account the specifics of solving boundary value problems. RBF parameters were learned by gradient descent method. The algorithm takes into account the differentiability of RBF and is based on the matrix–vector representation of the functional error (6).

In [15], it was proposed, and in [30, 31], a fast RBFN learning algorithm was learned, based on an effective optimization method, the trust region method (TRM) [32]. The method allows to simultaneously optimize a large number of parameters, has a high convergence rate even for poorly conditioned tasks and allows to overcome local minima.

The TRM algorithm is quite complicated, since at least it is found in limited areas, which requires at each step of the optimization process to solve the conditional optimization problem. Therefore, it is advisable to investigate the possibility of adaptation for learning RBFN of modern fast first-order methods and the Levenberg–Marquardt method. Of particular interest is the Levenberg–Marquardt method, which is simpler to implement than TRM and, as shown in [33], is equivalent to TRM.

### 3 Development of Levenberg–Marquardt Algorithm for Learning of Radial Basis Functions Networks for Solving PDE

The implementation of the Levenberg–Marquardt RBFN learning method for PDE solution will be considered on the example of the model problem described by the Laplace equation with Dirichlet boundary condition

$$\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = f(x_1, x_2), \quad (x_1, x_2) \in \Omega, \quad u = p(x_1, x_2), \quad (x_1, x_2) \in \partial\Omega, \quad (10)$$

The functional error for the model problem is the sum of the squared residuals along the internal and boundary sampling points

$$I = \left[ \sum_{i=1}^N (\Delta u_i - f_i)^2 + \lambda \cdot \sum_{j=1}^K (u_j - p_j)^2 \right], \quad (11)$$

where  $\Delta u_i$ —Laplacian at the point  $i$ ,  $r_i = \Delta u_i - f$ —residual of the  $i$ th internal sampling point,  $r_j = u_j - p_j$ —residual at the  $j$ th boundary sampling point.

In the Levenberg–Marquardt method, the correction  $\Delta \theta^{(k)}$  of the parameter vector  $\theta$  (7) is found from the solution of a system of linear algebraic equations

$$(\mathbf{J}_{k-1}^T \mathbf{J}_{k-1} + \mu_k \mathbf{E}) \Delta \theta^{(k)} = -\mathbf{g}_{k-1}, \quad (12)$$

where  $\mathbf{J}_{k-1}^T \mathbf{J}_{k-1} + \mu_k \mathbf{E}$ —an approximation of the Hessian matrix,  $\mathbf{E}$ —identity matrix,  $\mu_k$ —regularization parameter that changes at each step of learning,  $\mathbf{g} = \mathbf{J}^T \mathbf{r}$ —gradient vector of functional (11) according to the vector of  $\theta$  parameters,  $\mathbf{r} = [r_1 \ r_2 \ \dots \ r_n]^T$ —residual vector at internal and boundary sampling points,  $\mathbf{J}_{k-1}$ —Jacobi matrix calculated in  $k - 1$  iteration.

Let us represent the Jacobi matrix in block form  $\mathbf{J} = [\mathbf{J}_w \mid \mathbf{J}_{c_1} \mid \mathbf{J}_{c_2} \mid \mathbf{J}_a]$ , where

$$\mathbf{J}_w = \begin{bmatrix} \frac{\partial r_1}{\partial w_1} & \dots & \frac{\partial r_1}{\partial w_{n\text{RBF}}} \\ \frac{\partial r_2}{\partial w_1} & \dots & \frac{\partial r_2}{\partial w_{n\text{RBF}}} \\ \dots & \dots & \dots \\ \frac{\partial r_n}{\partial w_1} & \dots & \frac{\partial r_n}{\partial w_{n\text{RBF}}} \end{bmatrix}, \quad \mathbf{J}_{c_1} = \begin{bmatrix} \frac{\partial r_1}{\partial c_{11}} & \dots & \frac{\partial r_1}{\partial c_{n\text{RBF}1}} \\ \frac{\partial r_2}{\partial c_{11}} & \dots & \frac{\partial r_2}{\partial c_{n\text{RBF}1}} \\ \dots & \dots & \dots \\ \frac{\partial r_n}{\partial c_{11}} & \dots & \frac{\partial r_n}{\partial c_{n\text{RBF}1}} \end{bmatrix},$$

$$\mathbf{J}_{c_2} = \begin{bmatrix} \frac{\partial r_1}{\partial c_{12}} & \dots & \frac{\partial r_1}{\partial c_{n\text{RBF}2}} \\ \frac{\partial r_2}{\partial c_{12}} & \dots & \frac{\partial r_2}{\partial c_{n\text{RBF}2}} \\ \dots & \dots & \dots \\ \frac{\partial r_n}{\partial c_{12}} & \dots & \frac{\partial r_n}{\partial c_{n\text{RBF}2}} \end{bmatrix}, \quad \mathbf{J}_a = \begin{bmatrix} \frac{\partial e_1}{\partial a_1} & \dots & \frac{\partial e_1}{\partial a_{n\text{RBF}}} \\ \frac{\partial e_2}{\partial a_1} & \dots & \frac{\partial e_2}{\partial a_{n\text{RBF}}} \\ \dots & \dots & \dots \\ \frac{\partial e_n}{\partial a_1} & \dots & \frac{\partial e_n}{\partial a_{n\text{RBF}}} \end{bmatrix},$$

where  $n = N + K$ —total number of sampling points.

The elements of the Jacobi matrix are easy to calculate analytically. Elements of the  $\mathbf{J}_w$  matrix for internal sampling points are calculated by the formula

$$\frac{\partial e_i}{\partial w_j} = \frac{\partial(\Delta v_i - f_i)}{\partial w_j} = e^{-\frac{\|\mathbf{x}-\mathbf{c}_j\|^2}{2a_j^2}} \cdot \frac{\|\mathbf{x} - \mathbf{c}_j\|^2 - 2a_j^2}{a_j^4}.$$

For boundary sampling points, calculations are performed using the  $\frac{\partial e_i}{\partial w_j} = \exp\left(-\frac{\|\mathbf{x}-\mathbf{c}_j\|^2}{2a_j^2}\right)$  formula. The  $\mathbf{J}_{c_1}$  matrix elements for internal sampling points are of the form

$$\frac{\partial e_i}{\partial c_{j1}} = \frac{w_j}{a_j^4} \cdot e^{-\frac{\|\mathbf{x}-\mathbf{c}_j\|^2}{2a_j^2}} \cdot (x_1 - c_{j1}) \cdot \frac{\|\mathbf{x} - \mathbf{c}_j\|^2 - 4a_j^2}{a_j^2}.$$

For boundary points, matrix elements are written as  $\frac{\partial e_i}{\partial c_{j1}} = w_j \cdot e^{-\frac{\|\mathbf{x}-\mathbf{c}_j\|^2}{2a_j^2}} \cdot \frac{(x_1 - c_{j1})}{a_j^2}$ . Similarly, the elements of the  $\mathbf{J}_{c_2}$  matrix are calculated.

The elements of the  $\mathbf{J}_a$  matrix for internal sampling points are of the form

$$\frac{\partial e_i}{\partial a_j} = \frac{w_j}{a_j^5} \cdot e^{-\frac{\|\mathbf{x}-\mathbf{c}_j\|^2}{2a_j^2}} \cdot \left[ \frac{\|\mathbf{x} - \mathbf{c}_j\|^2}{a_j^2} \cdot \left( \|\mathbf{x} - \mathbf{c}_j\|^2 - 2a_j^2 \right) - 4 \cdot \left( \|\mathbf{x} - \mathbf{c}_j\|^2 - a_j^2 \right) \right].$$

For boundary points, matrix elements are written as  $\frac{\partial e_i}{\partial a_j} = w_j \cdot e^{-\frac{\|\mathbf{x}-\mathbf{c}_j\|^2}{2a_j^2}} \cdot \frac{\|\mathbf{x}-\mathbf{c}_j\|^2}{a_j^3}$ .

The condition for completing the learning process by the Levenberg–Marquardt method is a small value of the functional error (11).

The matrix  $\mathbf{J}_{k-1}^T \mathbf{J}_{k-1} + \mu_k \mathbf{E}$  of system (12) is dense symmetric and positive definite. Therefore, to solve system (12), one can use the Cholesky method [21]. A drawback of the Cholesky method is the use of a lengthy square root extraction operation when performing matrix decomposition. The  $\text{LDL}^T$  decomposition method [21] is free from this drawback, which represents the matrix in the form  $\mathbf{A} = \mathbf{LDL}^T$ , where  $\mathbf{L}$ —the lower triangular matrix with the unit main diagonal,  $\mathbf{D}$ —the diagonal matrix and  $T$ —the matrix transpose operation. When decomposing, the square root extraction operation is not applied.

In the Levenberg–Marquardt method, the regularization parameter  $\mu$  must change during the learning of the network. The learning process begins with a relatively large value of the parameter  $\mu$ . This means that at the beginning of the learning process, Hessian in (12) is close to the approximate value  $\mathbf{H} \approx \mu \mathbf{E}$ , and the correction vector is determined by the gradient descent method with a small step  $\Delta \boldsymbol{\theta}^{(k)} = -\mathbf{g}_{k-1} / \mu_k$ . As the functional error decreases, the parameter  $\mu$  decreases and the method approaches the Newton method with the Hessian approximation  $\mathbf{H} \approx \mathbf{J}^T \mathbf{J}$ . This ensures a high

convergence rate, since the Newton method near the minimum of the functional error has good convergence. In [33], it is recommended to start with some value of  $\mu_0$  and coefficient  $\nu > 1$ . The current value of  $\mu$  is divided by  $\nu$  if the functional error is reduced, or multiplied by  $\nu$  if the functional error is increased.

It was shown in [33] that the Levenberg–Marquardt method is equivalent to TRM, and the radius of the trust region is controlled by the parameter  $\mu$ . But unlike the well-known TRM implementations, the Levenberg–Marquardt method does not require solving a rather complicated conditional optimization problem at each learning iteration. That is, the Levenberg–Marquardt method, while maintaining the positive properties of the trust region method, is simpler.

The disadvantage of the Levenberg–Marquardt method is the poor conditionality of system (12), which depends on the RBF width and increases with increasing accuracy of calculations. It is known [34] that the matrix whose elements are RBF is poorly conditioned and the conditionality of the matrix depends on the width of the RBF. As the RBF width increases, the elements of the matrix  $\mathbf{J}_w$  tend to unity, and the elements of the matrices  $\mathbf{J}_c$  and  $\mathbf{J}_a$  tend to zero. The condition number of the matrix  $\mathbf{J}^T \mathbf{J}$  is increasing. The regularization parameter  $\mu$  improves the conditionality of system (12), but a decrease in the parameter  $\mu$  as the error decreases leads to a deterioration in conditionality.

## 4 Experiments

An experimental study was carried out using the example of problem (10) with  $f(x_1, x_2) = \sin(\pi x_1) \cdot \sin(\pi x_2)$ ,  $p(x_1, x_2) = 0$ . The problem was solved in a single square. The number of internal and boundary sampling points is  $N = 100$ ,  $K = 40$ . The penalty factor is  $\lambda = 10$ . The RBF centers were regularly located on a square grid with the number of centers at each coordinate equal to 8. Sampling points were randomly located in the solution region and on the region boundary. Weights were initiated by zero values. The initial width of all RBFs was constant, equal to 0.2. The experiments were carried out in the MATLAB R2019a system. To solve system (12), we used the MATLAB system solver. The RBFN learning by the Levenberg–Marquardt method was compared with the gradient descent learning and the accelerated Nesterov method [35]—the fastest first-order method.

Figure 1 shows the location of the centers, the symbol of the width (in the form of circles with radii equal to the width) of RBF and the weights using the MATLAB color palette before learning the network (Fig. 1a) and after learning (Fig. 1b). Figure 1 shows the importance of setting RBF parameters.

The dependence of the mean square residual of various algorithms on the iteration number is shown in Fig. 2.

The gradient descent method made it possible to solve the model problem with little accuracy. To solve with high accuracy, the method is practically not applicable.



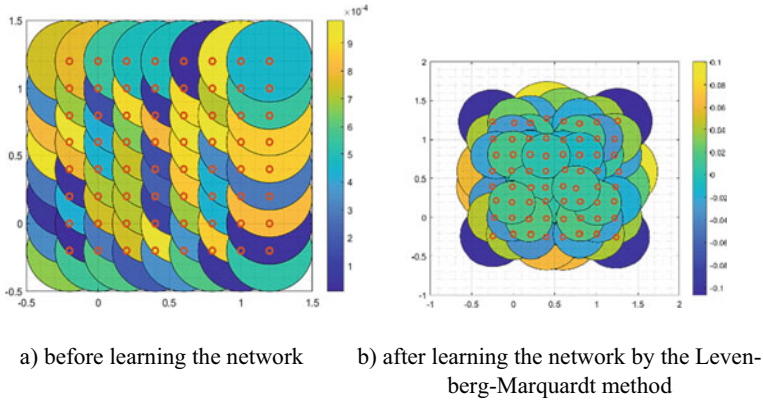


Fig. 1 The centers and width of RB functions in solving PDE

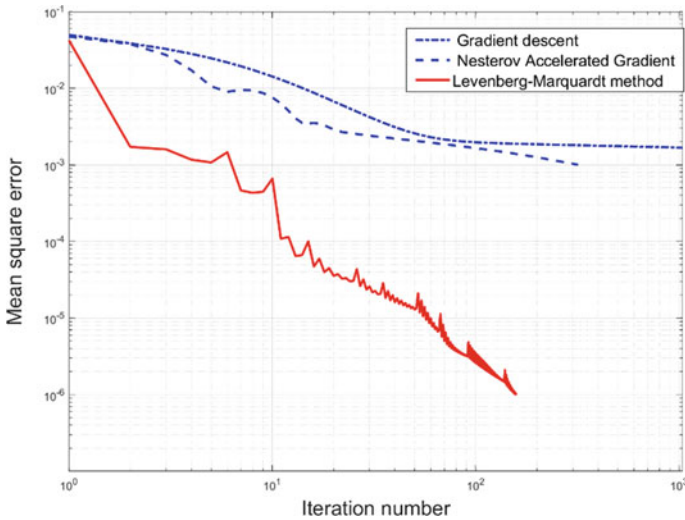


Fig. 2 Dependences of the mean quadratic residual of various algorithms on the iteration number

The Nesterov method provides somewhat greater accuracy. Only the Levenberg–Marquardt method allowed us to solve the problem with high accuracy in an acceptable time. The Levenberg–Marquardt method showed practically the same results compared to the trust region method [15], but the implementation of the Levenberg–Marquardt method is simpler. The disadvantages of the Levenberg–Marquardt method are the poor conditionality of the system that forms the correction of the parameters and the non-smooth nature of the convergence.

Thus, the algorithm of the Levenberg–Marquardt method showed a clear advantage over first-order algorithms and ensured accuracy at the level of known implementations of the trust region algorithm, but is simpler than these algorithms.

## 5 Conclusions

Networks of radial basis functions are a promising means of solving boundary value problems described by partial differential equations. But, the well-known methods of learning networks of radial basis functions do not provide quick learning of networks of radial basis functions. As a way to eliminate this drawback, it is proposed to improve the algorithms for learning networks.

For learning networks of radial basis functions intended for solving PDE, a learning algorithm based on the Levenberg–Marquardt method has been developed, which differs by taking into account the specifics of the network architecture and analytical calculation of parameters. The method made it possible to achieve the average quadratic discrepancy, which is not achievable by the known first-order algorithms, on the model problem. The proposed algorithm achieves a small error for the number of iterations equal to the number of iterations of the algorithm based on the trust region method, but is simpler than this algorithm, since it does not require solving the conditional optimization problem at each iteration.

## References

1. Grieves, M.: Digital twin: manufacturing excellence through virtual factory replication. *Nc-Race* 18 **95**, 6–15 (2014)
2. Madni, A., Madni, C., Lucero, S.: Leveraging digital twin technology in model-based systems engineering. *Systems* **7**(1), 7 (2019)
3. Farlow, S.J.: *Partial Differential Equations for Scientists and Engineers*. Dover Publications, USA (1993)
4. Mazumder, S.: Numerical methods for partial differential equations: finite difference and finite volume methods. In: *Numerical Methods for Partial Differential Equations: Finite Difference and Finite Volume Methods*, pp. 1–461. Elsevier, Amsterdam (2015)
5. Tolstykh, A.I., Shirobokov, D.A.: Meshless method based on radial basis functions. *Comput. Math. Math. Phys.* **45**(8), 1447–1454 (2005)
6. Vasilyev, A., Tarkhov, D., Malykhina, G.: Methods of creating digital twins based on neural network modeling. *Mod. Inf. Technol. IT-Educ.* **14**(3), 521–532 (2018)
7. Griebel, M., Schweitzer, M.A.: *Meshfree Methods for Partial Differential Equations III. Lecture Notes in Computational Science and Engineering* (2007)
8. Buhmann, M.D.: *Radial Basis Functions: Theory and Implementations*. Cambridge University Press, Cambridge (2004)
9. Chen, W., Fu, Z.-J.: *Recent Advances in Radial Basis Function Collocation Methods*. Springer, Berlin (2014)
10. Yadav, N., Yadav, A., Kumar, M.: *An Introduction to Neural Network Methods for Differential Equations*. Springer, Berlin (2015)

11. Aggarwal, C.C.: *Neural Networks and Deep Learning*. Springer International Publishing, Berlin (2018)
12. Jianyu, L., Siwei, L., Yingjian, Q., Yaping, H.: Numerical solution of elliptic partial differential equation by growing radial basis function neural networks. *Neural Netw.* **16**(5–6), 729–734 (2003)
13. Mai-Duy, N.: Solving high order ordinary differential equations with radial basis function networks. *Int. J. Numer. Meth. Eng.* **62**(6), 824–852 (2005)
14. Vasiliev, A.N., Tarkhov, D.A.: *Neural Network Modeling: Principles. Algorithms. Applications*. St. Petersburg Polytechnic University Publishing House (2009)
15. Gorbachenko, V.I., Zhukov, M.V.: Solving boundary value problems of mathematical physics using radial basis function networks. *Comput. Math. Math. Phys.* **57**(1), 145–155 (2017)
16. Gorbachenko, V.I., Lazovskaya, T.V., Tarkhov, D.A., Vasilyev, A.N., Zhukov, M.V.: Neural network technique in some inverse problems of mathematical physics. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9719, pp. 310–316. Springer, Berlin (2016)
17. Fasshauer, G.E., Zhang, J.G.: On choosing “optimal” shape parameters for RBF approximation. *Numer. Algorithms* **45**(1–4), 345–368 (2007)
18. Kansa, E.J.: Multiquadrics—a scattered data approximation scheme with applications to computational fluid-dynamics-I surface approximations and partial derivative estimates. *Comput. Math. Appl.* **19**(8–9), 127–145 (1990)
19. Kansa, E.J.: Multiquadrics—a scattered data approximation scheme with applications to computational fluid-dynamics-II solutions to parabolic, hyperbolic and elliptic partial differential equations. *Comput. Math. Appl.* **19**(8–9), 147–161 (1990)
20. Niyogi, P., Girosi, F.: On the relationship between generalization error, hypothesis complexity, and sample complexity for radial basis functions. *Neural Comput.* **8**(4), 819–842 (1996)
21. Watkins, D.S.: *Fundamentals of Matrix Computations*. Wiley, Hoboken (2005)
22. Goodfellow, I., Bengio, Y., Courville, A.: *Deep Learning*. MIT Press, Cambridge (2016)
23. Gill, P.E., Murray, W., Wright, M.H.: *Practical Optimization*. Emerald Group Publishing, UK (1982)
24. Jia, W., Zhao, D., Shen, T., Su, C., Hu, C., Zhao, Y.: A new optimized GA-RBF neural network algorithm. *Comput. Intell. Neurosci.* **2014**, 1–6, Article ID 982045 (2014)
25. Fletcher, R.: Function minimization by conjugate gradients. *Comput. J.* **7**(2), 149–154 (1964)
26. Polak, E., Ribiere, G.: Note sur la convergence de méthodes de directions conjuguées. *Revue Française d’informatique et de Recherche Opérationnelle. Série Rouge* **3**(16), 35–43 (1969)
27. Nocedal, J., Wright, S.: *Numerical Optimization*. Springer, Berlin (2006)
28. Zhang, L., Li, K., Wang, S.: An improved conjugate gradient algorithm for radial basis function (RBF) networks modelling. In: *Proceedings of the 2012 UKACC International Conference on Control, CONTROL 2012*, pp. 19–23 (2012)
29. Gorbachenko, V.I., Artyukhina, E.V.: Mesh-free methods and their implementation with radial basis neural networks. *Neirokomp’yutory: Razrabotka, Primentnine* **11**, 4–10 (2010) (in Russian)
30. Alqezweeni, M.M., Gorbachenko, V I., Zhukov, M.V., Jaafar, M.S.: Efficient solving of boundary value problems using radial basis function networks learned by trust region method. *Int. J. Math. Math. Sci.* **2018**, 1–4, Article ID 9457578 (2018).
31. Elisov, L.N., Gorbachenko, V.I., Zhukov, M.V.: Learning radial basis function networks with the trust region method for boundary problems. *Autom. Remote Control* **79**(9), 1621–1629 (2018)
32. Conn, A.R., Gould, N.I.M., Toint, P.L.: *Trust Region Methods*. Trust Region Methods. Society for Industrial and Applied Mathematics, USA (2000)
33. Marquardt, D.W.: An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.* **11**(2), 431–441 (1963)
34. Boyd, J.P., Gildersleeve, K.W.: Numerical experiments on the condition number of the interpolation matrices for radial basis functions. *Appl. Numer. Math.* **61**(4), 443–459 (2011)

35. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: 30th International Conference on Machine Learning, ICML 2013, pp. 2176–2184. International Machine Learning Society (IMLS) (2013)

# Predicting Personality from Image Preferences: Tendencies, Models and Implementation



Stanislav Krainikovsky , Mikhail Melnikov , and Roman Samarev 

**Abstract** This paper describes methods of predicting personality traits from image preferences. Methods of feature selection and machine learning were approved and the results of best models are described in the conclusion. Also the novel approach to the preference data analysis was applied: an attempt to discover general patterns in preference data analysis was made, and several hypotheses about tendencies and precision estimation were formulated and tested, using experiment with images from predefined gallery.

**Keywords** Psychometric data · Big Five · Image preferences · Machine learning · Regression

## 1 Introduction

In many areas, related to human resources, it is important to estimate personality traits of candidates. The classic way to do it is passing questionnaires. However, in some cases there is a need to have a preliminary estimation without asking questions in an explicit fashion. Some characteristics like preferences in images, colors, texts, or music may contain information about the personality. A lot of research has been made in this field [1–7]. For example, in [1] a correlation between Big Five (well-known personality model) and different kinds of digital footprints were estimated. In [2] Facebook “likes” and associated objects—posts, pictures, and words—were used to estimate the Big Five using linear regression. Authors showed, that after collecting more than 100 “likes” from particular users, the model prediction became more precise, than the same estimation, made by their friends or colleagues. The correlation coefficients between psychometric values, obtained from questionnaire assessment and model predictions were mostly above 0.4. Other media preferences were also used to estimate personality traits, such as music [8], or textual information from social media [3]. In general, correlation coefficients were not very high across

---

S. Krainikovsky (✉) · M. Melnikov · R. Samarev  
Dotin Inc., Fremont, CA, USA  
e-mail: [nov-ml@dotin.us](mailto:nov-ml@dotin.us)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021  
N. Voinov et al. (eds.), *Proceedings of International Scientific Conference on Telecommunications, Computing and Control*, Smart Innovation, Systems and Technologies 220, [https://doi.org/10.1007/978-981-33-6632-9\\_43](https://doi.org/10.1007/978-981-33-6632-9_43)

491

all of these researches and were in a range between 0.1 and 0.4. This may be the result of several factors. The first factor is the intrapersonal variance of psychometric assessment. Big Five scales have test-retest reliability coefficients of 0.7–0.9 [9], which means there is a fluctuation of accuracy due to this reason. Also, there are a lot of additional factors, such as cultural and geographic, which are not fixed in the experiments but may influence the result. In this case, the dependencies between media preferences and personality traits may be non-linear, and also probably contain a large amount of noise or outliers. Also, as a summarization of all research, one can say that there is no systematic view of how information from media preferences data can be retrieved and analyzed, and is there any general patterns, which can influence the result. Particularly, it is important to answer the following questions:

1. To what degree feature representation of the data can influence the correlation score between predicted values and psychometric data?
2. How many features are necessary to use in the model to get significant correlation scores?
3. To what degree the choice of the model can influence the precision? Are non-linear methods really more accurate, than simple linear regression?
4. What is the “ceiling” of accuracy in terms of correlation, and how much images are required to achieve it?

In the current article, some attempts were made to answer these questions and build the optimal model, using machine learning and feature engineering approach.

## 2 Materials and Methods

In the current study, several psychometric scales were used as a reference data. Big Five psychometric scale (NEO-PI, [10]) was utilized as a widely used standard for estimating personality traits. Big Five [11] is a well-known characteristic of personality with a strong presence of temperament traits. As it follows from the title, Big Five localizes personality as a point in a five-dimensional space with such dimensions as Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Other scales, such as Gardner’s Multiple intelligence [12], and Personal Globe inventory [13] also were included. As a preference data, a result of choosing images from predefined gallery was used.

### 2.1 Participants

The paid online survey on Amazon Mechanical Turk [14] was conducted from January 10, to January 30, 2019. It was available for all Internet users from any country. Anyone who had been registered in the service would be able to participate in the survey. Participants were paid a 3\$ donation. A total 3200 participants were

recruited to participate in this experiment. After refinement and data cleansing, 1400 participants were included in the dataset for analysis. Participants were mostly from USA (48%) and India (35%). Mean age of responders was 31.85. Minimal age was 16, maximal—71. Sex proportion was 526 females and 875 males. All demographic data were self-reported.

## 2.2 *Experimental Design*

All participants passed several blocks of survey:

1. 300 pictures from predefined gallery were exposed on 20 pages. A participant could choose from 1 to 5 pictures from each page. So, for each person, there were from 20 to 100 “liked” pictures.
2. Participants passed several psychological tests: NEO PI inventory [10], Multiple Intelligence assessment test [15], and PGI [13]. All inventories in total contained 148 questions. Answers to these questions were further used to calculate 46 target psychological characteristics.

Patterns of participants’ answers were examined to detect and filter out unscrupulous cases. Also there were 5 validation questions during the survey, used to check the attention (for example, “Press number six”), or provocative ones (for example, “I am traveling trans-Atlantic 5 or more times a month”). The validation test was considered to be passed if at least 3 from 5 answers were correct. 25% of participants with either unfair answers or non-relevant images were excluded from the subsequent analysis. Target psychological characteristics of the rest 1400 people were calculated on the backend using the special models based on authorized keys of inventory. The target parameters included 46 characteristics and consist of Big Five, Gardner’s multiple intelligence and Personal Globe Inventory metrics, including RIASEC (Holland codes [16, 17]).

## 3 Experiments and Results

### 3.1 *Feature Analysis*

So, for a set of liked pictures from each person, two types of features were extracted:

1. “Gallery tags”. Each image from the gallery of free-to-use pictures from the internet was manually tagged by several keywords by psychologist. Tags belonged to one of four groups indicating objects, activity, place, or emotion present in the image. There were 105 tags in total. For each participant’s liked image set, a collection of 105 scores was calculated, and every number

**Table 1** Correlation analysis of features

	Gallery tags	Gallery histograms
Min	0	0
Mean	0.091	0.053
Median	0.072	0.042
Max	0.346	0.264

contained, how often the particular tag appeared. Each score was normalized by the total amount of scores.

2. “Gallery histograms”—for each image a 96-dimensional vector of a color histogram was used. Three blocks from 32 bins per each color dimension (RGB) were calculated.

First of all, an analysis of correlations between psychometric scales and features was made. For all 46 psychometric parameters all features from each category an absolute value of correlation was measured, and the result is presented in Table 1.

### Models comparison

Several multiple regression models were applied to the data. Unlikely to previous research, a multi-regression was used, and different combinations of features were used as an input for training, namely, gallery tags and gallery histograms. In order to compare models to each other, several metrics were calculated. Coefficients, generated by models, were compared with scores from the target set of psychological characteristics obtained from the questionnaire data. Every component of psychometric scale was analyzed separately. The correlation metric were used as a standard for such comparison, and for the reason to compare the result with reliability scores of psychometric scales. First, the gradient boosting regression was applied to the input data (XGBoost algorithm [18]). The best set of parameters was defined separately for each psychometric parameter using randomized search with fivefold cross-validation. Accuracy scores and correlation coefficients shown below were validated on specially selected testing subset of the original set of users. So, training set contained 75% of users, while 25% were reserved for testing. Additional testing was made on the data, obtained two months later after the first survey (294 responders). The following tables contain results from particular feature set and different psychometric scales of Big Five (Table 2).

A linear regression model was also implemented for the same training and testing sets. The following results were obtained (Table 3).

### Feature selection

One of the important question is about the potential of the decreasing the number of features with not sufficient drop of the quality. In [2] authors showed, that they use more than 100 different “likes” in order to achieve the correlation score higher than 0.4. In this paper, an experiment of feature selection was applied. From 105 different tags, only 30 were selected by algorithm of correlation based and univariate linear



**Table 2** Gradient boosting model results

	Gallery tags feature set		Gallery histograms feature set	
	Test 1 correlation	Test 2 correlation	Test 1 correlation	Test 2 correlation
Min	0.048	0.030	0.036	0
Mean	0.282	0.319	0.258	0.222
Median	0.295	0.333	0.273	0.229
Max	0.497	0.521	0.452	0.404

**Table 3** Linear regression model results

	Gallery tags feature set		Gallery histograms feature set	
	Test 1 correlation	Test 2 correlation	Test 1 Correlation	Test 2 correlation
Min	0.067	0.075	0.025	0
Mean	0.264	0.282	0.262	0.216
Median	0.274	0.296	0.275	0.197
Max	0.465	0.485	0.487	0.447

**Table 4** Gradient boosting model results with reduced set of features

	Test 1 correlation	Test 2 correlation
Min	-0.015	-0.002
Mean	0.250	0.293
Median	0.274	0.307
Max	0.447	0.473

regression [19] using scikit-learn library on Python [20], and the same model, based on Gradient boosting, was applied to the gallery tags feature set (Table 4).

**Estimation the dependence between number of samples and model precision**

Another question is about the number of samples, required from person to achieve desirable quality of predictions. Though there were no special experiment with different gallery sizes, the following procedure was made post hoc: several sub-galleries were formed from initial set of 300 pictures. The sub-gallery of 50, 100, 150, 200, and 250 pictures were used in analysis. Other pictures, which were excluded from the sub-gallery, were not considered in the analysis, even if some of them were chosen by participant. The model, based on reduced set of features, was implemented, and the following results were obtained (Table 5).

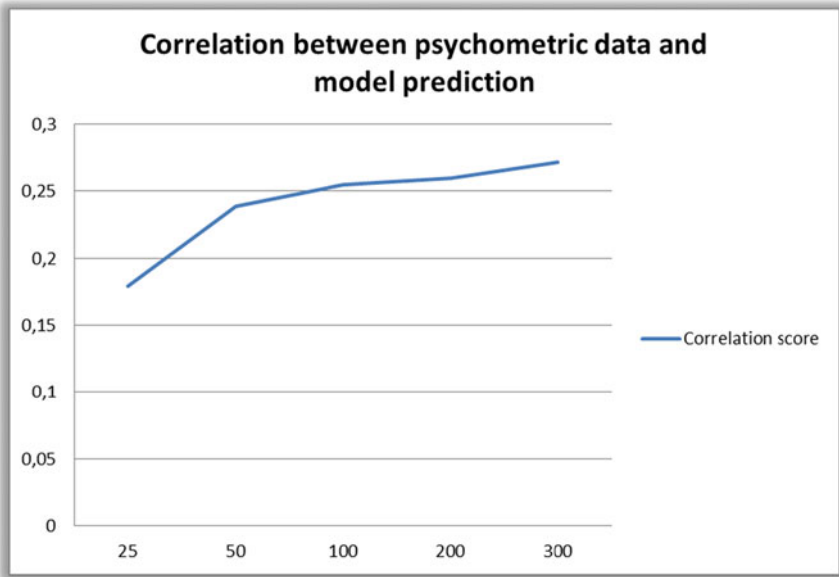
## 4 Discussion

Results of experiments show, that in general, correlation between predicted values and psychometric data corresponds to values of a lot of previous research. The main questions about preference data, which were set in the introduction, will be discussed in this section. First, feature representation of the image data influences the values of correlation, while using the same type of model. According to Table 2, the mean, median and maximal values of correlation scores distribution among 46 psychometric scales differs between semantic features (gallery tags) and average color profiles (gallery histograms). But the difference is not extremely high, so, one can say, different categories of extracted information makes a significant contribution to the result of prediction. Also, according to the Tables 2 and 3, the correlation scores, provided by linear and gradient boosting model are not significantly differ, but while comparing them to the correlation from single feature to psychometric data, we can say, that multiple feature models is better in general, than using single features in the prediction. But in the case of multi-regression, the low difference between linear and non-linear models allows to make a sentence, that the dependencies are not complicated, though have a lot of noise and outliers. In these conditions, such approaches like use complicated models, such as deep learning, can be not applicable for such class of datasets.

Next point is how much features is necessary to achieve certain values of correlation. Though initial use of 300 pictures in the gallery and 105 semantic tags provided 0.29 and 0.33 correlation scores on first and second testing in average, some of values were higher than 0.5. For better understanding of the data, these results should be compared with reduced sets of features and samples. According to Table 4, a set of 30 semantic features, used for the same experiment, provided 0.25 and 0.29 values of correlation—though the drop is not significant. But what is happening, when not only number of describing features, but also a number of chooses, provided by one participant, will decrease? In order to understand the tendency, we should return to the results of [2]. Authors discovered, that there is a direct dependency between number of Facebook “likes”, provided by user, and the correlation with psychometric scales. Our experiment differs in the type of features (using images instead of text) and different experiment conditions (using fixed gallery instead of story of social media footprints). But our hypothesis was that there would be the same dependency, and the presence of “ceiling”, where the trend would achieve in the asymptotic way if the number of samples will increase. So, Table 5 and Fig. 1 demonstrate that the same situation is presented in the conditions of our experiment. The implication is that after achieving more than 300–500 data pieces per participant, the quality of prediction will not change in sufficient way, though our current data are not enough to check this assumption.

**Table 5** Correlation values of Gradient boosting model with different gallery sizes

Gallery size	Test 1 correlation	Test 2 correlation
25	0.162	0.195
50	0.219	0.258
100	0.244	0.263
200	0.249	0.270
300	0.250	0.293



**Fig. 1** Correlation values of the gradient boosting model of 30 features and different gallery sizes (averaged of first and second tests)

## 5 Conclusion

The main result of this work is that using image preferences from predefined gallery allows us to estimate correlations between liked images and psychometric data, obtained by questionnaires. Corresponding correlation values depend on feature representation of the data, though to small degree as well as the result of linear multi-regression, and non-linear models, such as gradient boosting. None of forms of the experiment (feature set and model configuration) allowed to achieve correlation values higher, than 0.55, even while using optimization. This can be the base of the proposal about “ceiling” in the quality of the prediction in the interval 0.6–0.8 of correlation (the higher border is restricted by a mean value of psychometric scale reliability coefficients). And the hypothesis is this ceiling cannot be overcome by

collecting more than several hundred data pieces from each person. However, the appropriate feature selection can help to achieve correlation coefficients higher than 0.4 to some psychometric scales using only 30 features and less than 100 samples. These estimations may be of practical use while engineering real applications and planning the research. Our experimental design and dataset had sufficient limitations, so, these hypotheses should be validated and refined in the future experiments.

## References

1. Azucar, D., Marengo, D., Settanni, M.: Predicting the Big 5 personality traits from digital footprints on social media: a meta-analysis. *Personality Individ. Differ.* **124**, 150–159 (2018)
2. Youyou, W., Kosinski, M., Stillwell, D.: Computers judge personalities better than humans. *Proc. Natl. Acad. Sci.* **112**(4), 1036–1040 (2015)
3. Cutler, A., Kulis, B.: Inferring human traits from Facebook statuses. In: Staab, S., Koltsova, O., Ignatov, D. (eds.) *SocInfo 2018, LNCS*, vol. 11185, pp. 167–195. Springer, Cham (2018)
4. Farnadi, G., Sitaraman, G., Sushmita, S.: Computational personality recognition in social media. *User Model. User-Adap. Inter.* **26**, 2–3 (2016)
5. Hinds, J., Joinson, A.: Human and computer personality prediction from digital footprints. *Curr. Dir. Psychol. Sci.* **28**(2), 204–211 (2019)
6. Segalin, C., Lepri, B., Cristani, M., et al.: What your Facebook profile picture reveals about your personality. In: *ACM MM 2017, Proceedings of the 25th ACM International Conference on Multimedia*, pp. 460–468 (2017)
7. Tandra, T., Hendro, Suhartono, D.: Personality prediction system from Facebook users. *Procedia Comput. Sci.* **116**, 604–611 (2017)
8. Nave, G., Kosinski, M., Stillwell, D., et al.: Musical Preferences Predict Personality: Evidence from Active Listening and Facebook Likes. *Psychological Science* (2018)
9. Gnamb, T.: A meta-analysis of dependability coefficients (test-retest reliabilities) for measures of the Big Five. *J. Res. Personal.* **52**, 20–28 (2014)
10. Costa, P., McCrae, R.: The revised NEO personality inventory (NEO-PI-R). *The SAGE Handbook of Personality Theory and Assessment*, vol. 2, pp. 179–198 (2008)
11. John, O.P., Srivastava, S.: The Big-Five trait taxonomy: history, measurement, and theoretical perspectives. In: Pervin, L.A., John, O.P. (eds.) *Handbook of Personality: Theory and Research*, vol. 2, pp. 102–138. Guilford Press, New York (1999)
12. Gardner, H.: *Intelligence Reframed: Multiple Intelligences For The 21st Century*. Basic Books, New York (1999)
13. Tracey, T.J.G.: Personal globe inventory: measurement of the spherical model of interests and competence beliefs. *J. Vocat. Behav.* **60**, 113–172 (2002)
14. Amazon Mechanical Turk Homepage, <https://www.mturk.com>. Last accessed 2019/11/01
15. Sree Nidhi, S.K., Tay, C.: Multiple intelligence assessment based on Howard Gardner’s research. *Mr. Sreenidhi S.K., Ms. Tay Chinyi Helena.* **7**, 203–213 (2017)
16. Holland, J.L.: A study of measured personality variables and their behavioral correlates as seen in oil paintings. Unpublished doctoral dissertation, University of Minnesota, Minneapolis (1952)
17. Nauta, M.M.: The development, evolution, and status of holland’s theory of vocational personalities: reflections and future directions for counseling psychology. *J. Couns. Psychol.* **57**(1), 11–22 (2010)
18. XGBoost Homepage, <https://xgboost.readthedocs.io>, last accessed 2019/11/01
19. Hall, M.A.: Correlation-Based Feature Selection for Machine Learning, Ph.D. diss. Department of Computer Science, Waikato Univ (1998)
20. Scikit learn competition library Homepage, <http://scikit-learn.org>. Last accessed 2019/11/01

# Power Consumption Meter for Energy Monitoring and Debugging



Nikita Kulikov , Elena Yaitskaya , Arina Shvedova ,  
and Vladimir Zhalnin 

**Abstract** This paper presents the results of power consumption measurement research. The main purpose of the research is to develop a low-cost power monitoring device allowing detection of short-time power surges. Power meter should provide high accuracy of measurements along with portability and simplicity of use. This device helps to detect software bugs, caused by power anomalies in electronic equipment functioning. Fixing power bugs keeps application competitive on the market. The structure and electronic design of power meter is presented. The device prototype has been assembled and programmed using purchased components and original software. The testing method has been developed and applied for power meter prototype. The test results have been analyzed and modifications for the next power meter generation.

**Keywords** Power · UART · ADC · Microcontroller · Power anomaly · Energy · Data transmission · Frequency · Throughput · Software · Power supply · Oscillator · Single channel · Amplifier · Data sending protocol · Firmware · Low cost · Digital · Diagnosing · Configuring

## 1 Introduction

More and more devices are becoming wireless that means significant increase of independent power supply usage. One of the ways to achieve, the best functionality/size ratio is to optimize power consumption by finding expensive energy part of software or application [1, 2].

The other way to increase energy efficiency is to use application with lesser supply requirements. For example, of two applications providing similar features, it is more reasonable to choose the one with lower power consumption.

The aim of the study is to develop an economical prototype that solves the problem of enormous changes in power. Such developments are important for the

---

N. Kulikov (✉) · E. Yaitskaya · A. Shvedova · V. Zhalnin  
Bauman Moscow State Technical University, Moscow, Russia  
e-mail: [nikita@kulikof.ru](mailto:nikita@kulikof.ru)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021  
N. Voinov et al. (eds.), *Proceedings of International Scientific Conference on Telecommunications, Computing and Control, Smart Innovation, Systems and Technologies* 220, [https://doi.org/10.1007/978-981-33-6632-9\\_44](https://doi.org/10.1007/978-981-33-6632-9_44)

499

modern market. Resolution of this problem helps accelerate the development of the electronics industry.

## 2 Background

### 2.1 Method Selection

Circuit power consumption of electronic equipment consists of dynamic and leakage consumptions. Leakage consumption can be reduced by suppling voltage scaling, clock gating or using leakage controlling transistors [3]. Data transmission does not have a tangible impact on leakage. Therefore, the research is focused on dynamic consumptions.

Dynamic part can be calculated by Formula (1)

$$P_d = \frac{1}{2} f_C v_{DD} \sum_{i=1}^N a_i C_i V_{swi} \quad (1)$$

$a_i$ —switching activity

$C_i$ —capacitance

$V_{swi}$ — voltage swing.

Switching activity is a correlation between dynamic power consumption and amount of data sent. It is expressed as probability of transitions of logical components:

$$a = \lim_{T \rightarrow \infty} \frac{n_x(T)}{T \cdot f_C} = D(x)T_C \quad (2)$$

$n_x(T)$ —number of logic gate transitions in  $(-\frac{T}{2}, \frac{T}{2}]$  time interval,

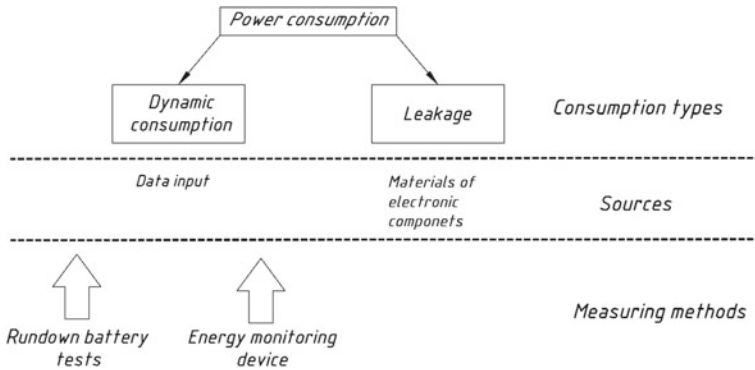
$(x)$ —transition density.

Transition density dependency of input data can be delivered as follows

$$D(y_i) = \sum_{k=1}^m G_k(D(x_{ik})), \quad (3)$$

where  $m$  is a number of inputs in an ideal circuit [4–9]. Figure 1 shows power consumption parts and measuring methods.

The most widely used method for energy cost measurements of applications and software are battery rundown tests. Usually, each test takes a lot of time and does not show any results precisely. That means that these tests should be run several times and debugging might last for months.



**Fig. 1** Power consumption types, sources and measurement methods

Similar research undertaken by many largest IT companies showed that power consumption meter characteristics as following:

- Accurate (resulted from taking measurements of small data samples transmission power);
- Portable (for easy measurements of different sub-circuits of the same electronic device);
- Non-destructive (highly recommended due to decreasing cost of consumables as wires, solder, etc.).

The results of the research had been quite acceptable apart from a very high cost of components. Therefore, another important criteria is the final cost of the developing device [10].

The research is devoted to the development of power meter using minimal economic and time resources. The main purpose of this device is to measure voltage output from any circuit node of electronic equipment and convert it into power values to detect software power bugs. Removing these bugs can dramatically increase battery full discharge time, as well as secure power advantages of application at market.

The high cost and complexity of integration are the main problems of automated power measurement devices. In the first studies, more than \$1500 were spent on prototypes, let alone the fact that their setup requires a lot of time, special software and high costs. Methods of power bugs detection are shown in Fig. 2 [11–14].

However, the market requires a cheap and affordable invention for testing to improve the quality of electronic equipment. The high cost makes it impossible to conduct many parallel measurements at different enterprises, since not all market participants are ready to invest in improving product quality [15]. Hence, it makes sense to use cheap components and accessible technologies. This study uses high-frequency ADCs and a Cortex microcontroller.

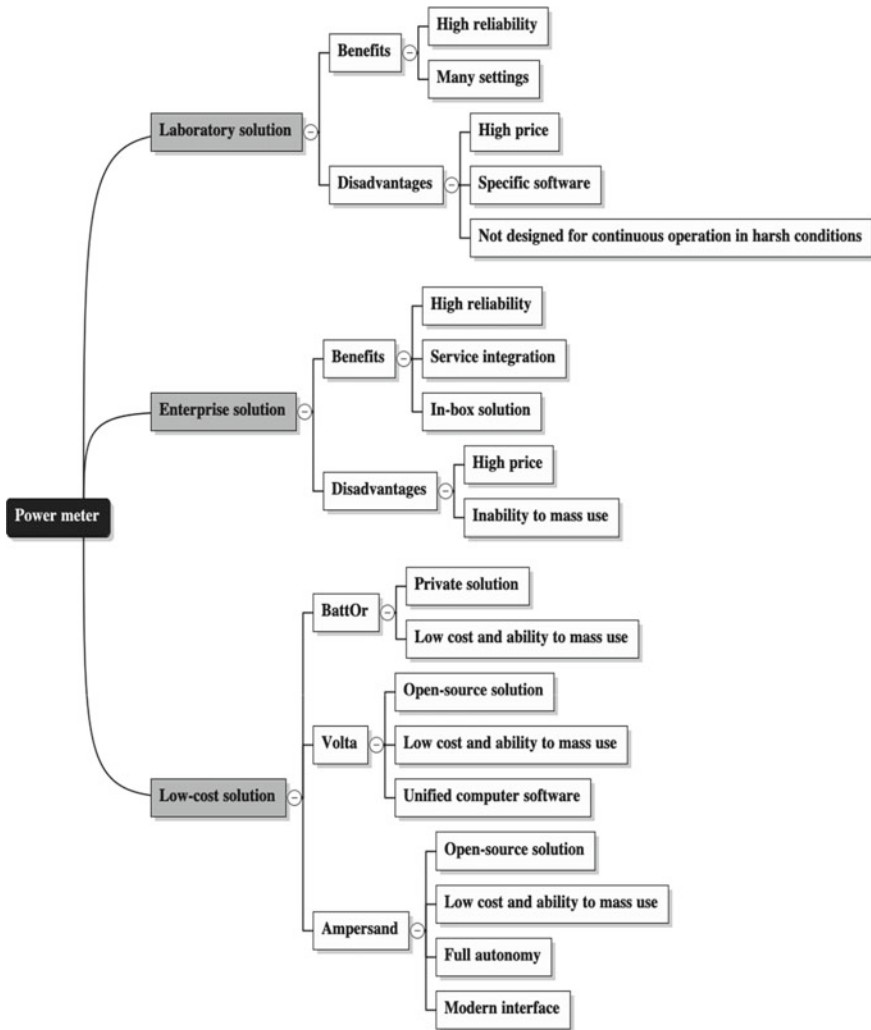


Fig. 2 Power measurement solutions

## 2.2 Data Collection

The main competitors for precise power consumption measurements include several leading companies, among them are IT-titans like Google, Yandex, Facebook. This interest can be explained by the need for power anomaly diagnostic in income generating projects such as Chromium, YouTube, Yandex.Browser. A higher sampling rate allows to examine the energy cost of minimal data transmission unit, for example data frame. Using one of the early power meter prototype developed by Google power draining anomaly have been found and resolved, frame rendering for Chromium have



been optimized. Google published its research in 2011, and further steps in energy efficiency development have been classified.

The current developments of power diagnostic system undertaken by Yandex have revealed some shortcomings:

- No full automatization;
- 2A high currents only;
- No support for modern connector types (USB Type-C).

The basic criteria remains the same like for the earlier developments such as low cost of both device components and manufacturing, and possibility of numerous measurements, as well as high speed of sampling [16].

### 3 Results

#### 3.1 Development

The main criteria for power monitoring device have been formulated as follows:

- USB Type-C connector for voltage supply;
- Independent power supply unit with galvanic isolation;
- Possibility to work with currents overcomes 2A value.

The main purpose of the energy monitor is to detect power debugs in programs reading measurements of power consumption. To achieve the best accuracy in detection, sample rate of analog-to-digital converter (ADC) must be the highest. Figure 3 shows electronic scheme of power consumption meter [17, 18].

Since there is no galvanic isolation between load circuit and controller, we can use embedded analog-to-digital converter, provided by STM32F051R8T6. STM32 controllers offer internal 12-bit ADC connected to APB bus. STM32 ADC operates

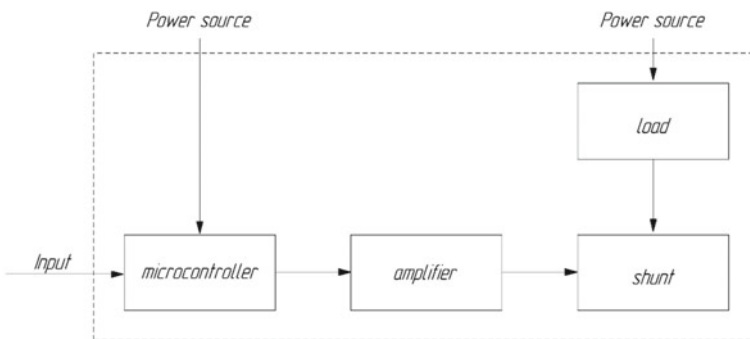


Fig. 3 Power consumption meter electronic scheme

at 14 MHz frequency. It can carry measurements from up to sixteen external sources and three internal. The measurements are stored in 16-bit register.

Best timing for internal ADC is around one microsecond that gives maximum one million samples per second (1 MS/s) or 1 MHz sampling speed. Some versions of STM32 can offer dual fast interleaved mode. While this mode is set, each conversion takes 7 clock cycles without data overlapping. Hence, recommender ADC frequency is 14 MHz sampling speed can be calculated as  $14 \text{ MHz}/7 = 2 \text{ MS/s}$ .

In-built ADC can provide significant sampling rate of around 2 MS/s. Therefore, measurement accuracy is limited only by UART throughput. Single-channel continuous conversion mode is installed on internal ADC of STM32F051R8T6.

The central unit in our device is STM32F051R8T6 controller. To load software, an in-circuit debugger and programmer ST-Link was used. Connectors for ST-Link are marked as XP3. USB Type-C connector serves for power supply. To stabilize supply voltage at the required 3.3 V voltage regulator, LM1084IT-3.3/NOPB is installed at USB power input. Apart from establishing a proper voltage value, power supply circuit needs some filtering capacitors (C1, C2, C5). USART circuit is separated from voltage supply with galvanic isolation to decrease interference and ensure ground separation [19]. The isolation is provided by MAX14850ASE + microchip, designed specially for USART. However, galvanic isolation is not enough for stable data transmission between power monitoring device and a computer. Both PC and the device should use the same frequency. To provide equal frequencies with only internal oscillator of STM32 controllers is not possible, so an external crystal oscillator was added to the circuit. Components C3 and C4 act as load capacitors.

Pins LOAD + and LOAD- are used for connecting power monitoring device and a circuit to be measured or load circuit (Fig. 4). Currents from load circuit flow through current-sensitive resistor R2 changing voltage drop on it. Voltage values from current-sensitive resistor must be increased with an operational amplifier in order to detect the slightest changes of the signal. Resistor R2 is connected to operational amplifier IC1 (+) and (-) inputs [20]. Variable resistor R3 purpose is to adjust output voltage of IC1. Voltage value from R2 being amplified by IC1 arrives at ADC inputs of STM32 controller U2 [21, 22].

An inexpensive solution was found in the high-performance microcontroller STM32 series. LM1084IT-3.3/NOPB was selected as one of the cheapest low voltage drop positive voltage regulator supporting 3.3 V and 5A current (Fig. 5). As a reliable element, the cheap single-pole AD623 power amplifier was chosen (Fig. 4). Yandex developers used this component in their prototype. The inclusion of microcontroller STM32 is shown in Fig. 6.

Description of infinite loop of the microcontroller software is shown in Fig. 7.

A data type *bool* equal to unit *8\_t* was introduced for the program to work, as well as constants TRUE and FALSE, which are 1 and 0, respectively. Using this interrupt algorithm, the variable “allowSend”, which is FALSE by default, is switched to TRUE, as shown in Fig. 8.

The pointer to “receiveByte” is created when configuring UART. This is done so that we can record there value of byte received before interrupt call. This mechanism was developed to form an optimal measurement data sending protocol. To indicate

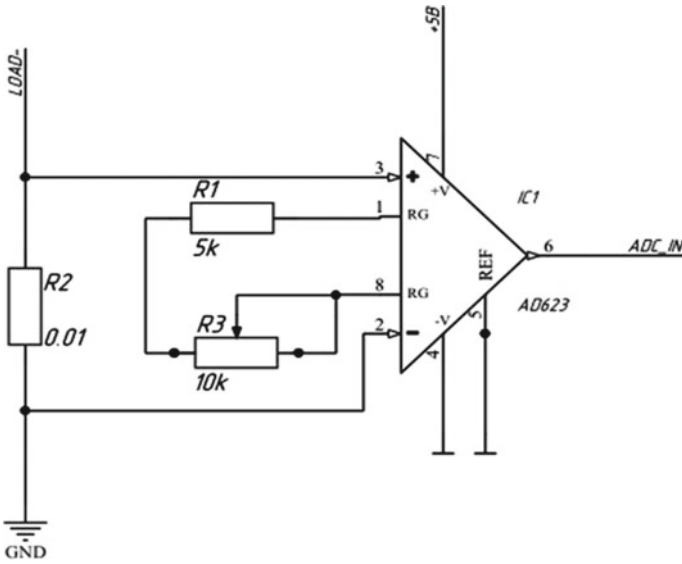


Fig. 4 Load inputs

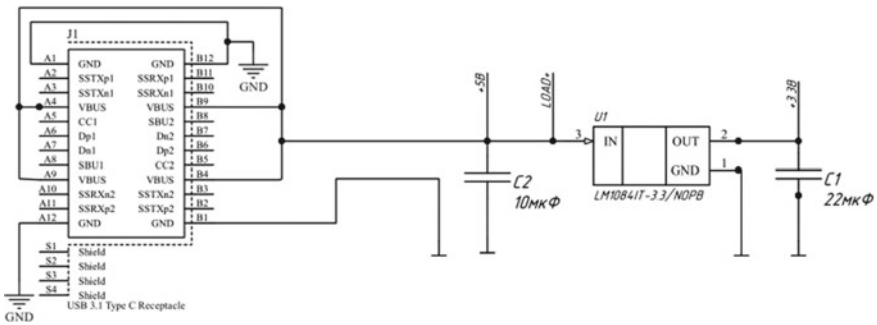


Fig. 5 Voltage regulator

the beginning of data reading, you need a reference word; further reading of data occurs in accordance with a predetermined indent (2 bytes—one value).

Software was developed using Java, as well as a special structure ListWithTimeLimit inherited from LinkedList. This structure’s purpose is to keep a certain amount of values, measured during time period defined by timeLimit variable. Thus, the difference between the first and the last saved values is lesser than timeLimit [23–25].

Power consumption curves were plotted with XChartlibrary. For establishing data transmission by USART port, JSerialComm library was used.

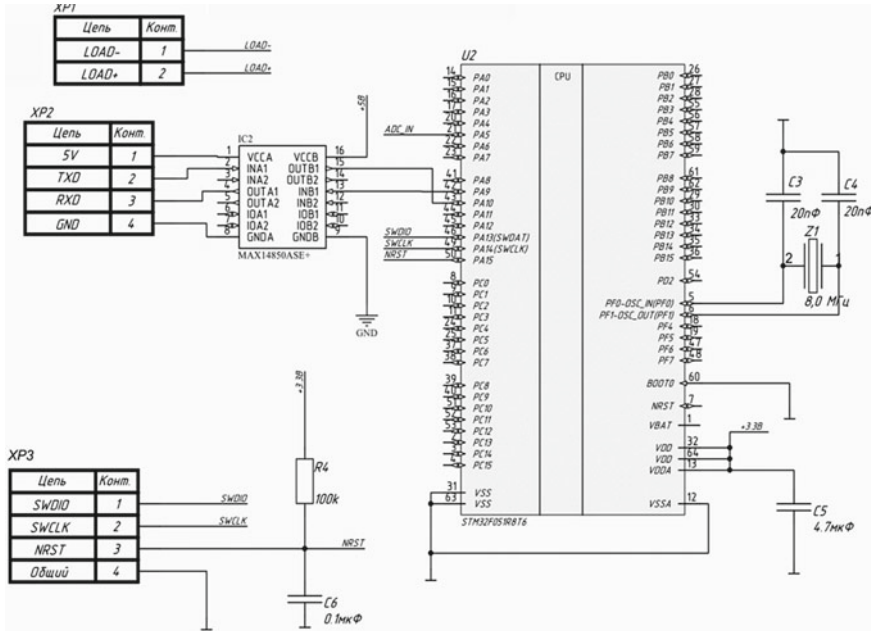


Fig. 6 Microcontroller STM32 activation

### 3.2 Experiment Results

Extra configured port is one of the most frequent power bugs of STM32 controllers. Therefore, the power meter was tested on STM32 microcontroller extra port anomaly. For test, STM32F051R8T6 controller was taken. First, several additional ports were configured and clock signal was supplied for them. This is a waste of energy to be detected. Then, these ports were powered of and second measurement was taken. Figure 8 illustrates experiment results. Comparing two plots, it can be concluded that the voltage signal of the first measurement rises earlier than the signal of the second test and stays high longer. Thus, power consumption of the first experiment is higher than of the second experiment. Power consumption meter has been tested by voltage signal. Results are presented in Fig. 9.

The *x*-axis represents the time in milliseconds, and the *y*-axis shows the voltage on the shunt in turn power can be calculated by a simple formula:

$$P = \frac{u^2}{R} \tag{4}$$

The calculations are not difficult to transfer to the microcontroller since the most important is the protocol and the voltage relieving mechanism. High sampling rate allows to identify the problem spot accurately, therefore can be understood when

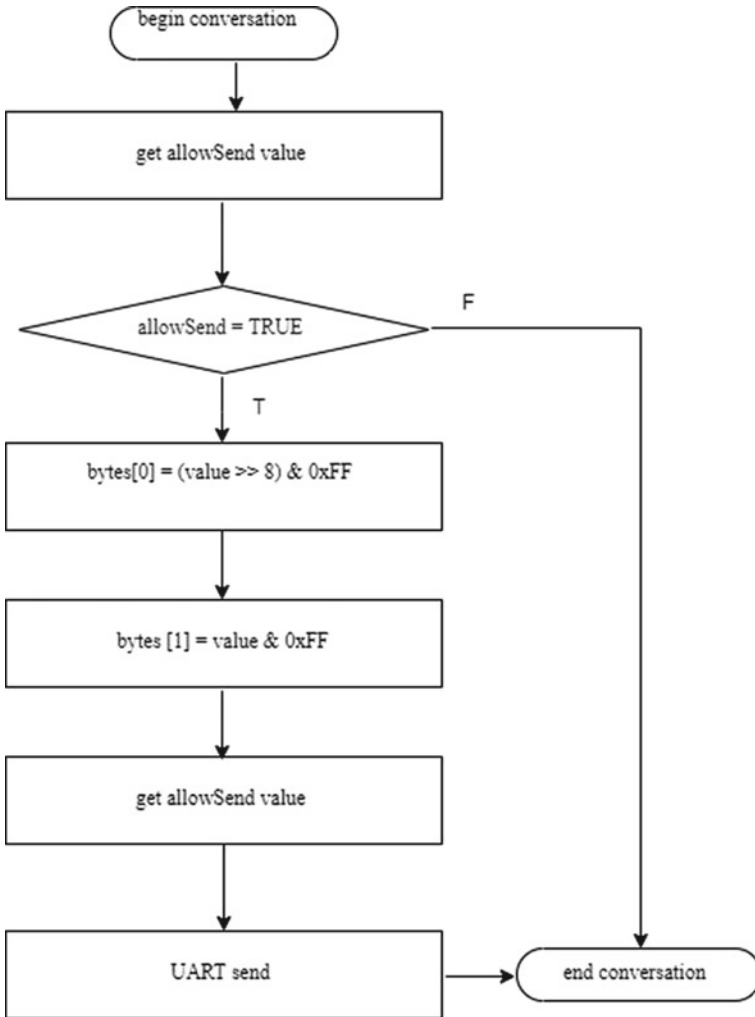


Fig. 7 Algorithm of power consumption measurements

increased consumption began. For example, on the STM32 microcontroller, it is possible to diagnose an erroneous inclusion of port clocking by this way.

### 4 Discussion

This work aims at helping not only specialists in the development of devices with increased energy consumption requirements, but also ordinary developers. It reduces

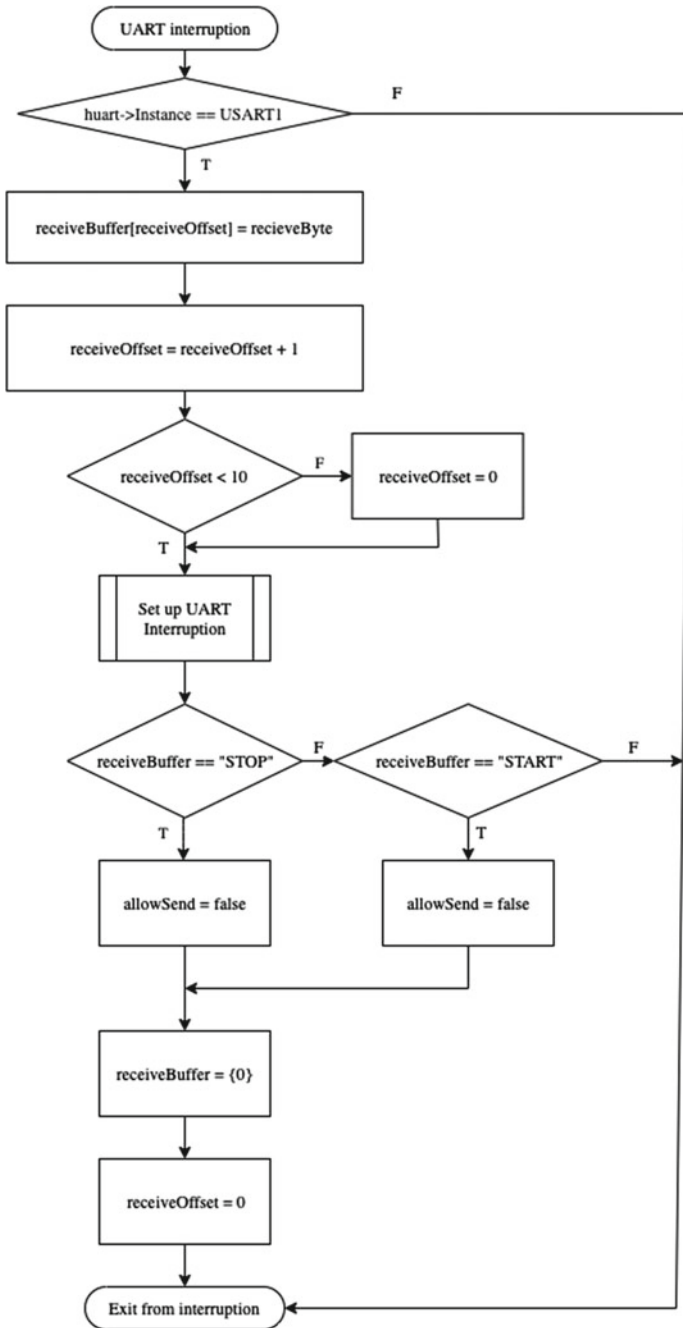
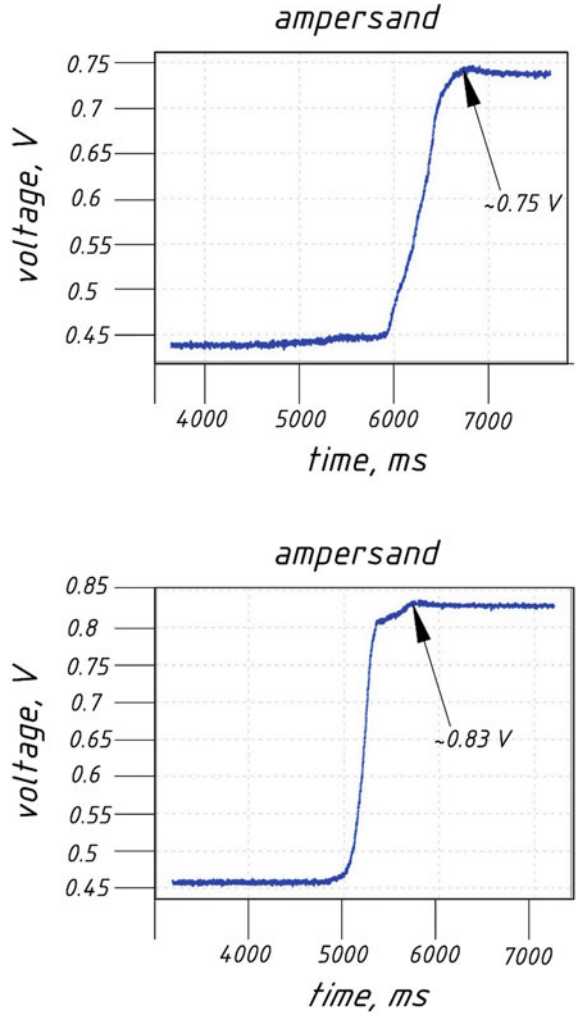


Fig. 8 Interrupt algorithm

Fig. 9 Testing results



the cost of the process of diagnosing anomalies in the firmware. This leads to leapfrog market development of quality software.

The devices can dislodge developers of laboratory equipment which more than \$1000 and create a new approach to researching low-cost solutions.

The cost of these elaborations allows the integration of the solution into IoT devices [26]. Also, integration with the Wi-Fi module and transmitting readings via the Internet is useful when using the fully offline mode. Byte protocols such as "protobuf" are successfully used in robotics and web services development for data transfer and can significantly increase data transmission efficiency of the power meter. Further research will be focused on the improvements in the field of digital gain adjustment, which can be done on the basis of transistors.

## 5 Conclusion

In this paper, the power consumption meter is developed. The device measures voltage values from any circuit node and converts them into power units. The purpose of the power meter is to monitor energy consumption and help software developers to detect power bugs. Suggested device provides sampling rate of about 9600 bps, high measurement accuracy (12-bit), extended functionalities in comparison to earlier open-source research. Power meter is successfully applied for extra configured port power anomaly detection.

Power consumption meter is a low-cost device (total cost is approximately \$30), which can reveal energy inefficiency of software and applications. Further upgrades are to improve data transmission between power meter and PC, thereby significantly increasing data sampling per second. The possible ways of improvement are data transmission by the Internet, either wire and wireless, and integration into the Internet-of-Things.

**Acknowledgements** Separate results of this research were obtained within the framework of RFFI Grant N. 19-07-00463.

## References

1. Carroll, A., Heiser, G.: An analysis of power consumption in a smartphone. USENIX Annual Technical Conference (2010)
2. Vinnal, T., Janson, K., Kalda, H.: Analysis of power consumption and losses in relation to supply voltage quality. In: 2009 13th European Conference on Power Electronics and Applications, EPE '09 (2009)
3. Lorenzo, R., Chaudhary, S.: Low leakage and minimum energy consumption in CMOS logic circuits. In: 2015 International Conference on Electronic Design, Computer Networks and Automated Verification, EDCAV 2015 (2015). <https://doi.org/10.1109/EDCAV.2015.7060536>
4. Sichani, A.S., Moreno, W.A.: Mathematical model for glitch power consumption to study its implication on power analysis attacks (2018). <https://doi.org/10.1109/lascas.2018.8554090>
5. Yuldashev, M.N., Vlasov, A.I.: Mathematical model of the general problem of state classification in wireless sensor networks. IOP Conf. Ser.: Mater. Sci. Eng. (2019). <https://doi.org/10.1088/1757-899X/498/1/012002>
6. Bircher, W.L., John, L.K.: Analysis of dynamic power management on multi-core processors. Proc. Int. Conf. Supercomput. (2008). <https://doi.org/10.1145/1375527.1375575>
7. Martinez, B., Montón, M., Vilajosana, I., Prades, J.D.: The power of models: modeling power consumption for IoT devices. IEEE Sens. J. (2015). <https://doi.org/10.1109/JSEN.2015.2445094>
8. Boicea, V.A.: Energy storage technologies: the past and the present. Proc. IEEE (2014). <https://doi.org/10.1109/JPROC.2014.2359545>
9. Najm, F.N.: Transition density: a new measure of activity in digital circuits. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. (1993). <https://doi.org/10.1109/43.205010>
10. Schulman, A., Schmid, T., Dutta, P., Spring, N.: Demo: phone power monitoring with BattOr. MobiCom 2011 (2011)
11. Balasubramanian, N., Balasubramanian, A., Venkataramani, A.: Energy consumption in mobile phones: a measurement study and implications for network applications. In: Proceedings of



- the ACM SIGCOMM Internet measurement conference, IMC (2009). <https://doi.org/10.1145/1644893.1644927>
12. Watteyne, T., Vilajosana, X., Kerkez, B., Chraim, F., Weekly, K., Wang, Q., Glaser, S., Pister, K.: OpenWSN: A standards-based low-power wireless development environment. *Eur. Trans. Telecommun.* (2012). <https://doi.org/10.1002/ett.2558>
  13. Tutuncuoglu, K., Yener, A.: Optimum transmission policies for battery limited energy harvesting nodes. *IEEE Trans. Wireless Commun.* (2012). <https://doi.org/10.1109/TWC.2012.012412.110805>
  14. Constandache, I., Bao, X., Azizyan, M., Choudhury, R.R.: Did you see Bob?: Human localization using mobile phones. In: Proceedings of the Annual International Conference on Mobile Computing and Networking, MOBICOM (2010). <https://doi.org/10.1145/1859995.1860013>
  15. Snowdon, D.C., Sueur, E.L., Petters, S.M., Heiser, G.: Koala a platform for OS-level power management. In: Proceedings of the 4th ACM European Conference on Computer Systems, EuroSys'09 (2009). <https://doi.org/10.1145/1519065.1519097>
  16. Todorov, G.N., Volkova, E.E., Vlasov, A.I., Nikitina, N.I.: Modeling energy-efficient consumption at industrial enterprises. *Int. J. Energy Econ. Policy* (2019). <https://doi.org/10.32479/ijeeep.7376>
  17. Frenkil, J.: Multi-level approach to low-power IC design. *IEEE Spectr.* **10**(1109/6), 648684 (1998)
  18. Rabaey, J., Chandrakasan, A., Nikolić, B.: Digital integrated circuits: a design perspective (2003). Retrieved from <http://online.sfsu.edu/mahmoodi/engr890/handouts/lecture8.pdf>
  19. High Voltage Power Monitor | Purchase.: (n.d.) Retrieved 12 Dec 2019, from <https://www.msoon.com/online-store>
  20. Sagahyoon, A.: Power consumption in handheld computers. In: IEEE Asia-Pacific Conference on Circuits and Systems, Proceedings, APCCAS (2006). <https://doi.org/10.1109/APCCAS.2006.342129>
  21. RM0091 Reference manual. (2017). Retrieved from <http://infocenter.arm.com>
  22. Zhenzhong, Y., Applications, C.K.-M. & I., & 2016, undefined. (n.d.). Design of robot simulated teaching box based on STM32. *En.Cnki.Com.Cn*. Retrieved from [http://en.cnki.com.cn/Article\\_en/CJFDTotal-WXJY201623012.htm](http://en.cnki.com.cn/Article_en/CJFDTotal-WXJY201623012.htm)
  23. Vlasov, A., on, M. Y.-2019 I. C., & 2019, undefined. (n.d.). Performance Analysis of Algorithms for Energy-Efficient Data Transfer in Wireless Sensor Networks. *Ieeexplore.Ieee.Org*. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8743087/>
  24. Energy-efficient algorithm for classification of states of wireless sensor network using machine learning methods. (n.d.). <https://doi.org/10.1088/1742-6596/1015/3/032153>
  25. Vlasov, A.I., Muraviev, K.A., Prudius, A.A., Uzenkov, D.A.: Load balancing in big data processing systems. *Int. Rev. Autom. Control* **12**(1), 42–47 (2019). <https://doi.org/10.15866/ireaco.v12i1.16808>
  26. Vlasov, AI, ... O. B.-2018 G. S., & 2018, undefined. (n.d.). Technological Platform for Innovative Social Infrastructure Development on Basis of Smart Machines and Principles of Internet of Things. *Ieeexplore.Ieee.Org*. Retrieved from <https://ieeexplore.ieee.org/abstract/document/8570062/>

# LoRaWAN Gateway Coverage Evaluation for Smart City Applications



Vadim Shpenst and Andrei Terleev

**Abstract** The meaning of the term smart city has undergone changes over the last decades. However, smart city services nowadays are gaining popularity worldwide. Russia is not an exception. Since early 2000s smart city market is growing steadily and, according to forecasts, will continue to grow. Today, the possible applications of smart city services cover a wide range of sectors—from energy production, distribution, and consumption to sustainable mobility and waste management. All the innovative services require millions of monitoring sensors and control devices to be connected between each other and to a management platform. Hence, new types of wireless communication networks that meet the specific requirements of smart city services are needed. LoRaWAN is the most promising (machine-to-machine (M2M) communication technology among other LPWANs such as NB-IoT and Sigfox. Conducted field study of LoRaWAN gateway coverage in different conditions shows the LoRaWAN attenuation coefficient in conditions of city center and outskirt and reveals the factors on which the signal attenuation coefficient depends.

**Keywords** Smart city · LPWAN · LoRaWAN

## 1 Introduction

The first mention of the term “Smart city” date back to the early 2000s [1]. The smart city concept initially described how to use IT infrastructure to create a virtual entity of a city in the information society [2]. Later, the smart city was mainly associated with the strengthening role of intellectual technologies in increasing the efficiency of urban development [3]. Nowadays, the term “Smart city” has transformed to a “smart sustainable city” (SSC), in which information and communication technologies and other tools, on the one hand, are used to improve the effectiveness of the functioning of the city and the provision of urban services and, as a consequence, the overall quality of life, as well as strengthening competitiveness, and on the other hand,

---

V. Shpenst · A. Terleev (✉)  
Saint Petersburg Mining University, St. Petersburg, Russia  
e-mail: [andrew.terleev@gmail.com](mailto:andrew.terleev@gmail.com)

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021  
N. Voinov et al. (eds.), *Proceedings of International Scientific Conference  
on Telecommunications, Computing and Control*, Smart Innovation, Systems  
and Technologies 220, [https://doi.org/10.1007/978-981-33-6632-9\\_45](https://doi.org/10.1007/978-981-33-6632-9_45)

513

satisfy the needs of present and future generations without negatively affecting the economic, social, and ecological components of a city [4].

There are no unified criteria to assess the level of city's smartness [5]. Therefore, several approaches used in Russian and international practice can be cited. For instance, the National Research Institute for Technology and Communications (NIITS) has developed the "Smart Cities Indicators" rating that is based on data obtained from public sources and considers 26 indicators characterizing the level of development of 7 key areas of the smart city [6]. Another approach is used by specialists of Skolkovo Business School: the smartness index is calculated for 15 biggest Russian cities, such as Moscow, Saint Petersburg, Kazan, Volgograd, Novosibirsk, Yekaterinburg, Nizhny Novgorod, Samara, Chelyabinsk, Omsk, Rostov-on-Don, Ufa, Krasnoyarsk, Perm, Voronezh. The calculation method considers 7 spheres of smart city services implementation: transport, healthcare, public administration, media, education, finance, trade [7].

It is rather difficult to estimate the actual scale of the global market for smart city technologies. Moreover, it is even more complicated to forecast how they will change in the medium and long term. However, some attempts to predict the evolution of the smart city market are still being made. Thus, according to the estimates of the research company Markets & Markets, the market volume in 2017 amounted to 424.68 billion US dollars, and by 2022 it will already reach 1.2 trillion US dollar [8]. Frost & Sullivan experts give another estimate: According to their forecasts, the smart city technology market will reach \$2.4 trillion by 2025 [9]. The market growth is also due to the fact that, as it develops, in addition to traditional IT companies and infrastructure giants, new players begin to emerge—small- and medium-sized technology firms, engineering and consulting companies. The estimates of the smart city market volume are varying vastly but it is the obvious fact that the number of smart city services will significantly grow in the nearest future.

Smart city services can be divided into groups according to the sphere of its implementation in the city's ecosystem. In Russia, the following four groups of services are in the spotlight nowadays:

1. **Smart energy.** The group covers the whole chain – from efficient and environmentally friendly energy production (both centralized and decentralized), through distribution system with minimum losses up to responsible consumption on the demand side [10].
2. **Smart and sustainable mobility.** The main benefits of the smart mobility introduction are reduced congestion of roads, reduction of negative impact on the environment, as well as reduction of energy consumption of vehicles [11].
3. **Smart waste management.** This smart city service implies a transparent control of all the waste flows within the city together with best available technics for waste recycling and utilization [12].
4. **Smart environmental solutions.** The package of technological solutions designed to provide better environmental protection for smart cities includes a whole range of components—these are environmental monitoring systems,

smart wastewater treatment systems, and renewable energy solutions. Such solutions contribute to improving the quality of the environment (air, soil, water), the transition to a more rational model of energy use and, as a result, improving the health status of citizens and the sanitary situation in the city as a whole [13].

It is important to understand that all the services require millions of monitoring sensors and actuators which are connected to the single or several management platforms. Today, the number of connected sensors is growing explosively and has already exceeded the number of people with access to the global network [14]. Figure 1 presents the forecast of number of connected devices and sensors until 2025 in comparison with number of people connected to the Internet.

The upcoming smart city revolution requires an infrastructure that can enable the effective interaction of millions of smart devices and sensors [15]. Such interaction has some specific requirements among which:

1. Transfer of small data packages;
2. Energy efficiency;
3. Ability to connect devices in remote areas;
4. High degree of data protection;
5. Interoperability.

In practice, energy efficiency is the biggest challenge from the list above. Connected end-devices have to be able to operate for a long period of time (several years) powered by imbedded battery with no connection to the grid. Otherwise, the maintenance of the end-devices will be difficult and expensive.

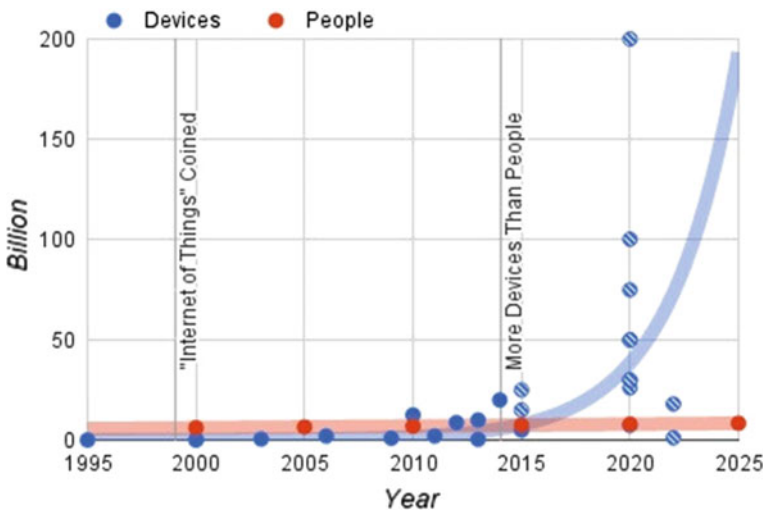


Fig. 1 Forecast of number of connected devices and sensors until 2025 ([www.brookings.edu](http://www.brookings.edu))

## 2 Methods

To effectively solve problems related to energy consumption, new types of Low Power Wide Area Networks (LPWAN) have appeared. Technologies that allow connecting autonomous devices to the global network appeared in 2015–2016 and gradually gaining popularity [16]. The most popular of them are: LoRaWAN, Narrowband IoT (NB-IoT), and Sigfox [17]. Comparison of the main technical characteristics of the networks is presented in Table 1.

Sigfox is practically not presented in Russia [18]. The first few NB-IoT networks were deployed in Moscow and Saint Petersburg in 2019 by local telecommunication companies [19]. In the same time, there are a number of public and private LoRaWAN networks in Russia. It can be explained by affordability of LoRaWAN gateways and end-devices compare to Sigfox and NB-IoT.

In frame of the research, the coverage of LoRaWAN gateway in different conditions has been studied. The conditions were (a) city center and (b) outskirts surrounded by coniferous forest. In both locations, LoRaWAN gateway with vertical omnidirectional antenna (864–876 MHz 10 dBi) was used (Fig. 2).

Coverage area was assessed by measuring the special network tester (Fig. 3) at pre-approved control points. This device sends a special signal to the LoRaWAN network, in response to which the network informs it of the number of gateways that have received this signal and the signal quality. The tester displays this data every time the button is pressed. The device is used to test LoRaWAN networks when they are deployed and configured.

The gateway was stationary at the height of 25 m above the ground, and the tester moved a predetermined distance from the gateway in the range from 50 to 1500 m in increments of 50 m. At each point, a series of 10 measurements of the signal level of the gateway by the tester and the tester signal by the gateway was carried out.

According to the results, the average received signal strength indicator (RSSI) value was determined. Then radio signal attenuation coefficients in different conditions were calculated according to Eq. (1) [20].

$$\text{RSSI} = \text{TSSI} + b * \log_{10}(x), \quad (1)$$

**Table 1** Comparison of LPWAN

Characteristic	LoRa	Sigfox	NB-IoT
Modulation method	CSS	–	OFDMA/DSSS
Range	ISM	ISM	Licensed
Rate	0.3–50 Kb/s	100 bit/s	1–200 Kb/s
Battery life	up to 10 years	–	up to 10 years
Frequency	868.8 MHz (Europe) 915 MHz (USA) 433 MHz (Asia)	868.8 MHz (Europe) 915 MHz (USA)	700/800/900 MHz



**Fig. 2** LoRaWAN gateway and antenna

where RSSI—Received Signal Strength Indicator, dBm;

TSSI—radio signal level at a distance of 1 m from the transmitting antenna, dBm;

$b$ —radio signal attenuation coefficient, dBm;

$x$ —distance between the gateway and the tester, m.

### 3 Results and Discussion

According to the measurements, the coverage area of the LoRaWAN gateway in the city center was 1500 m and in the outskirts surrounded by coniferous forest was 1050 m. After exceeding these distances, the signal was lost. The results of RSSI



**Fig. 3** Tester of LoRaWAN network (<https://www.euromobile.ru>)

measurements in the city center and in the outskirts surrounded by coniferous forest are shown in Figs. 4 and 5 accordingly.

The results of radio signal attenuation coefficients calculation are presented in Table 2.

## 4 Conclusions

Rapid development of smart city services requires new network solutions to allow millions of sensors and devices to communicate with each other. In the same time, smart city applications have a number of specific requirements. The main condition of effective functioning of distributed network of devices and sensors is energy efficiency. Nowadays, smart city services based on LoRaWAN technology has the best prospects among others LPWANs due to affordability on the Russian market and good characteristics.

According to the field study results, LoRaWAN signal attenuation in conditions of coniferous forest is higher compare to the city center. It can be explained by moisture absorption in the trees.

To provide full LoRaWAN coverage for smart city services, it is necessary to locate the gateways on a distance of 1500 m at the city center and of 1050 m at the outskirts at maximum. It means that to cover all the area of Saint Petersburg, approximately 400

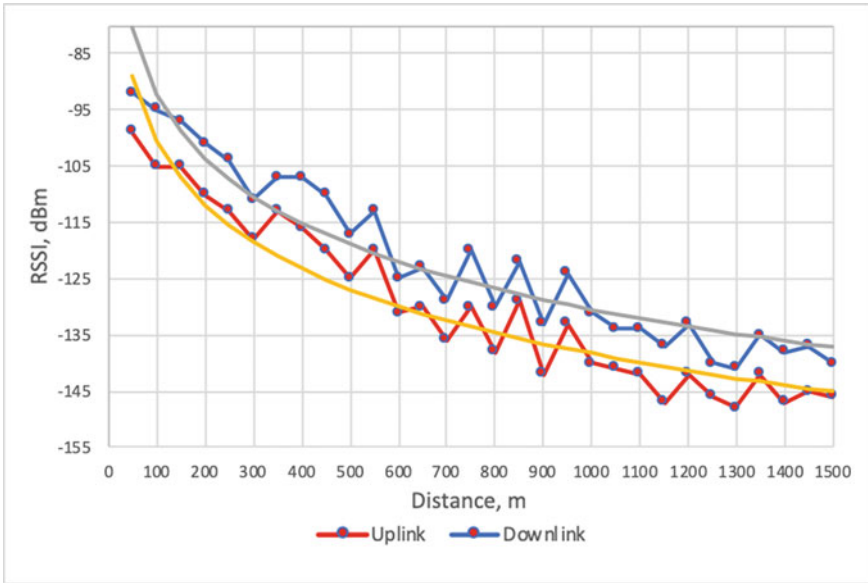


Fig. 4 Dependence of LoRaWAN signal level on distance in the city center

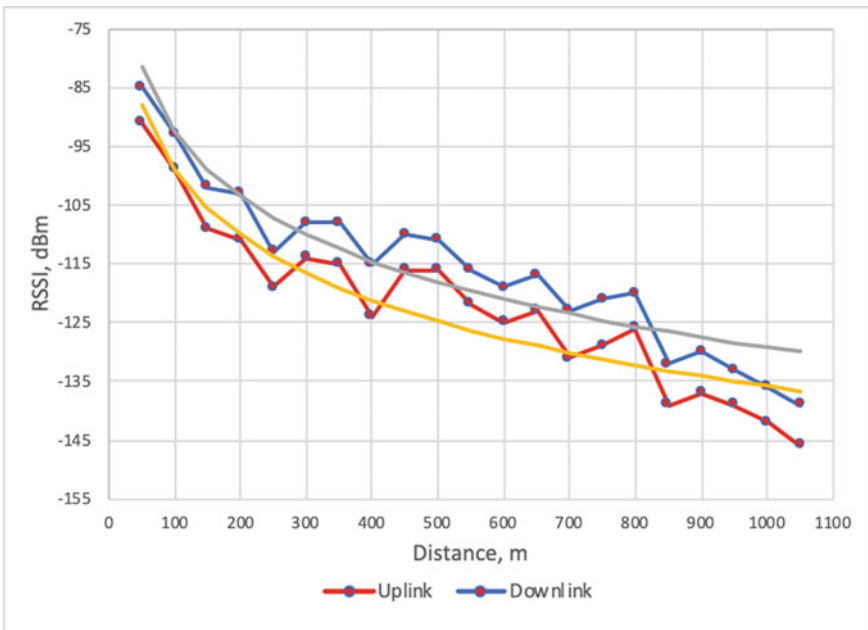


Fig. 5 Dependence of LoRaWAN signal level on distance in the outskirt surrounded by coniferous forest



**Table 2** LoRaWAN signal attenuation coefficient in the conditions under the study

Conditions		LoRaWAN signal attenuation coefficient (dBm)
City center	Uplink	−24.4
	Downlink	−15.4
Outskirt surrounded by coniferous forest	Uplink	−25.5
	Downlink	−19.4

LoRaWAN gateways are needed and 84,000 gateways are required for the Leningrad region.

## References

1. Sikora-Fernandez, D., Stawasz, D.: The concept of smart city in the theory and practice of urban development management. *Rom. J. Reg. Sci.* (2016)
2. Nam, T., Pardo, T.A.: Conceptualizing smart city with dimensions of technology, people, and institutions. In: *ACM International Conference Proceeding Series* (2011)
3. Van der Meer, A., Van Winden, W.: E-governance in cities: a comparison of urban information and communication technology policies. *Reg. Stud.* (2003). <https://doi.org/10.1080/0034340032000074433>
4. Bibri, S.E., Krogstie, J.: Smart sustainable cities of the future: an extensive interdisciplinary literature review (2017)
5. Anthopoulos, L., Janssen, M., Weerakkody, V.: Comparing smart cities with different modeling approaches. In: *WWW 2015 Companion—Proceedings of the 24th International Conference on World Wide Web* (2015)
6. Sozinov, M.: Indicators of smart cities NIITS. (2017)
7. SKOLKOVO Institute for Emerging Market Studies (IEMS): Digital Life of Russian Megapolises Model. *Dynamics Cases* (2016)
8. Markets&Markets: IoT in Smart Cities Market by Solution (Remote Monitoring, Network Management, Analytics, RTLS, Security), Service, Application (Smart Transportation, Buildings, Utilities, Healthcare and Public Safety), and Region—Global Forecast to 2023, <https://www.marketsandmarkets.com/Market-Reports/iot-smart-cities-market-215714954.html>
9. Frost&Sullivan: Smart City Adoption Timeline, <https://store.frost.com/smart-city-adoption-timeline.html>
10. Lund, H., Østergaard, P.A., Connolly, D., Mathiesen, B.V.: Smart energy and smart energy systems (2017)
11. Menuhova, T.: Automation of operational management of interregional cargo transportation. *Proc. Min. Inst.* **2011**, 80–85 (2015)
12. Esmaeilian, B., Wang, B., Lewis, K., Duarte, F., Ratti, C., Behdad, S.: The future of waste management in smart and sustainable cities: a review and concept paper (2018)
13. Artmann, M., Kohler, M., Meinel, G., Gan, J., Ioja, I.C.: How smart growth and green infrastructure can mutually support each other—a conceptual framework for compact and green cities. *Ecol. Indic.* (2019). <https://doi.org/10.1016/j.ecolind.2017.07.001>
14. Evans, D.: The Internet of Things: how the next evolution of the internet is changing everything (2011)
15. Hayat, P.: Smart cities: a global perspective. *India Q.* (2016). <https://doi.org/10.1177/0974928416637930>

16. Koucheryavy, A., Vladyko, A.: The prospects for research in the field of communications networks on the 2017–2020 years. *Telecom IT* **4**, 1–14 (2016)
17. Mekki, K., Bajic, E., Chaxel, F., Meyer, F.: Overview of Cellular LPWAN Technologies for IoT Deployment: Sigfox, LoRaWAN, and NB-IoT. In: 2018 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2018 (2018)
18. Sigfox Ltd.: Sigfox coverage, <https://www.sigfox.com/en/coverage>
19. MTS: MTS Built First Nation-Wide NB-IoT Network in Russia, <http://ir.mts.ru/ir-blog/mts-blog-details/2018/MTS-Built-First-Nation-wide-NB-IoT-Network-In-Russia/default.aspx>
20. Chukin, V.: Preliminary results of experiments to determine the conditions of radio communication between nodes of wireless sensor networks in the frequency range of EU868 in the urban environment and forest area of the Leningrad Region. In: The Twenty-Third Scientific Conference on Radiophysics Dedicated to the 100th Anniversary of the Birth of N.A. Zheleztsova., Nizhny Novgorod (2019)

# Fire Resistance Evaluation of Tempered Glass in Software ELCUT



Marina Gravit , Nikolay Klimin , Alina Karimova ,  
Evgenia Fedotova , and Ivan Dmitriev 

**Abstract** Modern windows and facade glazing elements, depending on the type and purpose of the building, perform various functions. They can be both direct sources of daylighting and elements fireproof structures in external building envelopes. The high temperature in case of fire causes deformations and loss performance properties of structures. Structural calculations and modeling of fire resistance of structures is a hot topic. Designers should know the fire endurance of the structure in the process of solutions development. In this paper, the temperature fields and stress fields under fire exposure are calculated in software package ELCUT. The various thermophysical properties of 6-mm-thick heat-strengthened soda-lime silica float glass as part of single-chamber and two-chamber packages with a cold-framed steel profile and various thickness of glass packs: 6 mm and 24 (with an air gap) taken into account. Berkeley Lab WINDOW software was used to calculate the overall thermal performance of the window. It is shown that ELCUT allows to make a model the fire test of structures and represent temperature fields and stress fields. The interrelation of the occurrence of mechanical stresses from temperature effects with the consideration of the thermophysical properties of materials has been revealed.

**Keywords** Glazed facades · Glass thermal stress · Glazing systems · Window breakage · Fire simulation · ELCUT

## 1 Introduction

Translucent structures are one of the characteristic features of the twenty-first century architecture. None of unique building or structure is complete without glazing.

---

M. Gravit (✉) · A. Karimova · E. Fedotova · I. Dmitriev  
Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia  
e-mail: [marina.gravit@mail.ru](mailto:marina.gravit@mail.ru)

N. Klimin  
Brand Glass Ltd., St. Petersburg, Russia

Fire-resistant glass is a glass product that can withstand the effects of thermal and mechanical loads during a fire, preventing the spread of fire and combustion products [1]. New technologies for the manufacture of glasses that can resist high temperatures are developed. Manufacturers of fire-resistant glass create new chemical compounds in accordance with the requirements of environmental safety [2–4].

Despite the variety of glass (reinforced, heat-strengthened, multilayer laminated) and their manufacturing techniques, there are a number of problems to ensure their fire resistance.

Ordinary single glass with a thickness of 2–4 mm is practically impracticable for use in fireproof translucent structures. The destruction of glass by fire depends on various parameters, such as the type of glass, boundary conditions, and restrictions on glass [5–8].

Features of insurance fire resistance of translucent structures discussed in the papers [9–11]. The features of the behavior of external translucent walls in case of fire and the main criteria for their destruction are described. The authors consider techniques that could predict the destruction of glazing during a fire. The requirements of regulatory documents in the field of fire resistance and fire hazard of external translucent walls establish constructive and technical solutions aimed at improving the fire resistance of structures.

The behavior of heat-strengthened glass in case of fire is considered in [12–14]. The maximum temperature of the water film maintained by heat-strengthened glass was investigated.

Despite the large amount of research on this topic, there are quite a few works that reveal the interrelation of thermal stresses and mechanical stresses arising in glass under fire exposure.

The authors were faced with the task of identifying the interrelation between the occurrence of mechanical stresses from temperature exposure, taking into account the various thermal and physical properties of the materials used and the geometric dimensions of the structures by model engineering structures in the software package ELCUT. Another task is comparing the fire endurance of single-chamber heat-strengthened glass and two-chamber heat-strengthened glass with an air gap.

## 2 Methods

The temperature in the furnace  $T_f$  (1) increases in accordance with the standard temperature conditions ISO 834 [15] as

$$T_f = T_0 + 345 \lg(8t + 1) \quad (1)$$

where

$t$  is the time from the start of the test in minutes,

$T_0$ —temperature before the test, °C.

Glass temperature can be calculated in accordance with the recommendations ISO 15099 [16]. The natural convection heat transfer coefficient for the indoor side,  $h_{cv,int}$ , is determined terms of the Nusselt number, Nu (2):

$$h_{cv,int} = Nu(\lambda/H) \tag{2}$$

where

$\lambda$  is the thermal conductivity of air, W/(m K).

Nu is calculated as a function of the corresponding Rayleigh number based on the height,  $H$ , of the glazing cavity,  $Ra_H$  (3)

$$Ra_H = (\rho^2 H^3 g \cdot C_p |T_2 - T_{int}|) / (T_m \cdot \mu \cdot \lambda) \tag{3}$$

where  $g$ —gravitational acceleration;

$\rho$ —is the air density, kg/m<sup>3</sup>;

$C_p$ —mass heat capacity at constant pressure, J/K;

$\mu$ —absolute viscosity, Pa s;

specific to average temperature  $T_m$  (4)

$$T_m = T_{int} + 1/4(T_2 - T_{int}) \tag{4}$$

For vertically set glass, the Nusselt number can be expressed as (5)

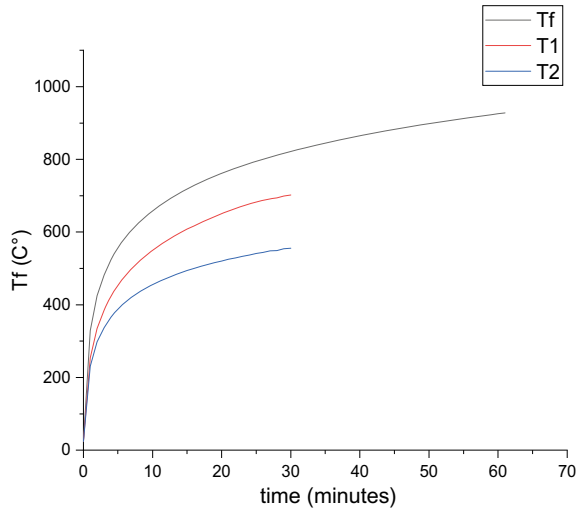
$$Nu = 0.13Ra_H^{1/3} \tag{5}$$

The natural convection heat transfer coefficient is a function of the temperature of the indoor side of the glass; therefore, an iterative procedure is required to calculate these parameters. The temperature of the inner and outer surfaces of the glass can be calculated. The calculation is performed in the WINDOW program [17]. The calculation results are shown in Fig. 1, where  $T_f$  is the temperature in the furnace;  $T_1$  is glass surface temperature from the heated side;  $T_2$  is glass surface temperature on the unheated side (Fig. 2).

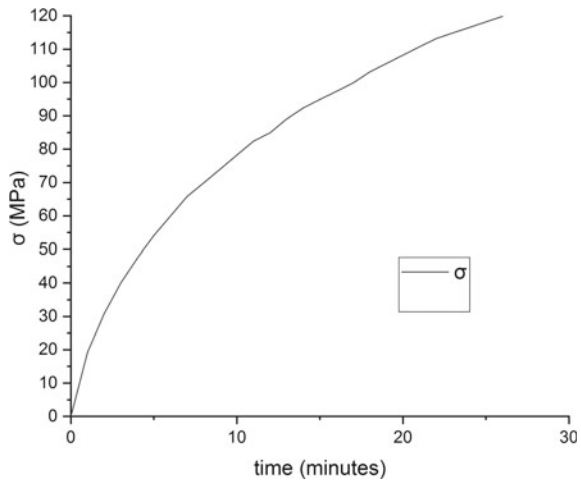
It can be seen from the figure that the glass is cooled by natural convection and the temperature of the outer layer does not reach the glass melting point of 730 °C. However, it can be seen that the temperature rises rapidly and significant differences in temperature occur in the glass, both inside the glass and between the central region and the edge of the glass. These temperature differences create mechanical stresses that can lead to the destruction of glass.

For a rectangular plate with free edges, the temperature stresses can be estimated by the formula as (6)

**Fig. 1** Calculation of the surface temperature of glass 6 mm thick during fire tests. The calculations are limited to a temperature of 800 °C [14]



**Fig. 2** Calculation of mechanical stresses in glass with a thickness of 6 mm during fire tests. The calculations are limited to a temperature of 800 °C [14]



$$\sigma = \alpha E \Delta T / (1 - \nu) \tag{6}$$

where  $\alpha$ —coefficient of expansion by heat of glass  $89 \times 10^{-7} \text{ K}^{-1}$ ,

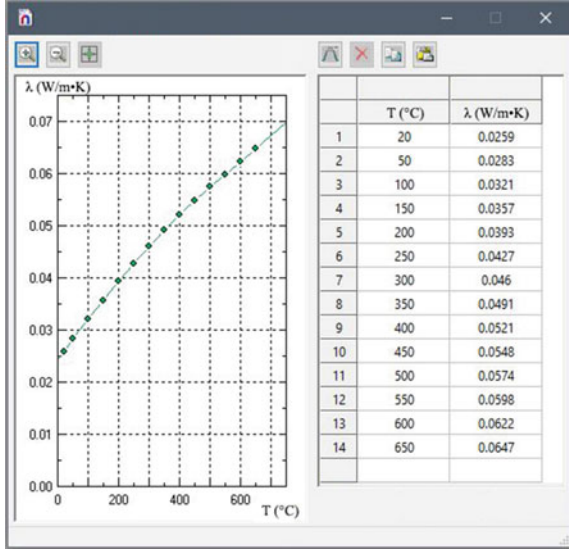
$E$ —Young’s of elasticity—72 GPa,

$\Delta T$ —temperature difference, K.

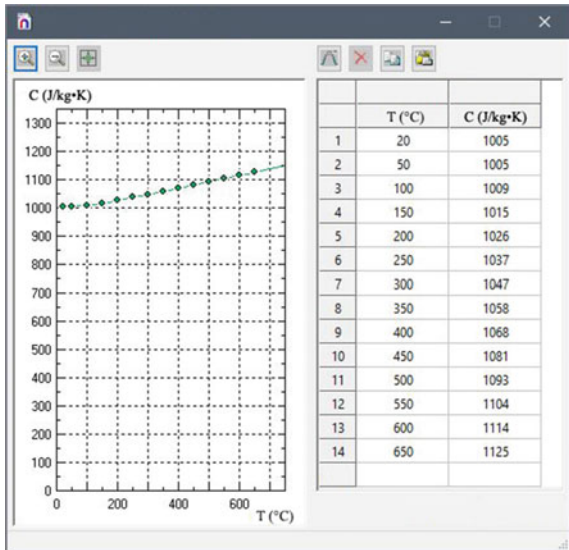
$\nu$ —Poisson’s ratio—0.23.

The maximum permissible difference between the indoor side of the glass temperature and the room temperature of 140 °C and the convection heat transfer coefficient

**Fig. 3** Dependence of thermal conductivity of air on temperature. Values are limited to 650 °C



**Fig. 4** Dependence of specific heat capacity of air on temperature. Values are limited to 650 °C



of the indoor side of the glass—a room of 10 W/m<sup>2</sup> K makes it possible to estimate the convective heat flux  $Q$  at 1.4 KW/m<sup>2</sup>. The reduced heat transfer resistance for such a hypothetical glass unit will be (7)

$$R_{total} = (T_f - T_2)/Q \tag{7}$$

The density of the thermal energy flux due to radiation is determined by the Stefan–Boltzmann law (8)

$$u = \varepsilon_f \varepsilon_g \sigma (T_f^4 - T_0^4) \varphi \quad (8)$$

where  $\varepsilon_f$ —emissivity factor, for a flame, equals 1,

$\varepsilon_g$ —coefficient for glass surface—0.89,

$\sigma$ —Stefan–Boltzmann constant  $5.67 \times 10^{-8}$  W/(m<sup>2</sup> K<sup>4</sup>),

$\varphi$ —angular irradiance coefficient that can be calculated based on recommendations [18].

### 3 Results and Discussion

The calculation of the temperature of the glass unit, taking into account the radiation and convective heat transfer, can be carried out using the ISO-15099 recommendations and in the ELCUT software package.

This software package allows to make models multiphysical processes with the calculation of the parameters of the process of transient heat transfer, while the results of solving the thermal task allow to estimate the temperature distribution in space and solve the nonlinear heat conduction task during heating of the tested structures and to determine the elastic stresses and strains.

The module of the task of calculating the temperature field of the ELCUT program is designed to model 2D temperature fields, solving tasks of steady-state heat transfer and transient Heat transfer [19–24].

Two models of the window frame design have been tested during 6 min. The first one is single-chamber tempered glass. The second one is two-chamber double-glazed window with an air gap. For each structure, two tasks were created: “Transient Heat transfer” and “Stress analysis.” Using the multiphysical connection of tasks, thermal calculation was connected with mechanical and mechanical stresses and strains were obtained in structures under the influence of heat.

Purposes of the model engineering is to obtain the temperature fields of the test structure at the 6th minute of the test and to determine the mechanical stresses created by the temperature difference in the thickness of the glass, which can lead to destruction under the conditions of the standard temperature conditions of the fire.

The model is made in a flat projection of the cross section of the window frame (Table 1).

Figures 5, 6, 7, 8, 9, 10, 11 and 12 show temperature and stress fields based on the model engineering result.

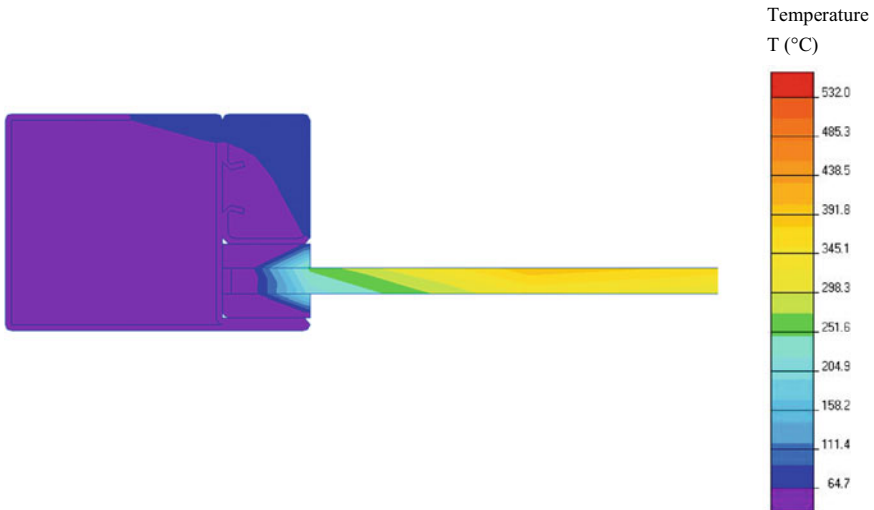
The results of solving the task “stress analysis” for tempered glass:

Model engineering has shown that the temperature of 532 °C is reached on the glass surface of single-chamber tempered glass after 6 min of temperature exposure.



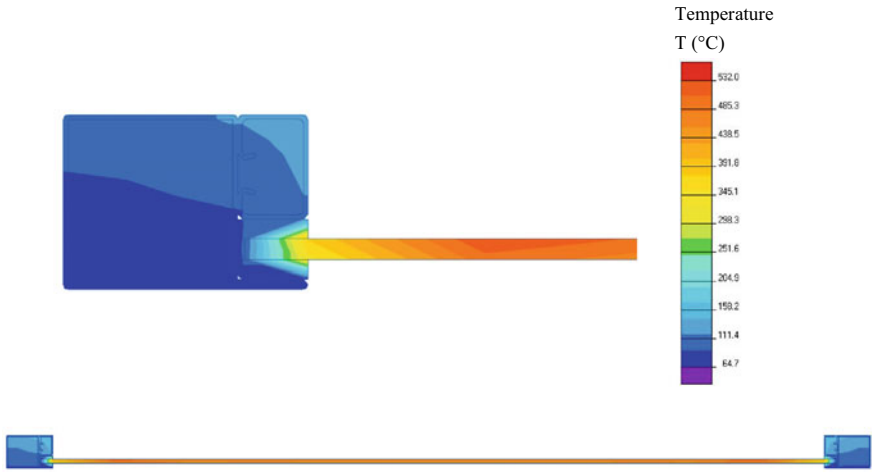
**Table 1** Thermophysical properties of materials

Layer	Material	Thickness (mm)	Thermal conductivity (W/m K)	Specific heat capacity (J/kg K)	Density (kg/m <sup>3</sup> )
Glass	Heat-strengthened glass	6	0.937	880	2500
Cold-framed steel	Steel	1.5	47	460	7800
Gasket kerafix	Calcium-magnesium-silica fibers	5.5	0.037	1650	160
Air gap	Air	12	Thermal conductivity and specific heat capacity are set depending on the temperature and are presented in Figs. 3 and 4, respectively.		1.2



**Fig. 5** Temperature field of a single-chamber double-glazed window with tempered glass after 3 min

This temperature is the yield point of the glass. The temperature on the heated side near the steel profile reaches 361 °C. The temperature on the unheated side is 498 °C. Also stresses of the order of 60 MPa arise on the glass surface which corresponds to the tensile strength of tempered glass. Therefore, after 6 min there is a loss of integrity of the structure with a single glazing. The same stress order was obtained as a result of calculation (Fig. 2).

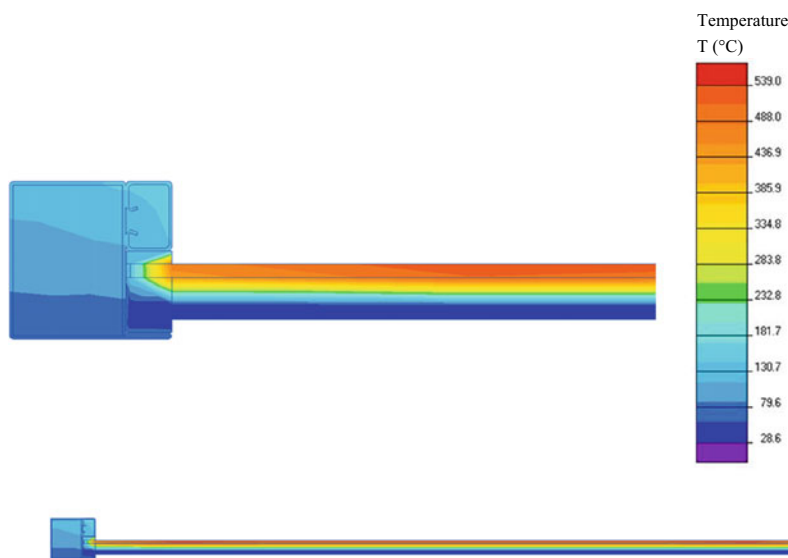


**Fig. 6** Temperature distribution over the junction and construction of a single-chamber double-glazed window with tempered glass after 6 min



**Fig. 7** Temperature field of a two-chamber double-glazed window with tempered glass after 3 min

The temperature of 539 °C is also reached On the surface of the glass of a two-chamber double-glazed window from the heated side after 6 min of temperature exposure. The temperature on the heated side near the steel profile reaches 356 °C. The unheated side is slightly heated to 42 °C. The same stresses of the order of 60 MPa arise on the glass surface of a two-chamber double-glazed window from the heated side. The stress on the unheated surface is zero, and the structure of the two-chamber double-glazed window does not lose its integrity.



**Fig. 8** Temperature distribution over the junction and construction of a two-chamber double-glazed window with tempered glass after 6 min

Results of model engineering can be compared with the experimental data obtained during fire tests in accordance with the national standard GOST R 53308 [25]. Brand Glass Ltd tested various types of fire-resistant glasses. The time of destruction of the outer tempered glass in 13 tests was 5–8 min, while the fire resistance of various types of the double-glazed windows (depending on the thickness and filling) was EIW 30, EIW 60 Figs. 13 and 14. Tests were carried out between 2009 and 2019 in various laboratories, including the FGBU VNIPO of EMERCOM of Russia branch.

## 4 Conclusions

In this paper, the temperature fields and stress fields under fire exposure are calculated in software package ELCUT. The various thermophysical properties of 6-mm-thick heat-strengthened soda-lime silica float glass as part of single-chamber and two-chamber packages with a cold-framed steel profile and various thickness of glass packs: 6 mm and 24 (with an air gap) taken into account.

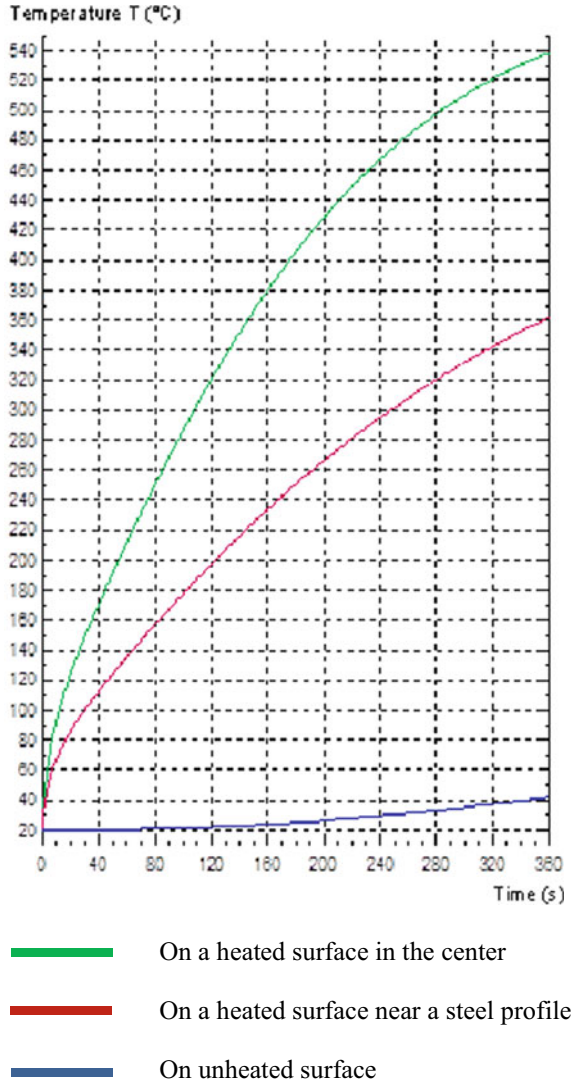
It is shown that ELCUT allows to make a model the fire test of structures, represent temperature fields and stress fields and, thus, estimate the fire resistance of the structure by the loss of thermal insulating capacity and structural integrity. It is represented that the results of model engineering in ELCUT converge with the results of analytical calculations in the WINDOW software.

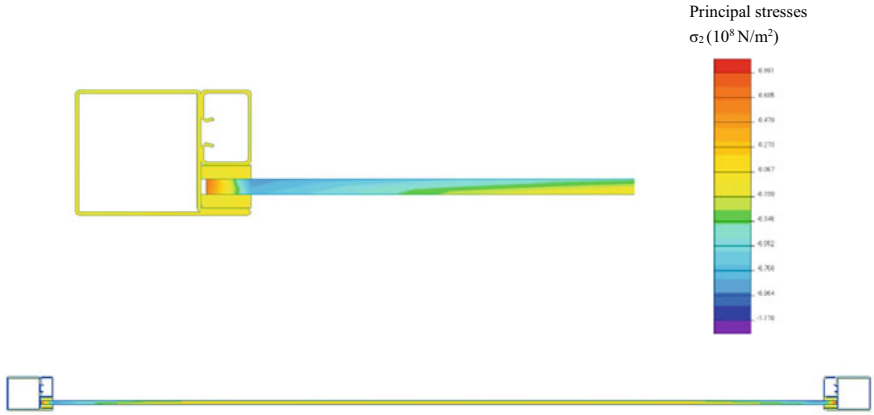
**Fig. 9** Growth curve of temperature in time for a single-chamber double-glazed window with tempered glass



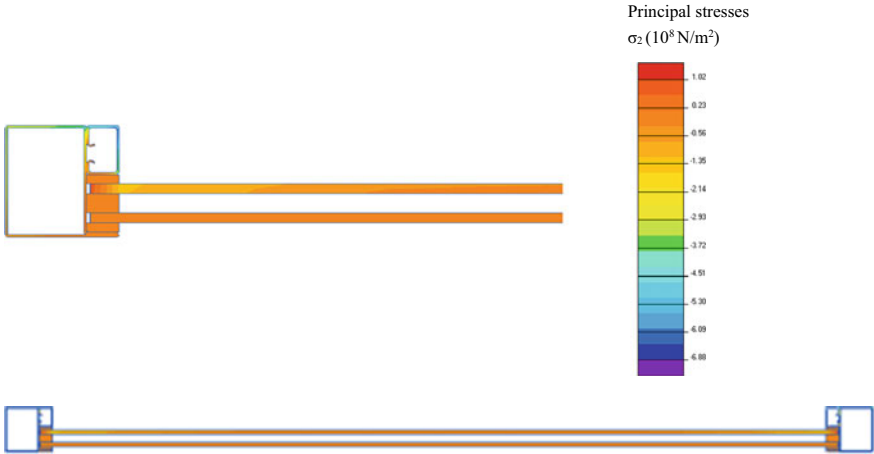
- On a heated surface in the center
- On a heated surface near a steel profile
- On unheated surface

**Fig. 10** Growth curve of temperature in time for a two-chamber double-glazed window with tempered glass





**Fig. 11** Main stresses  $\sigma_2$  of a single-chamber double-glazed window with tempered glass at a temperature effect after 6 min



**Fig. 12** Main stresses  $\sigma_2$  of a two-chamber double-glazed window with tempered glass at a temperature effect after 6 min



**Fig. 13** Tests of fire-resistant glass Brand Glass Paraflam, 1200 × 1000 × 28 mm, EIW 60

**Fig. 14** Tests of fire-resistant glass Brand Glass Paraflam 1200 × 1000 × 20 mm, EIW130



## References

1. EN 357:2004 Glass in building. Fire resistant glazed elements with transparent or translucent glass products. Classification of fire resistance. European Standard (2004)

2. Pukhkal, V., Mottaeva, A.B.: FEM modeling of external walls made of autoclaved aerated concrete blocks. *Mag. Civ. Eng.* **81**, 202–211 (2018). <https://doi.org/10.18720/MCE.81.20>
3. Shukhardin, A., Gravit, M., Dmitriev, I., Nefedov, G., Nazmeeva, T.: Fire simulation of light gauge steel frame wall system with foam concrete filling. *Adv. Intell. Syst. Comput.* **982**, 836–844 (2020). [https://doi.org/10.1007/978-3-030-19756-8\\_80](https://doi.org/10.1007/978-3-030-19756-8_80)
4. Galyamichev, A.V., Alkhimenko, A.I.: Design features of facade cassettes from thin ceramics. *Mag. Civ. Eng.* **69**, 64–76 (2017). <https://doi.org/10.18720/MCE.69.6>
5. Block, V.L.: *The Use of Glass in Buildings*. (2002). <https://doi.org/10.1520/stp1434-eb>
6. Dembele, S., Rosario, R.A.F., Wen, J.X.: Thermal breakage of window glass in room fires conditions—analysis of some important parameters. *Build. Environ.* **54**, 61–70 (2012). <https://doi.org/10.1016/j.buildenv.2012.01.009>
7. Peng, L., Ni, Z., Huang, X.: Review on the fire safety of exterior wall claddings in high-rise buildings in China. *Procedia Eng.* **62**, 663–670 (2013). <https://doi.org/10.1016/j.proeng.2013.08.112>
8. Kryukova, A.A., Vergizova, M.V., Gravit, M.V., Vaititckii, A.A., Nedryshkin, O.V.: The process of increasing the fire resistance of glass with preservation its operational properties. *Constr. Unique Build. Struct.* **1**, 17–26 (2017). <https://doi.org/10.18720/CUBS.52.2>
9. Shao, G., Wang, Q., Zhao, H., Wang, Y., Chen, H., Su, Y., Sun, J., He, L.: Maximum temperature to withstand water film for tempered glass exposed to fire. *Constr. Build. Mater.* **57**, 15–23 (2014). <https://doi.org/10.1016/j.conbuildmat.2014.01.094>
10. Xie, Q., Zhang, H., Wan, Y., Zhang, Q., Cheng, X.: Full-scale experimental study on crack and fallout of toughened glass with different thicknesses. *Fire Mater.* **32**, 293–306 (2008). <https://doi.org/10.1002/fam.968>
11. Jurinak, J.J., Summers, L.E.: Oilfield applications of colloidal silica gel. *SPE (Society Pet. Eng. Prod. Eng. (United States))*. <https://doi.org/10.2118/18505-PA>
12. Brinker, C.J., Scherer, G.W.: *Sol-Gel Science: The Physics and Chemistry of Sol-Gel Processing*. Academic Press Inc, San Diego (1990)
13. Hamouda, A.A., Amiri, H.A.A.: Factors affecting alkaline sodium silicate gelation for in-depth reservoir profile modification. *Energies* **7**, 568–590 (2014). <https://doi.org/10.3390/en7020568>
14. Curcija, D., Goss, W.P.: New correlations for convective heat transfer coefficient on indoor fenestration surfaces—compilation of more recent work. *Therm. VI Therm. Perform. Exter. Envel. Build.* **VI**, 567–572 (1995)
15. ISO 834-75: Elements of building constructions. Fire-resistance test methods. General requirements. Inter-State Scientific and Technical Commission on Standardisation in Construction (1996)
16. ISO 15099:2003: Thermal performance of windows, doors and shading devices. Detailed calculations. ISO (2003)
17. Curcija, C., Vidanovic, S., Hart, R., Jonsson, J., Mitchell, R.: *WINDOW Technical Documentation*. Lawrence Berkeley National Laboratory, California (2018)
18. EN 1991-1-2:2002 Eurocode 1: Actions on structures—Part 1-2: General actions. Actions on structures exposed to fire. European committee for standardization (2002)
19. Baiburin, A.K., Rybakov, M.M., Vatin, N.I.: Heat loss through the window frames of buildings. *Mag. Civ. Eng.* **85**, 3–14 (2019). <https://doi.org/10.18720/MCE.85.1>
20. Dmitriev, I., Lyulikov, V., Bazhenova, O., Bayanov, D.: Calculation of fire resistance of building structures in software packages. *E3S Web Conf.* **91**, 1–6 (2019). <https://doi.org/10.1051/e3s/conf/20199102007>
21. Dudin, M.O., Vatin, N.I., Barabanshchikov, Y.G.: Modeling a set of concrete strength in the program ELCUT at warming of monolithic structures by wire. *Mag. Civ. Eng.* **54**, 33–45 (2015). <https://doi.org/10.5862/MCE.54.4>
22. Klimin, N.N., Rivkind, V.Y., Pachin, V.A.: Collision efficiency calculation model as a software tool for microphysics of electrified clouds. *Meteorol. Atmos. Phys.* **53**, 111–120 (1994). <https://doi.org/10.1007/BF01031908>
23. Saknite, T., Serdjuks, D., Goremikins, V., Pakrastins, L., Vatin, N.I.: Fire design of arch-type timber roof. *Mag. Civ. Eng.* **64**, 26–39 (2016). <https://doi.org/10.5862/MCE.64.3>



24. Karkin, I.N., Subachev, S.V., S.A.A.: Algorithm for the identification of rooms in FDS-projects for fire modeling by the integral method. *Fire Explos. Saf.* **24**, 45–53 (2015). <https://doi.org/10.18322/pvb.2015.24.11.45-53>
25. State Standard 53308-2009. Building structures. Fire-resistance tests methods. Standartinform Publ, Moscow (2009)

# Author Index

## A

Abdikarimov, Rustamkhan, 109  
Alqezweeni, Mohie, 475  
Al-Windi, Basim, 49  
Andropova, Polina, 277  
Antonov, Alexander, 255  
Antonova, Tat'yana, 407  
Arefeva, Daria, 407  
Arseniev, Dmitry, 177, 343  
Averchenkova, Elena, 395  
Avlasko, Pavel, 419

## B

Baskakov, Dmitry, 333, 343  
Belenko, Viacheslav, 15  
Bordyugov, Semen, 419  
Borshchev, Igor, 99

## C

Cheremisin, Dmitriy, 277  
Chernenko, Valery, 15  
Chernyshev, Alexander, 185  
Chistyakova, Tamara, 119  
Chukanov, Viacheslav, 245  
Chuvatov, Mikhail, 185  
Coolen, Frank, 205

## D

Denisov, Victor, 385  
Deylid, Ivan, 27  
Dmitriev, Ivan, 523  
Drobintsev, Pavel, 67

Drozd, Oleg, 419  
Dyakov, Nikita, 407  
Dybov, Anton, 231

## E

Egiazarov, Tigran, 285  
Egorenko, Marina, 451, 461

## F

Fedotova, Evgenia, 523

## G

Ge, Dong, 195  
Glazunov, Vadim, 185  
Gorbachenko, Vladimir, 475  
Gravit, Marina, 523  
Gritskevich, Evgenij, 451, 461

## H

Hanin, Leonid, 185  
Hlopin, Sergei, 307

## I

Ilina, Elena, 407  
Ipatov, Oleg, 357  
Ivanov, Evgeniy, 99  
Ivanov, Vladimir, 177, 357

## J

Jakovlev, Dmitri, 165

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

N. Voinov et al. (eds.), *Proceedings of International Scientific Conference on Telecommunications, Computing and Control, Smart Innovation, Systems and Technologies* 220, <https://doi.org/10.1007/978-981-33-6632-9>

**K**

Kalinin, Maxim, 1, 15  
 Kamaletdinova, Iuliia, 165  
 Kapulin, Denis, 419  
 Karelsky, Pavel, 431  
 Karimova, Alina, 523  
 Karmanov, Igor, 445  
 Kasimov, Ernest, 219  
 Khlopin, Sergei, 293  
 Khodzhaev, Dadakhan, 109  
 Kim, Kiseon, 153  
 Kisilev, Ivan, 255  
 Klimin, Nikolay, 523  
 Klygach, Alexander, 177  
 Klyuyev, Arsentiy, 99  
 Kocherzhinskaya, Yuliya, 407  
 Kochovski, Petar, 67  
 Kohlert, Christian, 119  
 Konstantinov, Andrei, 205, 245  
 Korneyev, Vladimir, 445  
 Kostenko, Dmitri, 131  
 Kots, Mikhail, 245  
 Kovalev, Maxim, 219  
 Kovalevsky, Vladislav, 231  
 Kovzur, Max, 431  
 Kraeva, Svetlana, 293  
 Krainikovskiy, Stanislav, 491  
 Krasnov, Sergey, 385  
 Krasov, Andrey, 431  
 Krundyshev, Vasilii, 1, 15  
 Kulikov, Nikita, 499  
 Kurochkin, Leonid, 185  
 Kurochkin, Mikhail, 185

**L**

Levandovskii, Il'ya, 407  
 Lukashin, Aleksey, 321

**M**

Makarova, Diana, 451, 461  
 Makaruk, Roman, 119  
 Meldo, Anna, 205, 277  
 Melnikov, Mikhail, 491  
 Mikhalev, Igor, 321  
 Molodyakov, Sergey, 27  
 Monastyrev, Vitaly, 67  
 Mushchak, Nikita, 293, 307

**N**

Nesterov, Sergey, 285  
 Nikiforov, Igor, 37

Nikulin, Dmitry, 445  
 Normuminov, Bakhodir, 109  
 Novikov, Sergei, 451, 461

**O**

Onufriev, Vadim, 131, 231

**P**

Pankov, Pavel, 37  
 Polikanin, Alexey, 461  
 Popov, Mikhail, 321  
 Pozigun, Mikhail, 245  
 Prourzin, Vladimir A., 153

**R**

Reichert, Valery, 445

**S**

Sadykov, Ilya, 119  
 Sal'nikov, Vyacheslav, 365  
 Samarev, Roman, 491  
 Saradgishvili, Sergey, 267  
 Semenov, Konstantin, 365, 373  
 Sergeev, Sergey, 357  
 Shaburova, Aelita, 451, 461  
 Shel, Egor, 293, 307  
 Shergin, Sergey, 445  
 Shevlyakov, Georgy, 153, 165  
 Shkodyrev, Vyacheslav, 131, 333  
 Shpenst, Vadim, 513  
 Shvedova, Arina, 499  
 Sidnev, Alexandr, 77  
 Smolina, Elena, 285  
 Sokolova, Victoria, 205  
 Starobinskii, Egor, 293, 307  
 Strelkov, Sergey, 177

**T**

Tasyuk, Aleksander, 431  
 Terleev, Andrei, 513  
 Timofeev, Andrey, 385  
 Timofeev, Dmitrii, 321  
 Titov, Aleksandr, 385  
 Trinh, Quang-Kien, 141  
 Tropkina, Iuliia, 195  
 Tselishcheva, Anastasia, 373  
 Tutygin, Vladimir, 49  
 Tyutin, Boris, 27

**U**

Utkin, Lev, [205](#), [219](#)

**V**

Vassiliev, Alexei, [141](#)

Voronkov, Iliia, [267](#)

**Y**

Yaitskaya, Elena, [499](#)

Yulchiyev, Davron, [109](#)

**Z**

Zaborovskij, Vladimir, [255](#)

Zaitsev, Gleb, [141](#)

Zavjalov, Sergey, [195](#)

Zhalnin, Vladimir, [499](#)

Zhang, Yufeng, [37](#)

Zharkovskii, Aleksandr, [99](#)

Zhemelev, Georgiy, [77](#)

Zuyev, Igor, [431](#)

Zvyagintseva, Polina, [451](#), [445](#)