

Lecture Notes in Networks and Systems 702

Anatoly Alikhanov
Pavel Lyakhov
Irina Samoylenko *Editors*

Current Problems in Applied Mathematics and Computer Science and Systems

 Springer

Series Editor

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

Advisory Editors

Fernando Gomide, *Department of Computer Engineering and Automation—DCA, School of Electrical and Computer Engineering—FEEC, University of Campinas—UNICAMP, São Paulo, Brazil*

Okay Kaynak, *Department of Electrical and Electronic Engineering, Bogazici University, Istanbul, Türkiye*

Derong Liu, *Department of Electrical and Computer Engineering, University of Illinois at Chicago, Chicago, USA*

Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, *Department of Electrical and Computer Engineering, University of Alberta, Alberta, Canada*

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

Marios M. Polycarpou, *Department of Electrical and Computer Engineering, KIOS Research Center for Intelligent Systems and Networks, University of Cyprus, Nicosia, Cyprus*

Imre J. Rudas, *Óbuda University, Budapest, Hungary*

Jun Wang, *Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong*

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

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

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

For proposals from Asia please contact Aninda Bose (aninda.bose@springer.com).

Anatoly Alikhanov · Pavel Lyakhov ·
Irina Samoylenko
Editors

Current Problems in Applied Mathematics and Computer Science and Systems

Editors

Anatoly Alikhanov
Information Systems Department
Stavropol State Agrarian University
Stavropol, Russia

Pavel Lyakhov
Information Systems Department
Stavropol State Agrarian University
Stavropol, Russia

Irina Samoylenko
Information Systems Department
Stavropol State Agrarian University
Stavropol, Stavropol Territory, Russia

ISSN 2367-3370

ISSN 2367-3389 (electronic)

Lecture Notes in Networks and Systems

ISBN 978-3-031-34126-7

ISBN 978-3-031-34127-4 (eBook)

<https://doi.org/10.1007/978-3-031-34127-4>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2023

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

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

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

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

Preface

This volume contains the proceedings of International Conference on Current Problems of Applied Mathematics and Computer Systems and Science. The conference was held by North-Caucasus Centre for Mathematical Research, North-Caucasus Federal University.

The papers are focused on four sections: numerical methods in scientific computing; information and computation systems for distributed environments; data analysis and modular computing and actual problems of mathematical education. All presented papers include significant research achievements.

Numerical methods field contains such contributions as the modeling of the potential dependence on the permittivity at the metal–dielectric medium interface and modeling of temperature contribution to the interphase energy of the faces of cadmium crystals at the boundary with organic liquids; difference method for solving the Dirichlet problem for a multidimensional integro-differential equation of convection–diffusion; the third boundary value problem for a multidimensional convection–diffusion equation with memory effect; initial–boundary value problems for the loaded Hallaire equation with Gerasimov–Caputo fractional derivatives of different orders; grid method for solving local and nonlocal boundary value problems for a loaded moisture transfer equation with two fractional differentiation operators. Additionally, authors describe forced longitudinal oscillations of a rod with a mass at the end and determining the frequencies of free longitudinal vibrations of rods by analytical and numerical methods. The section also includes works on calculating the hyperbolic parameter of a two-dimensional lattice of linear comparison solutions as well as fast calculation of parameters of parallelepipedal nets for integration and interpolation. Particular attention was paid to transport in porous media. Important applied discoveries in the modeling of the adjustable DC voltage source for industrial greenhouse lighting systems, unassociated matrices number of the n order and a given determinant and the problem of restoring the unit of approximation in the model for studying functional dependence from approximate data were also presented in the section.

Information and computation systems for distributed environments issues is presented by analysis of influence of Byzantine robots with random behavior strategy on collective decision making in swarms; mathematical concept of a model for processing metadata of employee’s psycho-states for identifying him as an internal violator; modified CLNet; facemask wearing correctness detection using deep learning approaches; an approach to the implementation of nonparametric algorithms for controlling multidimensional processes in a production problem. The section presents works on the review and analysis of artificial intelligence methods. There are also a number of works on information security issues, including smart cities and companies with a focus on the application of the residue number system for data protection.

The field of data analysis and modular computing includes such significant contributions, as trust monitoring in a cyber-physical system for security analysis based on

distributed computing; application of the SIFT algorithm in the architecture of a convolutional neural network for human face recognition; temperature profile nowcasting using temporal convolutional network; guaranteed safe states in speed space for ships collision avoidance problem; an ensemble of UNet frameworks for lung nodule segmentation; no-reference metrics for images quality estimation in a face recognition task. Particular attention was paid to the analysis of medical data: neural network skin cancer recognition with a modified cross-entropy loss function; cloud-based service for recognizing pigmented skin lesions using a multimodal neural network system; application of bidirectional LSTM neural networks and noise pre-cleaning for feature extraction on an electrocardiogram. The section also featured a number of important works on modular computing: modulo 2^{k+1} truncated multiply-accumulate unit; RNS reverse conversion algorithm and parity detection for the wide arbitrary moduli set; improving the parallelism and balance of RNS with low-cost and 2^{k+1} modules. An essential part of the section consisted of works in the field of developing new approaches to digital signal processing: uncoded video data stream denoising in a binary symmetric channel; high-speed wavelet image processing using the Winograd method; fractionation discrete neural network of bidirectional associative memory and using soft decisions for error correction.

In addition, the volume contains study in the field of mathematical education: interactive methods in the study of the discipline “Mathematics” for nonmathematical specialties; applied mathematics and informatics bachelor’s and master’s educational programs continuity during updating educational standards and designing computer simulators in support of a propaedeutic course on neural network technologies.

The target audience of the conference proceedings includes postgraduates, lecturers at institutions of higher education and researchers who study mathematics and its applications in computer systems. Based on the conclusions and results offered, representatives of the targeted audience are likely to find essential knowledge and suggestions for future research.

Contents

Numerical Methods in Scientific Computing

Modeling of the Potential Dependence on the Permittivity at the Metal – Dielectric Medium Interface	3
<i>Aslan Apekov and Liana Khamukova</i>	
Difference Method for Solving the Dirichlet Problem for a Multidimensional Integro-Differential Equation of Convection-Diffusion	15
<i>Zaryana Beshtokova</i>	
The Problem of Restoring the Unit of Approximation in the Model for Studying Functional Dependence from Approximate Data	26
<i>E. Yartseva, L. Andruhiv, and R. Abdulkadirov</i>	
RNS Reverse Conversion Algorithm and Parity Detection for the Wide Arbitrary Moduli Set	36
<i>Vitaly Slobodskoy, Elizaveta Martirosyan, Valeria Ryabchikova, and Roman Kurmaev</i>	
Anomalous Solute Transport in an Inhomogeneous Porous Medium Taking into Account Mass Transfer	45
<i>T. O. Dzhiyanov, Sh Mamatov, and M. S. Zokirov</i>	
Determination of Relaxation and Flow Coefficients During Filtration of a Homogeneous Liquid in Fractured-Porous Media	54
<i>Erkin Kholiyarov and Mirzohid Ernazarov</i>	
Solution of the Anomalous Filtration Problem in Two-Dimensional Porous Media	68
<i>Jamol Makhmudov, Azizbek Usmonov, and Jakhongir Kuljonov</i>	
On Calculating the Hyperbolic Parameter of a Two-Dimensional Lattice of Linear Comparison Solutions	81
<i>N. N. Dobrovol'skii, N. M. Dobrovol'skii, I. Yu. Rebrova, and E. D. Rebrov</i>	
Inverse Problem of Contaminant Transport in Porous Media	87
<i>B. H. Khuzhayorov, E. Ch. Kholiyarov, and O. Sh. Khaydarov</i>	

Numerical Solution of Anomalous Solute Transport in a Two-Zone Fractal Porous Medium	98
<i>Bakhtiyor Khuzhayorov, Azizbek Usmonov, and Fakhridin Kholliiev</i>	
Initial-Boundary Value Problems for the Loaded Hallaire Equation with Gerasimov–Caputo Fractional Derivatives of Different Orders	106
<i>Murat Beshtokov</i>	
Grid Method for Solving Local and Nonlocal Boundary Value Problems for a Loaded Moisture Transfer Equation with Two Fractional Differentiation Operators	118
<i>Murat Beshtokov</i>	
Determining Frequencies of Free Longitudinal Vibrations of Rods by Analytical and Numerical Methods	131
<i>Kh. P. Kulterbaev, M. M. Lafisheva, and L. A. Baragunova</i>	
Forced Longitudinal Oscillations of a Rod with a Mass at the End	137
<i>Kh. P. Kulterbaev, M. M. Lafisheva, and L. A. Baragunova</i>	
On the Unassociated Matrices Number of the n Order and a Given Determinant	146
<i>Urusbi Pachev and Rezuhan Dokhov</i>	
Fast Calculation of Parameters of Parallelepipedal Nets for Integration and Interpolation	157
<i>N. N. Dobrovol'skii, N. M. Dobrovol'skii, Yu. A. Basalov, and E. D. Rebrov</i>	
Modeling of the Adjustable DC Voltage Source for Industrial Greenhouse Lighting Systems	167
<i>Vladimir Samoylenko, Vladimir Fedorenko, and Nikolay Kucherov</i>	
Modeling of Temperature Contribution to the Interphase Energy of the Faces of Cadmium Crystals at the Boundary with Organic Liquids	179
<i>Aslan Apekov, Irina Shebzukhova, Liana Khamukova, and Zaur Kokov</i>	
Information and Computation Systems for Distributed Environments	
Mathematical Concept of a Model for Processing Metadata of Employee's Psycho-States for Identifying Him as an Internal Violator (Insider)	189
<i>I. V. Mandritsa, V. I. Petrenko, O. V. Mandritsa, and T. V. Minkina</i>	

Analysis of Influence of Byzantine Robots with Random Behaviour Strategy on Collective Decision-Making in Swarms	205
<i>V. I. Petrenko, F. B. Tebueva, S. S. Ryabtsev, V. O. Antonov, and I.V Struchkov</i>	
Beamforming for Dense Networks-Trends and Techniques	217
<i>Nabarun Chakraborty, Aradhana Misra, and Kandarpa Kumar Sarma</i>	
Modified CLNet: A Neural Network Based CSI Feedback Compression Model for Massive MIMO System	233
<i>Dikshita Sarma, Mrinmoy Shandilya, Aradhana Misra, and Kandarpa Kumar Sarma</i>	
Facemask Wearing Correctness Detection Using Deep Learning Approaches	243
<i>Atlanta Choudhury and Kandarpa Kumar Sarma</i>	
An Approach to the Implementation of Nonparametric Algorithms for Controlling Multidimensional Processes in a Production Problem	250
<i>Vyacheslav Zolotarev, Maria Lapina, and Darya Liksonova</i>	
Analysis of Neural Networks for Image Classification	258
<i>Nikolay Vershkov, Mikhail Babenko, Viktor Kuchukov, Nataliya Kuchukova, and Nikolay Kucherov</i>	
Factors of a Mathematical Model for Detection an Internal Attacker of the Company	270
<i>I. V. Mandritsa, V. V. Antonov, and Siyanda L. Madi</i>	
Comparative Analysis of Methods and Algorithms for Building a Digital Twin of a Smart City	277
<i>Vladislav Lutsenko and Mikhail Babenko</i>	
Modification of the Projection Method to Correct Errors in RNS	288
<i>Egor Shiriaev, Viktor Kuchukov, and Nikolay Kucherov</i>	
An Overview of Modern Fully Homomorphic Encryption Schemes	300
<i>Ekaterina Bezuglova and Nikolay Kucherov</i>	
Model of Error Correction Device in RNS-FRNN	312
<i>Egor Shiriaev and Viktor Kuchukov</i>	
Review of Modern Technologies of Computer Vision	321
<i>Ekaterina Bezuglova, Andrey Gladkov, and Georgy Valuev</i>	

Data Analysis and Modular Computing

Discrete Neural Network of Bidirectional Associative Memory 335
Aleksey V. Shaposhnikov, Andrey S. Ionisyan, and Anzor R. Orazhev

Modulo $2^k + 1$ Truncated Multiply-Accumulate Unit 343
Maxim Bergerman, Pavel Lyakhov, and Albina Abdulsalyamova

Neural Network Skin Cancer Recognition with a Modified Cross-Entropy Loss Function 353
Ulyana Alekseevna Lyakhova

Application of the SIFT Algorithm in the Architecture of a Convolutional Neural Network for Human Face Recognition 364
Diana Kalita and Parviz Almamedov

High-Speed Wavelet Image Processing Using the Winograd Method 373
N. N. Nagornov, N. F. Semyonova, and A. S. Abdulsalyamova

Improving the Parallelism and Balance of RNS with Low-Cost and $2^k + 1$ Modules 381
Pavel Lyakhov

Uncoded Video Data Stream Denoising in a Binary Symmetric Channel 391
Anzor R. Orazhev

Cloud-Based Service for Recognizing Pigmented Skin Lesions Using a Multimodal Neural Network System 401
Ulyana Alekseevna Lyakhova, Daria Nikolaevna Bondarenko, Emiliya Evgenevna Boyarskaya, and Nikolay Nikolaevich Nagornov

Temperature Profile Nowcasting Using Temporal Convolutional Network 410
Nikolay Baranov

Application of Bidirectional LSTM Neural Networks and Noise Pre-cleaning for Feature Extraction on an Electrocardiogram 421
Mariya Kiladze

Trust Monitoring in a Cyber-Physical System for Security Analysis Based on Distributed Computing 430
Elena Basan, Maria Lapina, Alexander Lesnikov, Anatoly Basyuk, and Anton Mogilny

Guaranteed Safe States in Speed Space for Ships Collision Avoidance Problem 441
A. S. Zhuk

An Ensemble of UNet Frameworks for Lung Nodule Segmentation 450
Nandita Gautam, Abhishek Basu, Dmitry Kaplun, and Ram Sarkar

No-Reference Metrics for Images Quality Estimation in a Face Recognition Task 462
Aleksander S. Voznesensky, Aleksandr M. Sinitca, Evgeniy D. Shalugin, Sergei A. Antonov, and Dmitrii I. Kaplun

Using Soft Decisions for Error Correction 475
Tatyana Pavlenko and Oleg Malafey

Actual Problems of Mathematical Education

Interactive Methods in the Study of the Discipline “Mathematics” for Non-mathematical Specialties 489
Irina Lavrinenko, Natalia Semenova, and Valentina Baboshina

Applied Mathematics and Informatics Bachelor’s and Master’s Educational Programs Continuity During Updating Educational Standards 500
Irina Zhuravleva, Ludmila Andrukhiv, Elena Yartseva, and Valentina Baboshina

Designing Computer Simulators in Support of a Propaedeutic Course on Neural Network Technologies 509
Andrej Ionisyan, Irina Zhuravleva, Irina Lavrinenko, Aleksey Shaposhnikov, and Violetta Liutova

Author Index 521

About the Authors

Anatoly Alikhanov has obtained his Ph.D. in Physical and Mathematical Sciences and is Vice-Rector for Scientific and Research work of the North Caucasus Federal University. He is the Head of the Regional Scientific and Educational Mathematical Center “North Caucasian Center for Mathematical Research.” Earlier, he worked as Visiting Member of the dissertation council at Southeast University, Department of Mathematics, Nanjing, China. In 2016, he was invited for an internship in the field of numerical methods for solving fractional differential equations at the Nanjing University of China. A. Alikhanov is Member of the Editor Board of the Fractional Calculus and Applied Analysis Journal and Reviewer of over 30 reputable scientific journals in computational mathematics.



Pavel Lyakhov graduated in mathematics in Stavropol State University, in 2009, where he also received the Ph.D. degree in computer science, in 2012. He has been working at North-Caucasus Federal University, since 2012. He is currently the Head of the Department of Mathematical Modeling, North-Caucasus Federal University. His research interests include high-performance computing, quantum computing, residue number systems, digital signal processing, image processing and machine learning.

Irina Samoylenko, Ph.D., Associate Professor at Informational Systems Department, Stavropol State Agrarian University, Russia. She received M.S. degree in Applied Mathematics and Informatics and Ph.D. in System Analysis, Control and Processing of Information in North-Caucasus Federal University. Irina is a researcher in North-Caucasus Centre for Mathematical Research, North-Caucasus Federal University. She is a member of Association of Scientific Editors and Publishers, Russia. Her research interests include wireless sensor networks, IoT, optimization tasks and mathematical modeling.

Numerical Methods in Scientific Computing



Modeling of the Potential Dependence on the Permittivity at the Metal – Dielectric Medium Interface

Aslan Apekov^(✉)  and Liana Khamukova 

North Caucasus Center for Mathematical Research, North Caucasus Federal University,
Stavropol, Russia
aslkbbsu@yandex.ru

Abstract. The processes occurring at the interface of metal with organic dielectric media have a wide practical application in various devices. The features of the metal-organic interface make it possible to create materials with practically useful properties used in catalysis, energy storage, electronics, gas storage and separation, magnetism, nonlinear optics, etc. Knowing the course of the potential at the phase interface, it is possible to obtain interphase characteristics, including interphase energy. In this paper, the dependence of the potential on the permittivity is modeled in the framework of a modified version of the Frenkel–Gambosch–Zadumkin electron-statistical theory at the metal–dielectric medium boundary. The course of the dimensionless potential at the interface is obtained and it is shown that the greater the dielectric constant of the medium, the more the dimensionless potential drops at the physical interface. The coordinate of the Gibbs interface for the metal–dielectric medium system is obtained, which can be found from the condition of electroneutrality at this boundary. It is shown that with an increase in the value of the permittivity, the Gibbs coordinate increases, that is, it shifts towards the dielectric medium. The dependence of the interfacial energies of faces with different structures on the dielectric permittivity is shown.

Keywords: interfacial energy · electron-statistical method · polarization correction · dispersion interaction · alkali metals · dielectric medium

1 Introduction

Organometallic structures are widely used in practice [1]. The features of the metal-organic interface make it possible to create materials with practically useful properties used in catalysis [2], energy storage, electronics [3], storage and separation of gases [4, 5], magnetism, nonlinear optics, etc. [6–10].

Theoretical and experimental studies devoted to the influence of various factors on the magnitude and behavior of the surface characteristics of metal crystals are of great interest and their number is increasing every year. One of the important characteristics of the surface layer is the interfacial energy (IE). The IE of metals at the boundary with organic liquids and its orientation dependence have not been sufficiently studied. Within

the framework of the Frenkel–Gambosch-Zadumkin [11] theory, an electron–statistical method for calculating the interphase energy at the boundary with nonpolar organic liquids is proposed, the anisotropy of the interphase energy is obtained, the general laws of the dependence of the interphase energy on temperature, the atomic number of the metal and the permittivity of the organic liquid are established [12–20].

In metals at room temperature, the energy of thermal motion is sufficient for a part of the valence electrons to acquire sufficient energy to overcome the potential barrier and exit the atom. But this energy is not enough to get out of the metal. These electrons are in the metal and they form an electron liquid. Atoms of the metal that have lost valence electrons become positively charged and form a crystal lattice in the nodes of which they are located. Thus, the positive charge of the ion skeletons is placed in the nodes of the crystal lattice, which is immersed in an electronic liquid. The coordinate axis is drawn perpendicular to the interface and directed towards the dielectric medium. The physical interface has a coordinate $x = 0$ and is drawn tangentially to the surface ions in such a way that all positive ions of the solid metal belong entirely to the inner region occupied by the metal lattice (Fig. 1).

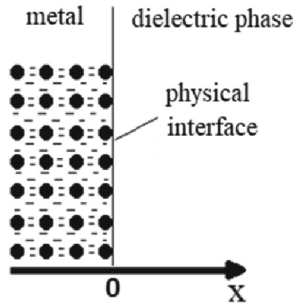


Fig. 1. The physical interface.

2 Materials and Methods

The physical properties of the interphase boundary are influenced by the crystal lattice of ionic skeletons, an electronic liquid that has an electron density $\rho(x)$, as well as molecules of a dielectric medium. To understand the properties of the interphase boundary, it is necessary to understand how the electron density behaves at this boundary. In this regard, we will try to simulate the behavior of the electron density at the interface between a metal crystal and a dielectric medium.

We will try to find the behavior of the electron density $\rho(x)$ and potential $V(x)$ near the metal–dielectric medium interface from the Thomas-Fermi equation for the inner and outer regions of the metal, taking into account the macroscopic permittivity ϵ_0 organic liquid:

$$\frac{d^2V}{dx^2} = 4\pi e[\rho(x) - \nu_+(x)] \quad \text{for } x \leq 0, \quad (1)$$

$$\frac{d^2V}{dx^2} = \frac{4\pi e}{\varepsilon_0} \rho(x) \quad \text{for } x > 0, \quad (2)$$

where $v_+(x)$ is the positive charge distribution density of metal ions. Or through the potential of Eqs. (1) and (2) we rewrite as

$$\frac{d^2V(x)}{dx^2} = 4\pi e\gamma \left(V^{3/2}(x) - V_i^{3/2} \right) \quad \text{for } x \leq 0, \quad (3)$$

$$\frac{d^2V(x)}{dx^2} = \frac{4\pi e\gamma}{\varepsilon_0} V^{3/2}(x) \quad \text{for } x > 0 \quad (4)$$

where eV_i is Fermi boundary energy, e is electron charge, $\gamma = 2^{3/2}/3\pi^2e^2a_0$, a_0 is radius of the first Bohr orbit of a hydrogen atom.

Equations (3) and (4) are crosslinked on the interface $x = 0$ by the conditions of continuity of potential V and its derivative $\frac{dV}{dx}$. In addition, the potential is equal to the Fermi potential in the depth of the metal and zero in the depth of the dielectric liquid

$$\begin{aligned} V &= V_i \quad \text{at } x = -\infty, \\ V &= 0 \quad \text{at } x = +\infty. \end{aligned}$$

To solve Eqs. (3) and (4), we proceed to the dimensionless potential $\chi(\beta) = \frac{V(x)}{V_i}$ and the dimensionless coordinate $\beta = \frac{x}{s}$, and also reduce these equations to a dimensionless form assuming $s^2 4\pi e\gamma V_i^{1/2} = 1$ (here, s is a linear parameter that reduces the Thomas–Fermi equation to a dimensionless form). Then the Eqs. (3) and (4) will take the form:

$$\chi''(\beta) = \chi^{3/2}(\beta) - 1 \quad \text{for } \beta \leq 0, \quad (5)$$

$$\chi''(\beta) = \frac{1}{\varepsilon_0} \chi^{3/2}(\beta) \quad \text{for } \beta > 0, \quad (6)$$

Equations (5) and (6) are solved under the following boundary conditions:

$$\begin{aligned} \chi(\beta) &= 0 \quad \text{at } \beta = +\infty; \\ \chi(\beta) &= 1 \quad \text{at } \beta = -\infty; \\ \chi'(\beta) &= 0 \quad \text{at } \beta = \pm\infty. \end{aligned}$$

Multiplying both parts of Eqs. (5) and (6) by $\chi'(\beta)$ and integrating twice, taking into account the boundary conditions, we obtain solutions:

$$\beta = - \int_{\chi(0, \varepsilon_0)}^{\chi(\beta, \varepsilon_0)} \frac{d\chi(\beta)}{\left(\frac{4}{5}\chi^{5/2}(\beta) - 2\chi(\beta) + \frac{6}{5}\right)^{1/2}} \quad \text{for } \beta \leq 0, \quad (7)$$

$$\chi_c(\beta, \varepsilon_0) = \frac{4^4}{(c - a\beta)^4} \quad \text{for } \beta > 0 \quad (8)$$

where $\chi(0, \varepsilon_0)$ is a dimensionless potential on the physical interface, depending on the dielectric constant of the liquid, $a = 2/\sqrt{5\varepsilon_0}$. Since on the physical interface $\chi_e(0, \varepsilon_0) = \chi_i(0, \varepsilon_0)$, then from (8) we get $c = 4\chi^{-1/4}(0, \varepsilon_0)$. Now the solution of the TF equation in the outer region of the metal will take the form:

$$\chi_e(\beta, \varepsilon_0) = \frac{\chi(0, \varepsilon_0)}{(1 + \beta/b)^4} \quad \text{for } \beta > 0 \quad (9)$$

where $b = \frac{2\sqrt{5\varepsilon_0}}{\chi^{1/4}(0, \varepsilon_0)}$. Bearing in mind (9), the solution of Eq. (5) is explicitly approximated, as in [21] in the form

$$\chi_i(\beta, \varepsilon_0) = 1 - \frac{A_0}{(1 - \beta/b)^n} \quad \text{for } \beta \leq 0 \quad (10)$$

Then A_0 and n are found from the condition of continuity of the potential $\chi(\beta, \varepsilon_0)$ and the first derivative $\chi'(\beta, \varepsilon_0)$ at the interface, i.e. we obtain

$$A_0 = 1 - \chi(0, \varepsilon_0), \quad n = \frac{4\chi(0, \varepsilon_0)}{1 - \chi(0, \varepsilon_0)}.$$

Then formula (10) will finally take the form

$$\chi_i(\beta, \varepsilon_0) = 1 - \frac{1 - \chi(0, \varepsilon_0)}{(1 - \beta/b)^n} \quad \text{for } \beta \leq 0 \quad (11)$$

Taking into account the continuity $\chi'(\varepsilon)$ on the physical interface, finding $\chi'(\varepsilon)$ from (5) and (6), an equation is compiled to determine $\chi(0, \varepsilon_0)$,

$$\frac{2}{5} \left(1 - \frac{1}{\varepsilon_0}\right) \chi^{5/2}(0, \varepsilon_0) - \chi(0, \varepsilon_0) + \frac{3}{5} = 0. \quad (12)$$

From expression (12) at $\varepsilon_0 = 1$, we obtain $\chi(0, 1) = 3/5$; $b = 2(125/3)^{1/4}$, the value of this quantity for the metal–vacuum boundary.

The calculation of interfacial properties at the boundary of the metal crystal face – dielectric medium is carried out using the Gibbs determination of the free surface energy relative to the equimolar interface of the metal – dielectric medium (Fig. 2).

The coordinate of the Gibbs interface for the metal – dielectric medium system is found from the condition of electroneutrality at this boundary as:

$$\int_{-\infty}^0 \left(1 - \chi_i^{3/2}(\beta, \varepsilon_0)\right) d\beta + \int_0^{\beta_G(\varepsilon_0)} \left(1 - \chi_e^{3/2}(\beta, \varepsilon_0)\right) d\beta = \int_{\beta_G(\varepsilon_0)}^{\infty} \chi_e^{3/2}(\beta, \varepsilon_0) d\beta, \quad (13)$$

where $\beta_G(\varepsilon_0)$ is coordinate of the Gibbs interface. Substituting (9) and (11) into expression (13), we find:

$$\beta_G(\varepsilon_0) = b \left\{ \frac{\chi^{3/2}(0, \varepsilon_0)}{5} + \int_{-\infty}^1 \left[1 - \left(1 - \frac{1 - \chi(0, \varepsilon_0)}{t^6}\right)^{3/2} \right] dt \right\}. \quad (14)$$

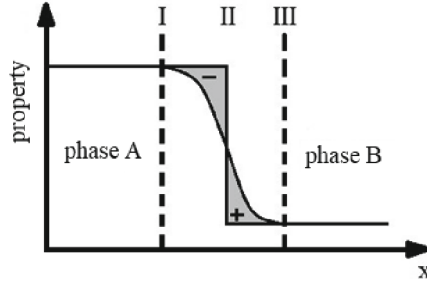


Fig. 2. Gibbs interface.

One of the important characteristics of the phase separation is the interphase energy. Experimentally measuring its values at the interphase boundaries of small dimension presents significant difficulties. Sometimes the interfacial tension is measured, and then using the relationship between interfacial tension and energy, the interfacial energy is calculated. But basically the values of the interphase energy are obtained theoretically.

The electron–statistical method for calculating the surface energy of metals at the metal – vacuum interface proposed by Zadumkin S.N., Shebzukhova I.G. was modified to determine the interphase energy at the metal–organic liquid interface in [12–20].

We divide the metal into Wigner-Seitz cells and replace each such cell with a sphere of equal volume, which we call an s-sphere. Then the interphase energy $f_{12}^{\omega}(hkl)$ for the crystal faces with Miller indices (hkl) can be represented as

$$f_{12}^{\omega}(hkl) = \eta n^{(0)}(hkl) \left(F^{(0)} - F^{(\infty)} \right) + \sum_{k=1}^{\infty} \left(F^{(k)} - F^{(\infty)} \right) n^{(k)}(hkl) + \int_0^{\infty} [\bar{\omega}_e(x) - \bar{\omega}_e(\infty)] dx, \quad (15)$$

where $\eta = \frac{1}{4}(1 + g)(2 - g)$, $g = r_i/R$, r_i is ion radius, $R - s$ is sphere radius, $F^{(0)}$ and $n^{(0)}(hkl)$ are the total free energy of the Wigner-Seitz cell and their number per 1 cm² of the plane located at distance $x = k \delta$ (hkl) on the surface $k = 0$ (δ (hkl) – interplane distance); $F^{(k)}$ is the free energy of an elementary ball on a plane k , $F^{(\infty)}$ is the free energy of an elementary ball inside a metal at $k = \infty$; $\eta F^{(0)}$ is free energy at $k = 0$; $\bar{\omega}_e(x)$ and $\bar{\omega}_e(\infty)$ are external energy densities of electron gas and saturated vapor at a distance x and $x = \infty$ from the surface. The dielectric medium in the outer region is taken into account, as mentioned above, through the macroscopic permittivity of the liquid.

Let us represent $F^{(k)}$ and $F^{(\infty)}$ in formula (15) as:

$$F^{(k)} = E^{(k)} + F_{osc}^{(k)} \quad \text{for} \quad F^{(\infty)} = E^{(\infty)} + F_{osc}^{(\infty)}, \quad (16)$$

where $E^{(k)}$ and $E^{(\infty)}$ are binding energies at $T = 0$ K per elementary ball; $F_{osc}^{(k)}$ and $F_{osc}^{(\infty)}$ are free energies of the vibrational motion of ions, taking into account the energy of zero oscillations and anharmonicity. Taking into account the comments made, if we neglect the excess energy density of saturated steam in the external space, expression (15) can

be rewritten as:

$$f_{12}^{\omega}(hkl) = \eta n^{(0)}(hkl) \left(E^{(0)} - E^{(\infty)} \right) + \sum_{k=1}^{\infty} \left(E^{(k)} - E^{(\infty)} \right) n^{(k)}(hkl) + \sum_{k=0}^{\infty} \left(F_{osc}^{(k)} - F_{osc}^{(\infty)} \right) n^{(k)}(hkl) + \int_0^{\infty} [\bar{\omega}_e(x) - \bar{\omega}_e(\infty)] dx + \Delta\omega_{eT}, \quad (17)$$

where $\Delta\omega_{eT}$ is correction related to temperature blurring of the Fermi level.

Interphase energy can be represented as the sum of internal, external and temperature contributions by entering the following notation

$$f_{12}^{\omega(i0)}(hkl) = \eta n^{(0)}(hkl) \left(E^{(0)} - E^{(\infty)} \right) + \sum_{k=1}^{\infty} \left(E^{(k)} - E^{(\infty)} \right) n^{(k)}(hkl), \quad (18)$$

$$f_{12}^{\omega(e0)} = \int_0^{\infty} [\bar{\omega}_e(x) - \bar{\omega}_e(\infty)] dx, \quad (19)$$

$$\Delta f_{12}^{\omega(T)} = \sum_{k=0}^{\infty} \left(F_{koi}^{(k)} - F_{koi}^{(\infty)} \right) n^{(k)}(hkl) + \Delta\omega_{eT} \quad (20)$$

where is the interfacial energy at $T = 0$ K

$$f_{12}^{\omega(0)}(hkl) = f_{12}^{\omega(i0)}(hkl) + f_{12}^{\omega(e0)}, \quad (21)$$

and at temperature T

$$f_{12}^{\omega}(hkl) = f_{12}^{\omega(0)}(hkl) + f_{12}^{\omega(T)}. \quad (22)$$

The excess energy inside the metal $\Delta E^{(k)} = E^{(k)} - E^{(\infty)}$ in k -th elementary ball compared to $k = \infty$ is composed of the excess free energy $\overline{\Delta E^{(k)}}$ of the electron gas of the metal, the excess energy of the interaction of the positive ion with the electron gas $\overline{\Delta W^{(k)}}$ and the electrostatic energy of the interaction of positively charged elementary balls $\overline{\Delta A^{(k)}}$ in the transition layer, that is

$$\Delta E^{(k)} = \overline{\Delta E^{(k)}} + \overline{\Delta W^{(k)}} + \overline{\Delta A^{(k)}} + \Delta E_p^{(k)} + \Delta E_g^{(k)} + \Delta E_{osc}^{(k)}, \quad (23)$$

and

$$\overline{\Delta E^{(k)}} = \Delta E_C^{(k)} + \Delta E_K^{(k)} + \Delta E_A^{(k)} + \Delta E_W^{(k)} + \Delta E_V^{(k)}, \quad (24)$$

$$\overline{\Delta W^{(k)}} = \Delta W_C^{(k)} + \Delta W_E^{(k)} + \Delta W_K^{(k)} + \Delta W_A^{(k)}. \quad (25)$$

Here $\Delta E_C^{(k)}$ is the excess electrostatic Coulomb energy of electron interaction in the k -th elementary ball; $\Delta E_K^{(k)}$ is the excess kinetic energy of the electron gas at absolute zero

temperature (excess Fermi energy); $\Delta E_A^{(k)}$ is the excess exchange energy of an electron gas in the Thomas-Fermi theory (Bloch excess energy); $\Delta E_W^{(k)}$ is the excess correlation energy of electrons, taking into account the interaction of electrons with antiparallel spins (excess Wigner energy); $\Delta E_V^{(k)}$ is an excess quantum correction to the kinetic energy of an electron gas in an elementary ball (the Weizsacker–Kirzhnitz correction, which is introduced due to the inhomogeneity of the potential in which electrons move near the surface); $\Delta W_C^{(k)}$ is the excess Coulomb energy of the interaction of a point positive ion with conduction electrons in the k -th elementary ball; $\Delta W_E^{(k)}$ is the excess electrostatic energy resulting from the fact that when an electron cloud of valence electrons overlaps with an electron cloud of an ion in the k -th elementary ball, a change in the electrostatic interaction is observed compared with the interaction of point charges; $\Delta W_K^{(k)}$ is the excess kinetic energy caused by the penetration of valence electrons into the electron cloud of the atomic skeleton and is due to the Pauli principle; $\Delta W_A^{(k)}$ is the excess energy of the exchange interaction of the electron gas with the electrons of the atomic skeleton in the k -th elementary ball.

If we substitute an expression for the course of the electron density at the interface into the Poisson equation and integrate it, then we can find an expression for the electric field strength at the metal–dielectric medium interface in the k -th elementary ball:

$$E^{(k)} = -\left(\frac{dV}{dx}\right). \quad (26)$$

We will proceed from the fact that the excess charge of the elementary ball is distributed evenly over the volume of the ion. Then, taking the potential and the electron density in dimensionless form, we can write an expression for the excess energy in the k -th elementary ball. The electrostatic energy $\Delta A^{(k)}$ of interaction of positively charged Wigner-Seitz cells in the transition layer (per elementary ball) can be approximately determined by the formula

$$\overline{\Delta A^{(k)}} = \frac{e}{2} \int_{\varepsilon_k} \Delta v_+(x) V(x) d\omega_k = \frac{e}{2} V_{iz} g^3 \overline{(1 - \chi_{ik}^{3/2})} \chi_{ik}. \quad (27)$$

It is also necessary to take into account the excess energy of the dispersion interaction between elementary s -spheres $\Delta E_g^{(k)}$, the deformation energy of the polarization of the atomic core by s -electrons $\Delta E_p^{(k)}$ and the oscillation of the electron density ΔE_{osc} .

The course of the electron density at the interface is defined as

$$\frac{\rho_{ik}(\beta)}{\rho(\infty)} = \chi_{ik}^{3/2}(\beta, \varepsilon_0), \quad (28)$$

Integrating the volumetric energy density by the volume of the elementary ball

$$\int_{\Omega} \rho_{ik}^n(x) d\Omega = \overline{\rho_{ik}^n(x)} \Omega \quad (29)$$

the expressions for the energy excesses in the k-th elementary ball can be written as

$$\Delta E_C^{(k)} = E_C^{(\infty)} \left(\overline{\chi_{ik}^3(\beta, \varepsilon_0)} - 1 \right), \quad (30)$$

$$\Delta E_K^{(k)} = E_K^{(\infty)} \left(\overline{\chi_{ik}^{5/2}(\beta, \varepsilon_0)} - 1 \right), \quad (31)$$

$$\Delta E_A^{(k)} = E_A^{(\infty)} \left(\overline{\chi_{ik}^2(\beta, \varepsilon_0)} - 1 \right), \quad (32)$$

$$\Delta E_W^{(k)} = E_W^{(\infty)} \left(\overline{\chi_{ik}^2(\beta, \varepsilon_0) \frac{1 + \rho^{1/3}(\infty)/\beta_0}{1 + \chi_{ik}^{1/2}(\beta, \varepsilon_0)\rho^{1/3}(\infty)/\beta_0}} - 1 \right), \quad (33)$$

$$\Delta E_V^{(k)} = \frac{k_V}{2} \int_{\Omega} \frac{1}{\rho(x)} (\nabla \rho(x))^2 d\Omega = \frac{9}{8} \frac{k_V}{s^2} z \chi_{ik}^{-1/2}(\beta, \varepsilon_0) \left(\frac{d\chi_{ik}(\beta, \varepsilon_0)}{d\beta} \right)^2, \quad (34)$$

$$\Delta W_C^{(k)} = W_C^{(\infty)} \left(\overline{\chi_{ik}^{3/2}(\beta, \varepsilon_0)} - 1 \right), \quad (35)$$

$$\Delta W_E^{(k)} = W_E^{(\infty)} \left(\overline{\chi_{ik}^{3/2}(\beta, \varepsilon_0)} - 1 \right), \quad (36)$$

$$\Delta W_K^{(k)} = W_K^{(\infty)} \left(\overline{\chi_{ik}^{3/2}(\beta, \varepsilon_0)} - 1 \right), \quad (37)$$

where $\beta_0 = \frac{0,1216}{a_0}$, $k_V = \frac{\hbar^2}{16\pi^2 m}$.

The individual components of the elementary ball energy at $k = 0$, included in $f_{12}^{\omega(i0)}$, were calculated using the formulas obtained in [22]

$$E_C^{(\infty)} = \frac{3Z^2 \cdot e^2}{5R}, \quad (38)$$

$$E_K^{(\infty)} = 1,105 \frac{Z^{5/3} \cdot e^2 \cdot a_0}{R^2}, \quad (39)$$

$$E_A^{(\infty)} = -0,4582 \frac{Z^{4/3} \cdot e^2}{R}, \quad (40)$$

$$E_W^{(\infty)} = -0,0172 \frac{Z \cdot e^2}{a_0} - 0,0577 \frac{Z^{4/3} \cdot e^2}{R}, \quad (41)$$

$$W_C^{(\infty)} = -\frac{3Z^2 \cdot e^2}{2R}, \quad (42)$$

$$W_E^{(\infty)} + W_K^{(\infty)} + W_A^{(a)(\infty)} = \frac{3Z \cdot e^2}{R^3} \left(\frac{Z \cdot r^2}{6} - \frac{5r^3}{32\pi^2 a_0} \right), \quad (43)$$

$$W_A^{(b)(\infty)} = \frac{\alpha \cdot Z^{4/3} \cdot e^2}{R^4}. \quad (44)$$

Substituting (30)–(37) and (38)–(44) into (18), we obtain an expression for the internal contribution at $T = 0$ K

$$\begin{aligned}
 f_{12}^{\omega(i0)} = & E_C^{(\infty)} \sum_{k=0}^{\infty} \left(\overline{\chi_{ik}^3}(\beta, \varepsilon_0) - 1 \right) n_s^{(k)} + E_K^{(\infty)} \sum_{k=0}^{\infty} \left(\overline{\chi_{ik}^{3/2}}(\beta, \varepsilon_0) - 1 \right) n_s^{(k)} + \\
 & + \left(E_A^{(\infty)} + W_A^{(b)(\infty)} \right) \sum_{k=0}^{\infty} \left(\overline{\chi_{ik}^2}(\beta, \varepsilon_0) - 1 \right) n_s^{(k)} + \\
 & + \left(W_C^{(\infty)} + W_E^{(\infty)} + W_K^{(\infty)} + W_A^{(a)(\infty)} \right) \sum_{k=0}^{\infty} \left(\overline{\chi_{ik}^{3/2}}(\beta, \varepsilon_0) - 1 \right) n_s^{(k)} + \\
 & + E_W^{(\infty)} \sum_{k=0}^{\infty} \left(\overline{\chi_{ik}^2(\beta, \varepsilon_0) \frac{1 + \rho^{1/3}(\infty)/\beta_0}{1 + \chi_{ik}^{1/2}(\beta, \varepsilon_0) \rho^{1/3}(\infty)/\beta_0} - 1} \right) n_s^{(k)} + \\
 & + \frac{k_V Z}{8 S^2} \sum_{k=0}^{\infty} n_s^{(k)} \chi_{ik}^{1/2}(\beta, \varepsilon_0) \left(\frac{d\chi_{ik}(\beta, \varepsilon_0)}{d\beta} \right)^2 + \\
 & + \frac{ezV_i}{2} g^3 \sum_{k=0}^{\infty} n_s^{(k)} \left(1 - \overline{\chi_{ik}^{3/2}}(\beta, \varepsilon_0) \right) \overline{\chi_{ik}(\beta, \varepsilon_0)} + f_{12}^{\omega(g)} + f_{12}^{\omega(p)} + f_{12}^{\omega(osc)}
 \end{aligned} \tag{45}$$

where $\chi_{ik}(\beta, \varepsilon_0) = 1 - \frac{1 - \chi(0, \varepsilon_0)}{(1 - \beta_k/b)^{\beta}}$. The stroke at the sum sign indicates that for $k = 0$ it is necessary to multiply all the terms by $\eta = (1 + f + g)^2(2 - f - g)/4$, $f = \beta_{\Gamma} s/R$ and $g = r/R$, where r is the radius of the ion, R is the radius of the s–sphere.

3 Results

Table 1 shows the obtained values of the Gibbs surface depending on the permittivity of the dielectric medium. We have obtained the course of the dimensionless potential at the metal–dielectric medium interface depending on the permittivity of the medium (Fig. 3).

According to the formula (22), the values of the interfacial energy of the faces are obtained depending on the permittivity of the contacting dielectric medium (Fig. 4).

4 Discussion

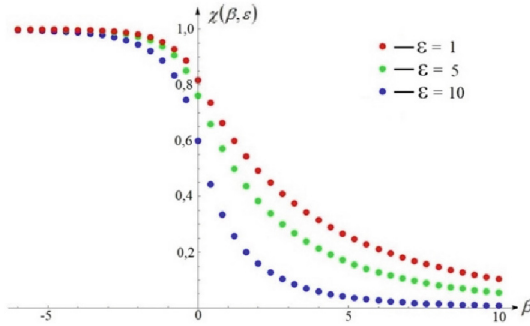
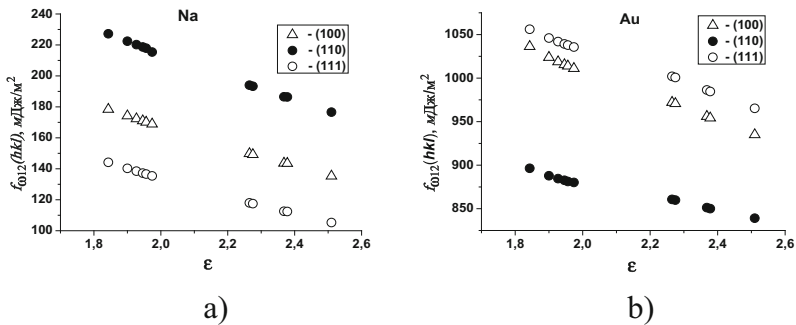
As can be seen from the table, with an increase in the value of the permittivity, the Gibbs coordinate increases, that is, it shifts towards the dielectric medium. As can be seen from Fig. 3, the greater the dielectric constant of the medium, the more the dimensionless potential drops at the physical interface.

The obtained results show:

- The presence of a dielectric liquid leads to a significant decrease in the surface energy of the faces of metal crystals. With an increase in the permittivity of an organic liquid, the value of the IE decreases.
- As the atomic number of the element in the group increases, the IE decreases.
- For metals with a VCC structure at the boundary with nonpolar organic liquids, the following pattern is observed for IE $f_{\omega_{12}}(110) > f_{\omega_{12}}(100) > f_{\omega_{12}}(111)$.
- For metals with the FCC structure, the IE of the faces is correlated as $f_{\omega_{12}}(111) > f_{\omega_{12}}(100) > f_{\omega_{12}}(110)$.

Table 1. The coordinate of the Gibbs interface at the metal – organic liquid interface

ε_0	$\beta_{\Gamma}(\varepsilon_0)$
1,843	0,269
1,900	0,290
1,927	0,300
1,946	0,307
1,974	0,317
2,265	0,416
2,275	0,419
2,368	0,449
2,378	0,452
2,510	0,493

**Fig. 3.** The course of the dimensionless potential at the boundary depending on the dielectric constant of the medium.**Fig. 4.** Interfacial energy of a) sodium with VCC structure and b) gold with FCC structure.

5 Conclusion

The dependence of the Gibbs surface and the interphase energy on the dielectric constant of the medium, as well as the course of the electron density at the interphase boundary, is shown. The obtained results demonstrate that the dielectric medium significantly affects the interfacial energy. This dependence can be used in the design of electronic devices. It is of interest to apply this approach to describe the interfacial boundary of different metals.

Acknowledgments. The work is supported by North-Caucasus Center for Mathematical Research under agreement №. 075-02-2022-892 with the Ministry of Science and Higher Education of the Russian Federation.

References

1. Czaja, A.U., Trukhan, N., Müller, U.: Industrial applications of metal–organic frameworks. *Chem. Soc. Rev.* **38**(5), 1284–1293 (2009)
2. Seo, J., et al.: A homochiral metal–organic porous material for enantioselective separation and catalysis. *Nature* **404**, 982–986 (2000)
3. Liu, Z., Kobayashi, M., Paul, B.C., Bao, Z., Nishi, Y.: Contact engineering for organic semiconductor devices via Fermi level depinning at the metal–organic interface. *Phys. Rev. B* **82**(3), 035311 (2010)
4. Mendoza-Cortes, J.L., Han, S.S., Goddard, W.A.: High H₂ Uptake in Li-, Na-, K-metalated covalent organic frameworks and metal organic frameworks at 298 K. *Phys. Chem. A* **116**(6), 1621–1631 (2012)
5. Furukawa, H., et al.: Ultrahigh porosity in metal–organic frameworks. *Science* **329**, 424–428 (2010)
6. Butova, V.V., Soldatov, M.A., Guda, A.A., Lomachenko, K.A., Lamberti, C.: Metal–organic frameworks: structure, properties, methods of synthesis and characterization. *Russ. Chem. Rev.* **85**(3), 280–307 (2016)
7. Stroppa, A., Barone, P., Jain, P., Perez-Mato, J.M., Picozziet, S.: Hybrid improper ferroelectricity in a multiferroic and magnetoelectric metal–organic framework. *Adv. Mater.* **25**(16), 2284–2290 (2013)
8. Ferrey, G.: Hybrid porous solids: past, present, future. *Chem. Soc. Rev.* **37**(1), 191–214 (2008)
9. Gibbons, N., Baumberg, J.: Optical minibands in metallodielectric superlattices. *Phys. Rev. B* **85**(16), 165422 (2012)
10. Bogdanov, A.A., Suris, R.A.: Effect of the anisotropy of a conducting layer on the dispersion law of electromagnetic waves in layered metal–dielectric structures. *JETP Lett.* **96**, 49–55 (2012)
11. Zadumkin, S.N.: A new version of the statistical electronic theory of surface tension of metals. *Fiz. Met. Metalloved.* **11**(3), 331–346 (1961)
12. Apekov, A.M., Shebzukhova, I.G.: Polarization correction to the interfacial energy of faces of alkali metal crystals at the borders with a nonpolar organic liquid. *Bull. Russ. Acad. Sci. Phys.* **82**(7), 789–792 (2018). <https://doi.org/10.3103/S1062873818070067>
13. Shebzukhova, I.G., Apekov, A.M.: Contribution of dispersion interaction of s-spheres into the interface energy of a-Li and a-Na crystals bounding to non-polar organic liquids. *Phys. Chem. Asp. Study Clusters Nanostruct. Nanomater.* **9**, 518–521 (2017)

14. Apekov, A.M., Shebzukhova, I.G.: Temperature contribution to the interfacial energy of the crystals faces of Sc, α -Ti and α -Co at the boundary with the organic liquids. Phys. Chem. Asp. Study Clusters Nanostruct. Nanomater. **8**, 19–25 (2016)
15. Shebzukhova, I.G., Apekov, A.M., Khokonov, K.B.: Orientation dependence of the interfacial energies of chromium and α -iron crystals at boundaries with nonpolar organic liquids. Bull. Russ. Acad. Sci. Phys. **81**(5), 605–607 (2017). <https://doi.org/10.3103/S1062873817050173>
16. Apekov, A.M., Shebzukhova, I.G.: Interface energy of crystal faces of IIA-type metals at boundaries with nonpolar organic liquids, allowing for dispersion and polarization corrections. Bull. Russ. Acad. Sci. Phys. **83**(6), 760–763 (2019)
17. Shebzukhova, I.G., Apekov, A.M., Khokonov, K.B.: Anisotropy of the interface energy of IA and IB metals at a boundary with organic liquids. Bull. Russ. Acad. Sci. Phys. **80**(6), 657–659 (2016). <https://doi.org/10.3103/S1062873816060307>
18. Shebzukhova, I.G., Apekov, A.M., Khokonov, K.B.: Interface energy of faces of manganese and vanadium crystals at boundaries with organic liquids. Bull. Russ. Acad. Sci. Phys. **79**(6), 749–751 (2015). <https://doi.org/10.3103/S1062873815060295>
19. Shebzukhova, I.G., Apekov, A.M., Khokonov, K.B.: Effect of an organic liquid on the surface energy of scandium and titanium. Bull. Russ. Acad. Sci. Phys. **78**(8), 804–806 (2014). <https://doi.org/10.3103/S1062873814080334>
20. Apekov, A.M., Shebzukhova, I.G.: Of the facets at the boundary between calcium/barium crystals and nonpolar organic liquids. Phys. Chem. Asp. Study Clusters Nanostruct. Nanomater. **10**, 20–26 (2018)
21. Smoluchowski, R.: Anisotropy of the electronic work function of metals. Phys. Rev. **60**(1), 661–674 (1941)
22. Gombash, P.: Statistical theory of the atom and its applications (1951)



Difference Method for Solving the Dirichlet Problem for a Multidimensional Integro-Differential Equation of Convection-Diffusion

Zaryana Beshtokova^(✉)

North-Caucasus Center for Mathematical Research, North-Caucasus Federal
University, 1 Pushkin Str., 355017 Stavropol, Russia
zarabaeva@yandex.ru

Abstract. The work is devoted to a numerical method for solving the Dirichlet problem for a multidimensional integro-differential convection-diffusion equation with variable coefficients. Using the method of energy inequalities for solving the first initial-boundary value problem, a priori estimates are obtained in differential and difference interpretations. The obtained estimates imply the uniqueness and stability of the solution of the original differential problem with respect to the right-hand side and initial data, as well as the convergence of the solution of the difference problem to the solution of the original differential problem at a rate of $O(|h| + \tau)$. For an approximate solution of the differential problem, an algorithm for the numerical solution was constructed, and numerical calculations of test examples were carried out, illustrating the theoretical calculations obtained.

Keywords: first initial-boundary value problem · a priori estimate · dif-ference scheme · parabolic equation · integro-differential equation · Dirichlet problem

1 Introduction

In the study of applied problems of continuum mechanics, heat and mass transfer, methods of mathematical modeling and computational mathematics are widely used. The diffusion transfer of one or another substance and the transfer caused by the movement of the medium, i.e., convective transfer, can be distinguished as the main ones in the study of many processes in moving media. In gas and hydrodynamics, one of the basic models of many processes is a boundary value problem for nonstationary convection-diffusion equations (i.e., second-order parabolic equations with lower terms) [1].

The work is supported by North-Caucasus Center for Mathematical Research under agreement N^o. 075-02-2023-938 with the Ministry of Science and Higher Education of the Russian Federation.

The work is devoted to a numerical method for solving the Dirichlet problem for a multidimensional integro-differential convection-diffusion equation with variable coefficients. Using the method of energy inequalities for solving the first initial-boundary value problem, a priori estimates are obtained in differential and difference interpretations. The obtained estimates imply the uniqueness and stability of the solution of the original differential problem with respect to the right-hand side and initial data, as well as the convergence of the solution of the difference problem to the solution of the original differential problem at a rate of $O(|h| + \tau)$.

Problems of this kind arise when describing the mass distribution function of drops and ice particles, taking into account the microphysical processes of condensation, coagulation (combining small drops into large aggregates), crushing and freezing of drops in convective clouds. This work is a continuation of a series of works by the author on the study of various boundary value problems for multidimensional differential equations of parabolic type with variable coefficients [2]–[3].

2 Materials and Methods

2.1 Statement of the Boundary Value Problem and a Priori Estimate in Differential Form

In a cylinder $\overline{Q}_T = \overline{G} \times [0 \leq t \leq T]$, the base of which is a p -dimensional rectangular parallelepiped $\overline{G} = \{x = (x_1, x_2, \dots, x_p) : 0 \leq x_\alpha \leq l_\alpha, \alpha = 1, 2, \dots, p\}$ with boundary Γ , $\overline{G} = G \cup \Gamma$, consider the problem

$$\frac{\partial u}{\partial t} = Lu + f(x, t), \quad (x, t) \in Q_T, \quad (1)$$

$$u|_{\Gamma} = 0, \quad 0 \leq t \leq T, \quad (2)$$

$$u(x, 0) = u_0(x), \quad x \in \overline{G}, \quad (3)$$

where $Lu = \sum_{\alpha=1}^p L_\alpha u$,

$$L_\alpha u = \frac{\partial}{\partial x_\alpha} \left(k_\alpha(x, t) \frac{\partial u}{\partial x_\alpha} \right) + r_\alpha(x, t) \frac{\partial u}{\partial x_\alpha} - q_\alpha(x, t) u - \int_0^{l_\alpha} \rho_\alpha(x, t) u dx_\alpha, \quad (4)$$

$$0 < c_0 \leq k_\alpha(x, t) \leq c_1, \quad |r_\alpha(x, t)|, \quad |q_\alpha(x, t)|, \quad |\rho_\alpha(x, t)| \leq c_2, \quad (4)$$

$$Q_T = G \times (0 < t \leq T], \quad \alpha = \overline{1, p}, \quad c_0, c_1, c_2 = \text{const} > 0.$$

We will assume that the coefficients of the equation and the boundary conditions of the problem (1)–(3) satisfy the conditions necessary in the course of the presentation, which ensure the desired smoothness of the solution $u(x, t)$ in the cylinder Q_T .

We will also use positive constants M_i ($i = 1, 2, \dots$) depending only on the input data of the original problem (1)–(3).

Assuming the existence of a regular solution of the differential problem (1)–(3) in the cylinder \bar{Q}_T , we obtain an a priori estimate, for which we use the method of energy inequalities. We multiply Eq. (1) scalarly by u and obtain the energy identity

$$\begin{aligned} \left(\frac{\partial u}{\partial t}, u \right) &= \left(\sum_{\alpha=1}^p \frac{\partial}{\partial x_\alpha} \left(k_\alpha(x, t) \frac{\partial u}{\partial x_\alpha} \right), u \right) + \left(\sum_{\alpha=1}^p r_\alpha(x, t) \frac{\partial u}{\partial x_\alpha}, u \right) - \\ &- \left(\sum_{\alpha=1}^p q_\alpha(x, t) u, u \right) - \left(\sum_{\alpha=1}^p \int_0^{l_\alpha} \rho_\alpha(x, \tau) u(x, t) dx_\alpha, u \right) + \left(f(x, t), u \right). \end{aligned} \quad (5)$$

We introduce the scalar product and the norm in the following form:

$$(u, v) = \int_G uv \, dx, \quad \|u\|_0^2 = \int_G u^2 \, dx, \quad u_x^2 = \sum_{\alpha=1}^p (u_{x_\alpha})^2.$$

Taking into account the obtained transformations, from (5), taking into account (2), we obtain an inequality:

$$\begin{aligned} \left(\frac{\partial u}{\partial t}, u \right) &= \int_G u \frac{\partial u}{\partial t} \, dx = \frac{1}{2} \frac{\partial}{\partial t} \|u\|_0^2, \\ \left(\sum_{\alpha=1}^p \frac{\partial}{\partial x_\alpha} \left(k_\alpha(x, t) \frac{\partial u}{\partial x_\alpha} \right), u \right) &= - \sum_{\alpha=1}^p \int_G k_\alpha(x, t) \left(\frac{\partial u}{\partial x_\alpha} \right)^2 \, dx. \end{aligned}$$

Using the ε -Cauchy inequality, we estimate the terms on the right-hand side

$$\begin{aligned} \left(\sum_{\alpha=1}^p r_\alpha(x, t) \frac{\partial u}{\partial x_\alpha}, u \right) &\leq \sum_{\alpha=1}^p \left(r_\alpha(x, t) \frac{\partial u}{\partial x_\alpha}, u \right) \leq \varepsilon \|u_x\|_0^2 + M_1(\varepsilon) \|u\|_0^2, \\ - \left(\sum_{\alpha=1}^p \int_0^{l_\alpha} \rho_\alpha(x, \tau) u(x, t) dx_\alpha, u \right) &\leq \left(\frac{1}{2}, u^2 \right) + \\ &+ \left(\frac{1}{2}, \left(\sum_{\alpha=1}^p \int_0^{l_\alpha} \rho_\alpha(x, \tau) u(x, t) dx_\alpha \right)^2 \right) \leq \\ &\leq \frac{1}{2} \|u\|_0^2 + M_1 \left(1, \sum_{\alpha=1}^p \int_0^{l_\alpha} u^2(x, t) dx_\alpha \right) \leq M_2 \|u\|_0^2, \\ - \left(\sum_{\alpha=1}^p q_\alpha(x, t) u, u \right) &\leq c_2 \|u\|_0^2, \\ \left(f(x, t), u \right) &\leq \frac{1}{2} \|f\|_0^2 + \frac{1}{2} \|u\|_0^2, \end{aligned}$$

where $G' = \{x' = (x_1, x_2, \dots, x_{\alpha-1}, x_{\alpha+1}, \dots, x_p) : 0 < x_k < l_k, k = 1, 2, \dots, \alpha-1, \alpha+1, \dots, p\}$, $dx' = dx_1 dx_2 \cdots dx_{\alpha-1} dx_{\alpha+1} \cdots dx_p$, $x = (x_1, x_2, \dots, x_{\alpha-1}, x_\alpha, x_{\alpha+1}, \dots, x_p) = (x_\alpha, x')$.

Taking into account the obtained transformations and (2), from (5) we obtain the inequality

$$\frac{1}{2} \frac{\partial}{\partial t} \|u\|_0^2 + \sum_{\alpha=1}^p \int_G k_\alpha(x, t) \left(\frac{\partial u}{\partial x_\alpha} \right)^2 dx \leq \varepsilon \|u_x\|_0^2 + M_3(\varepsilon) \|u\|_0^2 + \frac{1}{2} \|f\|_0^2. \quad (6)$$

We integrate (6) over τ from 0 to t , then we get

$$\begin{aligned} \|u\|_0^2 + \int_0^t \|u_x\|_0^2 d\tau &\leq \varepsilon M_4 \int_0^t \|u_x\|_0^2 d\tau + M_5(\varepsilon) \int_0^t \|u\|_0^2 d\tau + \\ &+ M_6 \left(\int_0^t \|f\|_0^2 d\tau + \|u_0(x)\|_0^2 \right). \end{aligned} \quad (7)$$

Choosing $\varepsilon = \frac{1}{2M_4}$, from (7) we find

$$\|u\|_0^2 + \int_0^t \|u_x\|_0^2 d\tau \leq M_7 \int_0^t \|u\|_0^2 d\tau + M_8 \left(\int_0^t \|f\|_0^2 d\tau + \|u_0(x)\|_0^2 \right). \quad (8)$$

Based on the Gronwall lemma (Lemma 1.1 [4, p. 152]), from (8) we obtain the inequality

$$\|u\|_0^2 + \int_0^t \|u_x\|_0^2 d\tau \leq M(T) \left(\int_0^t \|f\|_0^2 d\tau + \|u_0(x)\|_0^2 \right), \quad (9)$$

where $M(T)$ depends only on the input data of the original problem (1)–(3).

3 Constructing a Difference Scheme

In a closed cylinder \bar{Q}_T we introduce a uniform grid [5]:

$$\bar{\omega}_{h\tau} = \bar{\omega}_h \times \bar{\omega}_\tau = \{(x_i, t_j), x \in \bar{\omega}_h, t \in \bar{\omega}_\tau\},$$

$$\bar{\omega}_h = \prod_{\alpha=1}^p \bar{\omega}_{h_\alpha}, \quad \bar{\omega}_{h_\alpha} = \{x^{i_\alpha} = i_\alpha h_\alpha, i_\alpha = 0, 1, \dots, N_\alpha, N_\alpha h_\alpha = l_\alpha\},$$

$$\bar{\omega}_\tau = \{t_j = j\tau, j = 0, 1, \dots, m, m\tau = T\}.$$

On the uniform grid $\bar{\omega}_{h\tau}$, we associate the differential problem (1)–(3) with the difference scheme of the first order of approximation in h and τ :

$$y_{\bar{t}} = \Lambda(\bar{t})y + \varphi, \quad (x, t) \in \omega_{h\tau}, \quad (10)$$

$$y|_{\gamma_h} = 0, \quad t \in \bar{\omega}_\tau, \quad (11)$$

$$y(x, 0) = u_0(x), \quad x \in \bar{\omega}_h, \quad (12)$$

where $\Lambda(\bar{t}) = \sum_{\alpha=1}^p \Lambda_\alpha(\bar{t})$,

$$\Lambda_\alpha(\bar{t})y = (a_\alpha y_{\bar{x}_\alpha})_{x_\alpha} + r_\alpha^+ y_{x_\alpha} + r_\alpha^- y_{\bar{x}_\alpha} - d_\alpha y - \sum_{i_\alpha=1}^{N_\alpha} p_{i_\alpha}^\alpha y_{i_\alpha}^{j+\frac{\alpha}{p}} h_\alpha,$$

$$y_{\bar{x}_\alpha} = \frac{y_i - y_{i-1}}{h_\alpha}, \quad y_{x_\alpha} = \frac{y_{i+1} - y_i}{h_\alpha}, \quad a_\alpha^{(+1_\alpha)} = a_{\alpha, i_\alpha+1}, \quad y_{\bar{t}} = \frac{y^j - y^{j-1}}{\tau},$$

$$y = y^j, \quad \bar{y} = y^{j-1}, \quad r_\alpha = r_\alpha^+ + r_\alpha^-, \quad |r_\alpha| = r_\alpha^+ - r_\alpha^-, \quad r_\alpha^+ = 0.5(r_\alpha + |r_\alpha|) \geq 0, \\ r_\alpha^- = 0.5(r_\alpha - |r_\alpha|) \leq 0, \quad x^{-0.5\alpha} = x_1, \dots, x_{\alpha-1}, x_\alpha - 0.5h_\alpha, x_{\alpha+1}, \dots, x_p,$$

$$\bar{t} = (j + 0.5)\tau = t_j + 0.5\tau = t_{j+0.5}, \quad t_{j+\frac{\alpha}{p}} = t_j + \frac{\alpha\tau}{p} = \left(j + \frac{\alpha}{p} \right) \tau,$$

$$a_\alpha = k_\alpha(x^{-0.5\alpha}, \bar{t}_j), \quad r_\alpha = r_\alpha(x, \bar{t}_j), \quad d_\alpha = q_\alpha(x, \bar{t}_j), \quad p^\alpha = \rho_\alpha(x, \bar{t}_j),$$

$\varphi = f(x, t_j)$, τ, h are steps of grid.

4 Stability and Convergence of the Difference Scheme

To solve problem (10)–(12), we obtain an a priori estimate with the help of the method of energy inequalities. We introduce the scalar product in the following form:

$$\begin{aligned}
 (u, v) &= \sum_{x \in \omega_h} u(x)v(x)h_1h_2 \cdots h_p = \\
 &= \sum_{i_1=1}^{N_1-1} \sum_{i_2=1}^{N_2-1} \cdots \sum_{i_p=1}^{N_p-1} u(i_1h_1, i_2h_2, \dots, i_ph_p)v(i_1h_1, i_2h_2, \dots, i_ph_p)h_1h_2 \cdots h_p; \\
 (u, v)_\alpha &= \sum_{i_1=1}^{N_1-1} \sum_{i_2=1}^{N_2-1} \cdots \sum_{i_\alpha=1}^{N_\alpha-1} \cdots \sum_{i_p=1}^{N_p-1} u(x)v(x)h_1h_2 \cdots h_p = \\
 &= \sum_{i_\beta \neq i_\alpha} \left(\sum_{i_\alpha=1}^{N_\alpha-1} u(x)v(x)h_\alpha \right) H/h_\alpha, \\
 (u, v]_\alpha &= \sum_{i_1=1}^{N_1-1} \sum_{i_2=1}^{N_2-1} \cdots \sum_{i_\alpha=1}^{N_\alpha} \cdots \sum_{i_p=1}^{N_p-1} u(x)v(x)h_1h_2 \cdots h_p = \\
 &= \sum_{i_\beta \neq i_\alpha} \left(\sum_{i_\alpha=1}^{N_\alpha} u(x)v(x)h_\alpha \right) H/h_\alpha, \quad H = \prod_{\alpha=1}^p h_\alpha.
 \end{aligned}$$

In the function space, we define the norms and introduce them in the form:

$$\begin{aligned}
 (u, u) &= \|u\|^2, \quad (u, u] = \|u\|^2, \quad (u, v] = \sum_{\alpha=1}^p (u, v]_\alpha, \quad (u, v) = \sum_{\alpha=1}^p (u, v)_\alpha, \\
 \|Y_{\bar{x}}\|^2 &= \sum_{\alpha=1}^p \|Y_{\bar{x}_\alpha}\|^2.
 \end{aligned}$$

Let us now multiply the difference Eq. (10) scalarly by $2\tau y$:

$$2\tau(y_{\bar{t}}, y) = 2\tau(\Lambda(\tilde{t})y, y) + 2\tau(\varphi, y). \quad (13)$$

We transform the sums included in the identity (13), taking into account the conditions (11) and the formula $2zz_{\bar{t}} = (z^2)_{\bar{t}} + \tau(z_{\bar{t}})^2$:

$$2\tau(y_{\bar{t}}, y) = (1, y^2) - (1, \check{y}^2) + \tau^2(1, y_{\bar{t}}^2), \quad (14)$$

$$\begin{aligned}
 (\Lambda(\tilde{t})y, y) &= \left(\sum_{\alpha=1}^p \Lambda_\alpha(\tilde{t})y, y \right) = \sum_{\alpha=1}^p \left(\Lambda_\alpha(\tilde{t})y, y \right)_\alpha = \sum_{\alpha=1}^p \left(\left((a_\alpha y_{\bar{x}_\alpha})_{x_\alpha}, y \right)_\alpha + \right. \\
 &\quad \left. + (r_\alpha^+ y_{x_\alpha}, y)_\alpha + (r_\alpha^- y_{\bar{x}_\alpha}, y)_\alpha - (d_\alpha y, y)_\alpha - \left(\sum_{i_\alpha=1}^{N_\alpha} p_{i_\alpha}^\alpha y_{i_\alpha}^{j+\frac{\alpha}{p}} h_\alpha, y \right)_\alpha \right). \quad (15)
 \end{aligned}$$

Applying the first difference Green's formula in (15) and substituting the expressions transformed in this way into identity (13), taking into account (14), we find

$$\begin{aligned}
 &(1, y^2) - (1, \check{y}^2) + \tau^2(1, y_{\bar{t}}^2) + 2\tau \sum_{\alpha=1}^p (a_\alpha, y_{\bar{x}_\alpha}^2]_\alpha = \\
 &= 2\tau \sum_{\alpha=1}^p (r_\alpha^+ y_{x_\alpha}, y)_\alpha + 2\tau \sum_{\alpha=1}^p (r_\alpha^- y_{\bar{x}_\alpha}, y)_\alpha - 2\tau \sum_{\alpha=1}^p (d_\alpha y, y)_\alpha -
 \end{aligned}$$

$$- 2\tau \sum_{\alpha=1}^p \left(\sum_{i_\alpha=1}^{N_\alpha} p_{i_\alpha}^\alpha y_{i_\alpha}^{j+\frac{\alpha}{p}} h_\alpha, y \right)_\alpha + 2\tau(\varphi, y). \quad (16)$$

Using the Cauchy ε -inequality and Lemma 1 from [6], we estimate the sums included in the identity (16):

$$\begin{aligned} \sum_{\alpha=1}^p (a_\alpha, y_{\bar{x}_\alpha}^2)_\alpha &\geq c_0 \sum_{\alpha=1}^p (1, y_{\bar{x}_\alpha}^2)_\alpha = c_0(1, y_{\bar{x}}^2) = c_0 \|y_{\bar{x}}\|^2, \\ \sum_{\alpha=1}^p (r_\alpha^+ y_{x_\alpha}, y)_\alpha + \sum_{\alpha=1}^p (r_\alpha^- y_{\bar{x}_\alpha}, y)_\alpha &\leq 2c_2 \|y_{\bar{x}}\| \|y\| \leq c_2 \left(\varepsilon \|y_{\bar{x}}\|^2 + \frac{1}{4\varepsilon} \|y\|^2 \right), \\ - \sum_{\alpha=1}^p (d_\alpha y, y)_\alpha &\leq c_2 \|y\|^2, \\ -2 \sum_{\alpha=1}^p \left(\sum_{i_\alpha=1}^{N_\alpha} p_{i_\alpha}^\alpha y_{i_\alpha}^{j+\frac{\alpha}{p}} h_\alpha, y \right)_\alpha &\leq (1, y^2) + \sum_{\alpha=1}^p \left(1, \left(\sum_{i_\alpha=1}^{N_\alpha} p_{i_\alpha}^\alpha y_{i_\alpha}^{j+\frac{\alpha}{p}} h_\alpha \right)_\alpha^2 \right) \leq \\ &\leq \|y\|^2 + M_1 \sum_{\alpha=1}^p \left(1, \sum_{i_\alpha=1}^{N_\alpha} \left(y_{i_\alpha}^{j+\frac{\alpha}{p}} \right)^2 h_\alpha \right) \leq M_2 \|y\|^2, \\ (\varphi, y) &\leq \frac{1}{2} \|y\|^2 + \frac{1}{2} \|\varphi\|^2. \end{aligned}$$

Taking into account the obtained estimates, after simple transformations from (16) we find

$$\|y\|^2 - \|\check{y}\|^2 + 2\tau c_0 \|y_{\bar{x}}\|^2 + \tau^2 \|y_{\bar{t}}\|^2 \leq \varepsilon M_3 \|y_{\bar{x}}\|^2 \tau + M_4(\varepsilon) \|y\|^2 \tau + M_5 \|\varphi\|^2 \tau. \quad (17)$$

Choosing $\varepsilon = \frac{c_0}{M_3}$, from (17) we find the inequality

$$\|y\|^2 - \|\check{y}\|^2 + \|y_{\bar{x}}\|^2 \tau \leq M_6 \|y\|^2 \tau + M_7 \|\varphi\|^2 \tau. \quad (18)$$

Summing (18) over j' from 1 to j , we get

$$\|y^j\|^2 + \sum_{j'=1}^j \|y_{\bar{x}}\|^2 \tau \leq M_6 \sum_{j'=1}^j \|y^{j'}\|^2 \tau + M_7 \left(\sum_{j'=1}^j \|\varphi^{j'}\|^2 \tau + \|y^0\|^2 \right).$$

From the last inequality we have

$$\begin{aligned} \|y^j\|^2 + \sum_{j'=1}^j \|y_{\bar{x}}\|^2 \tau &\leq M_6 \|y^j\|^2 \tau + M_6 \sum_{j'=1}^{j-1} \|y^{j'}\|^2 \tau + \\ &+ M_7 \left(\sum_{j'=1}^j \|\varphi^{j'}\|^2 \tau + \|y^0\|^2 \right). \end{aligned} \quad (19)$$

Choosing $\tau_0 = \frac{1}{2M_6}$, from (19) we get that for all $\tau \leq \tau_0$ the inequality holds

$$\|y^j\|^2 + \sum_{j'=1}^j \|y_{\bar{x}}\|^2 \tau \leq M_8 \sum_{j'=1}^{j-1} \|y^{j'}\|^2 \tau + M_9 \left(\sum_{j'=1}^j \|\varphi^{j'}\|^2 \tau + \|y^0\|^2 \right). \quad (20)$$

Based on an analogue of the Gronwall lemma [7, p. 171] for the grid function from the inequality (20) we obtain the a priori estimate

$$\|y^j\|^2 + \sum_{j'=1}^j \|y_{\bar{x}}\|^2 \tau \leq M(T) \left(\sum_{j'=1}^j \|\varphi^{j'}\|^2 \tau + \|y^0\|^2 \right), \quad (21)$$

where $M(T) = \text{const} > 0$, independent of $|h|$ and τ .

Let $u(x, t)$ is a solution to the problem (1)–(3), $y(x_i, t_j) = y_i^j$ is a solution to the difference problem (10)–(12). Denote by $z_i^j = y_i^j - u_i^j$ approximation error, where $u_i^j = u(x_i, t_j)$. Then, substituting $y = z + u$ into (10)–(12), we get the problem for z :

$$z_{\bar{t}} = A(\bar{t})z + \Psi, \quad (x, t) \in \omega_{h\tau}, \quad (22)$$

$$z|_{\gamma_h} = 0, \quad (23)$$

$$z(x, 0) = 0, \quad x \in \bar{\omega}_h, \quad (24)$$

where $\Psi = O(|h| + \tau)$ is the approximation error on the solution of the problem (1)–(3).

Using the a priori estimate (21) for the solution of the problem (22)–(24), we obtain

$$\|z^j\|^2 + \sum_{j'=1}^j \|z_{\bar{x}}\|^2 \tau \leq M \sum_{j'=1}^j \|\psi^{j'}\|^2 \tau, \quad (25)$$

where $M = \text{const} > 0$, independent of $|h|$ and τ .

5 Algorithm for the Numerical Solution of the Problem

We rewrite the problem (1)–(3) at $0 \leq x_\alpha \leq l_\alpha$, $\alpha = 1, 2$, $p = 2$, then we get

$$\begin{aligned} \frac{\partial u}{\partial t} &= \frac{\partial}{\partial x_1} \left(k_1(x_1, x_2, t) \frac{\partial u}{\partial x_1} \right) + \frac{\partial}{\partial x_2} \left(k_2(x_1, x_2, t) \frac{\partial u}{\partial x_2} \right) + \\ &+ r_1(x_1, x_2, t) \frac{\partial u}{\partial x_1} + r_2(x_1, x_2, t) \frac{\partial u}{\partial x_2} - q_1(x_1, x_2, t)u - \\ &q_2(x_1, x_2, t)u - \int_0^{l_1} \rho_1(x, t)u dx_1 - \int_0^{l_2} \rho_2(x, t)u dx_2 + f(x_1, x_2, t), \end{aligned} \quad (26)$$

$$\begin{cases} u(0, x_2, t) = u(l_1, x_2, t) = 0, & 0 \leq t \leq T, \\ u(x_1, 0, t) = u(x_1, l_2, t) = 0, & 0 \leq t \leq T, \end{cases} \quad (27)$$

$$u(x_1, x_2, 0) = u_0(x_1, x_2). \quad (28)$$

Consider a grid $x_\alpha^{(i_\alpha)} = i_\alpha h_\alpha$, $\alpha = 1, 2$, $t_j = j\tau$, where $i_\alpha = 0, 1, \dots, N_\alpha$, $h_\alpha = l_\alpha/N_\alpha$, $j = 0, 1, \dots, m$, $\tau = T/m$. Introduce one fractional step $t_{j+\frac{1}{2}} = t_j + 0.5\tau$.

Denote by $y_{i_1, i_2}^{j+\frac{k}{2}} = y^{j+\frac{k}{2}} = y(i_1 h_1, i_2 h_2, (j + 0.5k)\tau)$, $k = 0, 1$, $j = 0, 1, 2, \dots, j_0$ grid function.

For the numerical solution of the two-dimensional problem (26)–(28), we construct a locally one-dimensional scheme, then we have

$$\begin{cases} \frac{y^{j+\frac{1}{2}} - y^j}{\tau} = \tilde{A}_1 y^{j+\frac{1}{2}} + \varphi_1, \\ \frac{y^{j+1} - y^{j+\frac{1}{2}}}{\tau} = \tilde{A}_2 y^{j+1} + \varphi_2, \end{cases} \quad (29)$$

$$\begin{cases} y_{0, i_2}^{j+\frac{1}{2}} = y_{N_1, i_2}^{j+\frac{1}{2}} = 0, \\ y_{i_1, 0}^{j+1} = y_{i_1, N_2}^{j+1} = 0, \end{cases} \quad (30)$$

$$y_{i_1, i_2}^0 = u_0(i_1 h_1, i_2 h_2), \quad (31)$$

$$A_\alpha(\tilde{t})y^{j+\frac{\alpha}{p}} = \left(a_\alpha y_{\bar{x}_\alpha}^{j+\frac{\alpha}{p}}\right)_{x_\alpha} + r_\alpha^+ y_{x_\alpha}^{j+\frac{\alpha}{p}} + r_\alpha^- y_{\bar{x}_\alpha}^{j+\frac{\alpha}{p}} - d_\alpha y^{j+\frac{\alpha}{p}} - \sum_{i_\alpha=0}^{N_\alpha} p_\alpha y_{i_\alpha}^{j+\frac{\alpha-1}{p}} \bar{h}_\alpha,$$

$$\varphi_\alpha = \frac{1}{2}f(x_1, x_2, t_{j+0.5\alpha}) \text{ or } \varphi_1 = 0, \varphi_2 = f(x_1, x_2, t_{j+1}), \alpha = 1, 2,$$

Let us present the calculation formulas for solving scheme (29)–(31).

At the first stage, we find the solution $y_{i_1, i_2}^{j+\frac{1}{2}}$. To do this, for each value of $i_2 = \overline{1, N_2 - 1}$, the problem is solved

$$A_{1(i_1, i_2)} y_{i_1-1, i_2}^{j+\frac{1}{2}} - C_{1(i_1, i_2)} y_{i_1, i_2}^{j+\frac{1}{2}} + B_{1(i_1, i_2)} y_{i_1+1, i_2}^{j+\frac{1}{2}} = -F_{1(i_1, i_2)}^{j+\frac{1}{2}}, \quad 0 < i_1 < N_1, \quad (32)$$

$$y_{0, i_2}^{j+\frac{1}{2}} = y_{N_1, i_2}^{j+\frac{1}{2}} = 0,$$

where

$$A_{1(i_1, i_2)} = \frac{(a_1)_{i_1, i_2}}{h_1^2} - \frac{(r_1^-)_{i_1, i_2}}{h_1}, \quad B_{1(i_1, i_2)} = \frac{(a_1)_{i_1+1, i_2}}{h_1^2} + \frac{(r_1^+)_{i_1, i_2}}{h_1},$$

$$C_{1(i_1, i_2)} = A_{1(i_1, i_2)} + B_{1(i_1, i_2)} + \frac{1}{\tau} + \frac{1}{p}(d_1)_{i_1, i_2},$$

$$F_{1(i_1, i_2)}^{j+\frac{1}{2}} = \frac{1}{\tau} y_{i_1, i_2}^j + \varphi_{1(i_1, i_2)} - \sum_{i_1=0}^{N_1} p_1 y_{i_1, i_2}^j \bar{h}_1,$$

At the second stage, we find the solution y_{i_1, i_2}^{j+1} . For this, as in the first case, the problem is solved for each value of $i_1 = \overline{1, N_1 - 1}$

$$A_{2(i_1, i_2)} y_{i_1, i_2-1}^{j+1} - C_{2(i_1, i_2)} y_{i_1, i_2}^{j+1} + B_{2(i_1, i_2)} y_{i_1, i_2+1}^{j+1} = -F_{2(i_1, i_2)}^{j+1}, \quad 0 < i_2 < N_2, \quad (33)$$

$$y_{i_1,0}^{j+1} = y_{i_1,N_2}^{j+1} = 0,$$

where

$$\begin{aligned} A_{2(i_1,i_2)} &= \frac{(a_2)_{i_1,i_2}}{h_2^2} - \frac{(r_2^-)_{i_1,i_2}}{h_2}, & B_{2(i_1,i_2)} &= \frac{(a_2)_{i_1,i_2+1}}{h_2^2} + \frac{(r_2^+)_{i_1,i_2}}{h_2}, \\ C_{2(i_1,i_2)} &= A_{2(i_1,i_2)} + B_{2(i_1,i_2)} + \frac{1}{\tau} + \frac{1}{p}(d_2)_{i_1,i_2}, \\ F_{2(i_1,i_2)}^{j+1} &= \frac{1}{\tau} y_{i_1,i_2}^{j+\frac{1}{2}} + \varphi_{2(i_1,i_2)} - \sum_{i_2=0}^{N_2} p_2 y_{i_1,i_2}^j \bar{h}_2, \end{aligned}$$

As can be seen, the non-local (integral) source in Eq. (1) leads to a violation of the tridiagonal structure of the coefficient matrix of the difference scheme. Therefore, to solve the obtained systems of difference equations (32), (33) by the sweep method [5, p. 40] it is possible to obtain a tridiagonal structure of the coefficient matrix of difference schemes, provided we take the value of the desired grid function from the lower layer when calculating the sum in Eq. (10).

6 Results

Theorem 1. *Let conditions (4) be satisfied, then the a priori estimate (9) is valid for the solution of problem (1)–(3).*

The a priori estimate (9) implies the uniqueness of the solution of the original problem (1)–(3) and continuous dependence of the problem solution on the input data at each time layer in the $\|u\|_1^2 = \|u\|_0^2 + \int_0^t \|u_x\|_0^2 d\tau$ norm.

Theorem 2. *Let conditions (4) be satisfied, then in the class of sufficiently smooth coefficients of the equation and boundary conditions for the solution of the difference problem (10)–(12) for sufficiently small $\tau \leq \tau_0(c_0, c_1, c_2)$, a priori estimate (21) is valid at each time layer in the grid norm $\|y^j\|_1^2 = \|y^j\|^2 + \sum_{j'=1}^j \|y_{\bar{x}}\|^2 \tau$.*

The a priori estimate implies the uniqueness and stability of the solution of the difference problem (10)–(12) with respect to the right-hand side and the initial data on the layer, as well as the convergence of the scheme (10)–(12) with the rate $O(|h| + \tau)$ in the grid norm $\|z^j\|_1$, where $|h| = h_1 + h_2 + \dots + h_p$.

Comment. To solve the resulting systems of difference Eqs. (32), (33), iterative solution methods can also be used. When solving the resulting algebraic system $Ay = f$ by the iterative method, where A is the coefficient matrix of the system of difference equations, it is necessary that the following conditions for the convergence of the iterative process should be satisfied:

$$\sum_{j=1, j \neq i}^N \left| \frac{a_{ij}}{a_{ii}} \right| < 1, \quad \|Ay_k - f\| \leq \varepsilon \|Ay_0 - f\|, \quad \|y_k - y\| \rightarrow 0, \quad k \rightarrow \infty.$$

Table 1. The error in the norm $\|\cdot\|_{L_2(\bar{w}_{h\tau})}$ when decreasing grid size for problem (1)–(3)

\bar{h}	Maximum error	CO
1/80	0.006416155	0.389613492
1/160	0.004430162	0.534349581
1/320	0.002775766	0.674473568
1/640	0.001609425	0.786340645
1/1280	0.000883100	0.865896625
1/2560	0	0.865896625

Table 2. The error in the norm $\|\cdot\|_{C(\bar{w}_{h\tau})}$ when decreasing grid size for problem (1)–(3)

\bar{h}	Maximum error	CO
1/80	0.0203034746	0.305960241
1/160	0.0149247895	0.444016058
1/320	0.0097120049	0.581871473
1/640	0.0061415431	0.699166006
1/1280	0.0035404042	0.794687085
1/2560	0	0.794687085

7 Numerical Results

The coefficients of the equation and boundary conditions of problem (1)–(3) are selected so that the exact solution of the problem for $p = 2$ is the function

$$u(x, t) = t^3(x_1^3 - l_1x_1^2)(x_2^3 - l_2x_2^2).$$

In Tables 1,2 we present the maximum value of the error ($z=y-u$) and the computational order of convergence (CO) in the norms: $\|\cdot\|_{L_2(\bar{w}_{h\tau})}$ and $\|\cdot\|_{C(\bar{w}_{h\tau})}$, where $\|y\|_{C(\bar{w}_{h\tau})} = \max_{(x_i, t_j) \in \bar{w}_{h\tau}} |y|$, when $\bar{h} = h_1 = h_2 = \tau$, while the mesh size is decreasing. The error is being reduced in accordance with the order of approximation $O(\bar{h} + \tau)$.

The order of convergence is determined by the following formula $CO_1 = \log_2 \frac{\|z_1\|}{\|z_2\|}$, where z_1 and z_2 errors corresponding to steps $0, 5\bar{h}, \bar{h}$.

8 Conclusion

The work is devoted to a numerical method for solving the Dirichlet problem for a multidimensional integro-differential convection-diffusion equation with variable coefficients. Using the method of energy inequalities for solving the first initial-boundary value problem, a priori estimates are obtained in differential and difference interpretations. The obtained estimates imply the uniqueness and stability of

the solution of the original differential problem with respect to the right-hand side and initial data, as well as the convergence of the solution of the difference problem to the solution of the original differential problem at a rate of $O(|h| + \tau)$. For an approximate solution of the differential problem, an algorithm for the numerical solution was constructed, and numerical calculations of test examples were carried out, illustrating the theoretical calculations obtained.

References

1. Samarsky, A.A., Vabishchevich, P.N.: Numerical methods for solving problems of convection-diffusion. Editoreal URSS, Moscow (1999)
2. Beshtokova, Z.V., Shkhanukov-Lafishev, M.K.: Locally one-dimensional difference scheme for the third boundary value problem for a parabolic equation of the general form with a nonlocal source. *Differ. Eq.* **54**, 870–880 (2018)
3. Beshtokova, Z.V., Lafisheva, M.M., Shkhanukov-Lafishev, M.K.: Locally one-dimensional difference schemes for parabolic equations in media possessing memory. *Comput. Math. Math. Phys.* **58**(9), 1477–1488 (2018)
4. Ladyzhenskaya, O.A.: Boundary value problems of mathematical physics. Nauka, Moscow (1973)
5. Samarskiy, A.A.: *Teoriya raznostnykh skhem.* [Theory of difference schemes]. Moscow: Nauka (1983)
6. Andreev, V.B.: The convergence of difference schemes which approximate the second and third boundary value problems for elliptic equations. *USSR Comput. Math. Math. Phys.* **8**(6), 44–62 (1968)
7. Samarskii, A.A., Gulin, A.V.: *Stability of Difference Schemes.* Nauka, Moscow (1973)



The Problem of Restoring the Unit of Approximation in the Model for Studying Functional Dependence from Approximate Data

E. Yartseva¹ , L. Andruhiv¹ , and R. Abdulkadirov²

¹ North-Caucasus Federal University, Stavropol, Russia

² North-Caucasus Center for Mathematical Research, Stavropol, Russia
ruslanabdulkadirovstavropol@gmail.com

Abstract. In this paper, we demonstrate the application of the analytical apparatus of representing functions by their singular integrals for developing numerical methods by interpreting approximate data in the problem of correctly restoring the studied functional dependencies. Such approach significantly extends the substantive basis of the apparatus for approximating functions in problems, where it is necessary to build a model of the functional dependence, which depends on approximate data. The main goal of this article is to provide the restoring of the real-valued functions $f = \{f(x_0), f(x_1), \dots, f(x_m)\}$, associated with discrete sequence of points $\{x_k\}_{k=0}^m$ on $[a, b]$, using generalized kernel $(K_n f)(x)$. We will demonstrate the theoretical calculations of restoring of the unit approximation, which is able to increase the accuracy of approximations. In the end of our research, the application of proposed approximation is used for solving optimization problem, where rising of the accuracy with minimal time computations plays an important role.

Keywords: approximation of functions · singular integrals · regular algorithms · optimization problem · approximate unit

1 Introduction

It is well known, that one of the most dispersed operation on integrable functions is their convolution [1], such as $\int f(x-y)g(y)dy \triangleq f * g(x)$, where f and g in L_1 . Such operation meets in the harmonic analysis, because of $(f * g)^\wedge = \hat{f} \cdot \hat{g}$, where « \wedge » means a Fourier transformation. For operation « $*$ », observed as algebraic operation on the set of elements in L_1 , it is necessary to build the unit. The appropriate model in this case can be approximate unit, which is a sequence $\{K_n(x)\}_{n=1}^\infty$ of elements in L_1 , such that

$$\sup_n \|K_n\| < \infty, \quad \lim_{n \rightarrow \infty} \int_{\Omega} K_n(x) dx = 1, \quad \lim_{n \rightarrow \infty} \int_{\delta \leq |x| \leq |\Omega|} K_n(x) dx = 0, \quad (1)$$

In the last equation of (1) δ means any fixed number, which satisfy $0 < \delta < |\Omega|$. The notation «approximate unit» is defined, because there are the following limits.

$$\lim_{n \rightarrow \infty} \|K_n * f - f\| = 0, \quad f \in C(\Omega), \tag{2}$$

$$\lim_{n \rightarrow \infty} \|D(K_n * f) - Df\| = 0, \quad f \in C^{(1)}(\Omega). \tag{3}$$

Equations (2) and (3) belong to spaces $C^{(k)}$ and L_p . There is a significant private case of approximate unit, where elements of $\{K_n(x)\}_{n=1}^\infty$ are positive. This case is named as unit distribution, and Eq. (1) becomes

$$\lim_{n \rightarrow \infty} \int K_n(x) dx = 1, \quad \lim_{n \rightarrow \infty} \int_{\delta \leq |x| \leq |\Omega|} K_n(x) dx = 0. \tag{4}$$

The limits of sequences $\{K_n(x)\}_{n=1}^\infty$, which exist almost everywhere in Ω , are called singular integrals of summable functions [2]. The representation of functions by singular integrals can be effectively used in problems of applied analysis [3], constructive theory of functions [4], approximation theory and so on. In section “Methods and Materials” we present applications in the development of numerical methods of the interpretation by approximate data for correct recovery of studied functional dependence.

2 Materials and Methods

For greater clarity of the theory we present, as the initial sequence $\{K_n(x)\}_{n=1}^\infty$ generating the required distribution of unity, a sequence of functions of the form $\left\{ \frac{n}{\pi d} \cdot \frac{1}{n^2(x-y)^2 + d^2} \right\}_{n=1}^\infty$, where $d > 0$ is a certain parameter of the model kernel of the corresponding singular integral

$$(K_n f)(x) = \int_a^b K_n(x, y) f(y) dy = \int_a^b \frac{n}{\pi d} \cdot \frac{1}{n^2(x-y)^2 + d^2} f(y) dy, \quad n = 1, 2, \dots \tag{5}$$

Recall that such integral is called the Poisson integral of a given function $f(x)$ defined on the interval $\Omega = [a, b]$. For further convenience, we write it in the parametrized form

$$(K_\tau f)(x) = \int_a^b K_\tau(x, y) f(y) dy = \int_a^b \frac{\tau}{\pi} \cdot \frac{1}{n^2(x-y)^2 + \tau^2} f(y) dy, \tag{6}$$

where $\tau = \frac{d}{n}$ ($\tau \in (0, 1)$). According to the fact that the kernel of the integral (6) depends on the difference between the arguments x and y , by introducing new integration variable $t = x - y$, ($dy = -dt$, $t \in [-(b-x), x-a]$, $x \in [a, b]$), we can use the expression

$$(K_\tau f)(x) = \frac{\tau}{\pi} \int_{t_1(x)}^{t_2(x)} \frac{1}{t^2 + \tau^2} f(x-t) dt, \tag{7}$$

where $t_1(x) = -(b - x)$ and $t_2(x) = x - a$. It is clear that the integral operator K_τ is convolution operator. Since

$$\frac{\tau}{\pi} \int_{t_1(x)}^{t_2(x)} \frac{dt}{t^2 + \tau^2} = \frac{1}{\pi} \operatorname{arctg} \frac{t}{\tau} \Big|_{t_1(x)}^{t_2(x)} = \frac{1}{\pi} \left[\operatorname{arctg} \frac{x-a}{\tau} + \operatorname{arctg} \frac{b-x}{\tau} \right], \tag{8}$$

then $\lim_{\tau \rightarrow 0^+} \int_{t_1(x)}^{t_2(x)} K_\tau(t) dt = 1$ for all $x \in [a, b]$, therefore, the kernel normalization condition in representations (5)–(7) is satisfied, it remains to check conditions (4). Consider an integral of the form $\int_{t_1}^{-\delta} \cdot dt + \int_{-\delta}^{+\delta} \cdot dt + \int_{+\delta}^{t_2} \cdot dt = \int_{t_1}^{t_2} K_\tau(t) dt$. According to formula (8), we have

$$\begin{aligned} \int_{t_1(x)}^{-\delta} K_\tau(t) dt &= \frac{1}{\pi} \left[-\operatorname{arctg} \frac{\delta}{\tau} + \operatorname{arctg} \frac{b-x}{\tau} \right], \\ \int_{-\delta}^{\delta} K_\tau(t) dt &= \int_{|t| \leq \delta} K_\tau(t) dt = \frac{2}{\pi} \operatorname{arctg} \frac{\delta}{\tau}, \\ \int_{\delta}^{t_2(x)} K_\tau(t) dt &= \frac{1}{\pi} \left[\operatorname{arctg} \frac{x-a}{\tau} - \operatorname{arctg} \frac{\delta}{\tau} \right]. \end{aligned} \tag{9}$$

Since $\int_{|t| \geq \delta} K_\tau(t) dt = \int_{t_1}^{-\delta} \cdot dt + \int_{+\delta}^{t_2} \cdot dt$, then, taking into account (9), we obtain the equality

$$\int_{|t| \geq \delta} K_\tau(t) dt = \frac{1}{\pi} \left[\operatorname{arctg} \frac{x-a}{\tau} - 2\operatorname{arctg} \frac{\delta}{\tau} + \operatorname{arctg} \frac{b-x}{\tau} \right]. \tag{10}$$

For $\tau \rightarrow 0^+$, all the terms on the right side of (10) have a limit $\frac{\pi}{2}$, which proves the equality

$$\lim_{\tau \rightarrow 0^+} \int_{|t| \geq \delta} K_\tau(t) dt = 0, \tag{11}$$

for all $\delta > 0$ and $x \in (a, b)$. So the sequence $K_n(x - y) = \left\{ \frac{n}{d} \cdot \frac{1}{n^2(x-y)^2 + d^2} \right\}_{n=1}^\infty$ is a distribution of unity. Now it remains to show the validity of the limit equalities (2) and (3), which largely predetermine the possible applications of singular integrals of functions in applied analysis [1, 3].

First, we note that $f(x) \in C(\Omega)$ in (2), where $\Omega = [a, b]$, then there is a constant M_f such that for any $x \in \Omega$ $\max_x |f(x)| \leq M_f$ (uniform boundedness of continuous functions

defined on a bounded support). Such assumption considerably simplifies the proof of equality (2). Let us introduce the notation $\int_{t_1(x)}^{t_2(x)} K_\tau(t)dt = \xi_\tau(x)$. It is clear that, in accordance with (8), $\xi_\tau(x) \leq 1$ for $\forall(x, \tau)$. When $\tau \rightarrow 0^+$ deviation $((K_\tau f)(x) - f(x))$ at all points $x \in [a, b]$ will be close in magnitude to the expression $((K_\tau f)(x) - f(x))\xi_\tau(x)$, which is written as the following integral

$$I_\tau(x, f) = \int_{t_1(x)}^{t_2(x)} K_\tau(t)[f(x-t) - f(x)]dt. \tag{12}$$

Integral (12) can have the representation as $I_\tau(x, f) = \int_{|t| \geq \delta} \cdot dt + \int_{|t| \leq \delta} \cdot dt$. For the first integral, we use the estimate

$$\left| \int_{|t| \geq \delta} K_\tau(t)[f(x-t) - f(x)]dt \right| \leq \int_{|t| \geq \delta} K_\tau(t)[f(x-t) - f(x)]dt \leq 2M_f \int_{|t| \geq \delta} K_\tau(t)dt.$$

For the second integral, due to the smallness of $\delta > 0$, we can accept $|f(x-t) - f(x)| \leq \xi$, where $\xi > 0$ and possibly depends on δ . As a result we obtain $\left| \int_{|t| \leq \delta} K_\tau(t)[f(x-t) - f(x)]dt \right| \leq \xi \int_{|t| \leq \delta} K_\tau(t)dt$. Taking into account (11) $\lim_{\tau \rightarrow 0^+} |(K_\tau f)(x) - f(x)\xi_\tau(x)| \leq \lim_{\tau \rightarrow 0^+} \sup_{x \in \Omega} |(K_\tau f)(x) - f(x)\xi_\tau(x)| \leq \xi$, since $\lim_{\tau \rightarrow 0^+} \xi_\tau(x) = 1$, it implies $\lim_{\tau \rightarrow 0^+} \|(K_\tau f) - f\|_C = 0$ for any continuous function. This shows that the sequence $\{K_n(x)\}_{n=1}^\infty$, generating the sequence of integrals $\{K_n f\}_{n=1}^\infty$, can be presented as a representation of the unit. Using the formula for differentiating, we receive

$$(\psi(u))'_u = \left(\int_{a(u)}^{b(u)} f(x, u)dx \right)'_u = \int_{a(u)}^{b(u)} \frac{\partial f}{\partial u} f(x, u)dx + f[b(u), u]b'(u) - f[a(u), u]a'(u)$$

and, according to (7),

$$(D(K_\tau f))(x) = \int_{t_1(x)}^{t_2(x)} K_\tau(t) \frac{\partial f}{\partial x}(x-t)dt + K_\tau(t_2(x))f(a) - K_\tau(t_1(x))f(b). \tag{13}$$

Without loss of generality, we can assume that $f(a) = f(b) = 0$ in the last expression. If for the investigated continuous function $f(x)$, defined on the interval $[a, b]$, $f(a) \neq 0$ and $f(b) \neq 0$, then passing to the new function $\hat{f}(x)$, according to the rule

$$\hat{f}(x, f) = f(x) - \left[f(a) \frac{b-x}{b-a} + f(b) \frac{x-a}{b-a} \right],$$

we obtain required boundary values $\hat{f}(a, f) = 0$ and $\hat{f}(b, f) = 0$. It is clear that $\hat{f}(x, f)$ in some cases can be considered as a certain functional in the class $C(\Omega)$, which is emphasized above by the corresponding notation. In view of this remark, in accordance with (13), the following equality $(D(K_\tau \hat{f})) (x) = (K_\tau (D\hat{f})) (x)$ can be asserted for all $x \in (a, b)$, where $f(x)$ has a derivative $(Df)(x)$ at the point x . Thus two convergent processes take place, namely, $(K_\tau f)(x) \rightarrow f(x)$ for any $\tau \rightarrow 0^+$ continuous function and $(K_\tau (Df))(x) = (D(K_\tau f))(x) \rightarrow (Df)(x)$ under the condition that $f(x)$ is differentiable in x . For subsequent applications of these processes, it is important to emphasize that we are talking about uniform convergence, as follows from the above proofs. This feature of approximation sequences $\{(K_\tau f)(x)\}$ and $\{(D(K_\tau f))(x)\}$ plays a decisive role in solving problems in the theory of approximation of functions from approximate data [5]. From a practical point of view, we focus primarily on the processing of empirical data, which should include functions of limited change and absolutely continuous functions. First and second classes mentioned here include measurable functions, and therefore include a wider class of domain Ω ($f \in C_\Sigma(\Omega)$) integrable functions. If we proceed from the theory of Lebesgue integration, then in the simplest case we can talk about functions from the class $L_1(\Omega)$ and, in special cases, from the class $L_p(\Omega)$ ($p > 1$).

Since the items of the sequence $\{K_n(x)\}_{n=1}^\infty$ forming the approximate unit are positive (4), then $\|K_n\|_{L_1} \leq 1$. For $n \rightarrow 1$ $\|K_n\|_{L_1} \rightarrow 1$, however, $\lim_{n \rightarrow \infty} \|K_n\|_p = \infty$. The similar relation holds for all sequences $\{K_n(x)\}_{n=1}^\infty$, that form the distribution of unity. As for the sequence $\{(K_n f)(x)\}$, if $\{K_n(x)\}_{n=1}^\infty \in L_1$ and $f \in L_p$ ($1 \leq p < \infty$), then $\{(K_n f)(x)\} \in L_p$ also

$$\|(K_n f)\|_{L_p} \leq \|K_n\|_{L_1} \cdot \|f\|_{L_p}. \quad (14)$$

Applying the Lebesgue integration theory, for any integrable function $f(x) \in L_1(\Omega)$, the conditions are satisfied at almost all points $x \in \Omega = [a, b]$.

$$\int_0^h (f(x+t) - f(x))dt = o(h), \quad \int_0^h (f(x-t) - f(x))dt = o(h), \quad (15)$$

for $h \rightarrow 0^+$. Such points $x \in \Omega$ are usually called the Lebesgue points of the function $f(x)$, and their set is called the Lebesgue set. It is clear that if $f(x)$ is continuous on Ω , then all Lebesgue points coincide with the points of its continuity. Before turning to the solution of the problem posed above, we point out that it is convenient to replace the integrability criterion of $f(x)$ in the form (15) by the condition.

$$\int_0^h |f^*(y)|dy = o(h), \quad (16)$$

where

$$f^*(y) = \frac{1}{2}[f(x+y) + f(x-y) - 2f(x)]. \quad (17)$$

Let us introduce the integral $I^*(h) = \int_0^h |f^*(y)|dy$, considering variable h and setting $0 < h \leq \delta$, where δ is a fixed sufficiently small number. The circumstance, that $f^*(y)$ satisfies condition (16), means that for each δ one can choose a sufficiently small number ε , such that $I^*(h) \leq \varepsilon h$. The behavior of the integral $\int_0^\delta K_n(y)f^*(y, x)dy$ is important for the convergence of the sequence $\{(K_n f^*)(x)\} n \rightarrow \infty$, hence we estimate its value using inequality

$$\left| \int_0^\delta K_n(y)f^*(y, x)dy \right| \leq \int_0^\delta |f^*(x, y)|K_n(y)dy. \tag{18}$$

For the integral on the right in (18), we integrate by parts

$$\int_0^\delta |f^*(x, y)|K_n(y)dy = K_n(y)I^*(y)|_0^\delta - \int_0^\delta I^*(y)K'_n(y)dy. \tag{19}$$

According to (18) $I^*(\delta) \leq \varepsilon h$, $K_n(y)I^*(y)|_0^\delta = \varepsilon \delta K_n(\delta)$. Further, for $y \leq \delta$ and $I^*(\delta) \leq \varepsilon y$, the second integral in (19) can be estimated from the inequality $\int_0^\delta I^*(y)|K'_n(y)|dy \leq \varepsilon \int_0^\delta y|K'_n(y)|dy$, which gives us

$$\left| \int_0^\delta f^*(x, y)K_n(y)dy \right| \leq \varepsilon \left[\delta K_n(\delta) + \int_0^\delta y|K'_n(y)|dy \right]. \tag{20}$$

For small ε and δ , the right side in (20) is arbitrarily small for all n . Omitting the quite obvious proof of this assumption, let us turn to the example of the model kernel $K_\tau(t) = \frac{\tau}{\pi} \frac{1}{t^2 + \tau^2}$ in (20)

$$\begin{aligned} \left[\delta K_\tau(\delta) + \int_0^\delta y|K'_\tau(y)|dy \right] &= \frac{\tau}{\pi} \frac{1}{\delta^2 + \tau^2} \delta + \frac{\tau}{\pi} \int_0^\delta y \frac{2y}{[y^2 + \tau^2]^2} dy \\ &= \frac{\tau \delta}{\pi[\delta^2 + \tau^2]} + \frac{\tau}{\pi} \left[-\frac{\delta}{\delta^2 + \tau^2} + \frac{1}{\tau} \arctg \frac{\delta}{\tau} \right] = \frac{1}{\pi} \arctg \frac{\delta}{\tau}. \end{aligned}$$

Hence

$$\left| \int_0^\delta f^*(x, y)K_\tau(y)dy \right| \leq \frac{\varepsilon}{\pi} \arctg \frac{\delta}{\tau}. \tag{21}$$

Since for any fixed $\delta \lim_{\tau \rightarrow 0^+} \arctg \frac{\delta}{\tau} = \frac{\pi}{2}$, then the smallness of the right side in (21) for $\tau \rightarrow 0^+$ is obvious. It remains to consider the estimate of the integral

$\int_{|t|>\delta} K_\tau(y)f^*(x, y)dy$ for $\tau \rightarrow 0^+$. It is no longer possible to use a pointwise estimate of the type $|f(x + y) - f(x)| \leq M_f$, as was done in the case of a set of continuous functions. For functions integrable in the domain Ω , in order to characterize their properties, only integral (average) criteria can be used. Due to the positivity of the function $K_\tau(y)$, it is convenient to use the mean value theorem to estimate this integral

$\left| \int_{|t|>\delta} K_\tau(y)f^*(x, y)dy \right| = K_\tau(\xi) \int_{|t|>\delta} |f^*(x, y)|dy$. For functions from the class $L_1(\Omega)$, this integral exists (it is bounded to a certain number $M(f)$). As for the multiplier $K_\tau(\xi)$, where $a \leq \xi \leq b$, then $\lim_{\substack{\tau \rightarrow 0^+ \\ (\xi \neq 0)}} K_\tau(\xi) = 0$. For a function $K_\tau(t) = \frac{\tau}{\pi(t^2 + \tau^2)}$, the fulfillment

of this condition for $t = \xi \neq 0$ is quite obvious. Thus, it is proved that the limit element of the sequence $\{(K_n f^*)(x)\}_{n=1}^\infty$ exists for any integrable function $f(x)$ defined on a finite support Ω and is equal to zero.

3 Results

Note that the logical constructions performed above can also be carried out for the integral $\int K_\tau(t)[f(x - t) - f(x)]dt$, as was done earlier for continuous functions. The choice of a function $f^*(x, y)$ with subsequent evaluation of the integral $\int K_\tau(y)f^*(x, y)dy$ is determined by the fact that it is desirable for applications to immediately limit the set of summable functions. The function $f^*(x, y)$ in analysis plays the role of a characteristic of the analytic properties of the original $f(x)$ on Ω . In particular, the quantity $\omega(h, x) = f(x + h) + f(x - h) - 2f(x)$ is called the modulus of continuity $f(x)$ at the point x . For integrable functions, we the integral modulus of continuity $\|\omega(h)\|_{L_1} = \|f(x + h) + f(x - h) - 2f(x)\|_{L_1}$. If this value does not exceed the values $M_1(f)$, then some parametrized subset is selected from the set $C_\Sigma(\Omega)$, determined by the pair $(h, M_1(f))$. This approach is used in the presentation of the theory of the function $f(x)$ from the class $C_\Sigma(\Omega)$ of non-differentiable in the general sense, when a direct consideration of the limit relation (3) is not possible. The corresponding practical interpretations of (3) require the introduction of the concepts of “generalized differentiation”, which is beyond the scope of this work.

It is known that singular integrals [1, 6] of functions find application in the constructive theory of functions [4], applied analysis [3], and computational mathematics [7]. The application of singular integrals is especially important in those applied problems of the constructive theory of functions, where the functions are represented approximately, for example, by a vector of values $f = \{f(x_0), f(x_1), \dots, f(x_m)\}$ associated with some discrete sequence of points on the segment $[a, b]$, say $\{x_k\}_{k=0}^m$. In this situation, the integral representation considered in the work

$$(K_n f)(x) = \int_a^b K_n(x, x')f(x')dx' = f_n(x) \rightarrow f(x) \quad (n \rightarrow \infty) \tag{22}$$

will pass into the mapping $f_{n,m}(x) = \sum_{k=0}^m f_k \bar{K}_{m,k}^{(n)}(x) \rightarrow f(x)$ ($n \rightarrow \infty$, $m \rightarrow \infty$), where $\{\bar{K}_{m,k}^{(n)}(x)\}$ is a certain system of functions generated by the kernel $K_n(x, x')$ on a discrete set $\{x_k\}_{k=0}^m$, and determined by one or another method of constructing quadratures. It is assumed that the items of the functional sequence $\{\bar{K}_{m,k}^{(n)}(x)\}_{k=0}^m$ for each pair (n, x) are continuous functions on the segment $[a, b]$, and then there is a sequence $\{f_{n,m}(x)\}$ of generalized polynomials of order not higher than m associated with the sequence $\{\bar{K}_{m,k}^{(n)}(x)\}_{k=0}^m$. In this regard, we should introduce a notation $\{P_{m,k}(x, f, \tau)\}_{k=1}^{\infty}$ and consider the problem of the theory of the best approximation of a function $f \in C(\Omega)$ by the indicated polynomials. It is clear that for continuous functions one can expect uniform convergence of approximation polynomials $P_{m,k}(x, f, \tau)$ [9] to $f(x)$ as $m \rightarrow \infty$ and $\tau \rightarrow 0^+$. In the case of functions, summable on the set Ω , we can talk about convergence almost everywhere. These assertions are quite clear from the provisions that were stated in the proof of the above limit relations for singular integrals of functions from the classes $C(\Omega)$ and $C_{\Sigma}(\Omega)$.

4 Discussion

If we assume, that the function $f(x)$ on can be represented by its σ -approximation $f_{\sigma}(x)$ such that $\|f(x) - f_{\sigma}(x)\| \leq \sigma \|f(x)\|$, then the problem of restoring the approximate unit (22) is formed as an optimization problem of the form.

$$\inf_{\tau \in (0,1)} \inf_{f \in C(\Omega)} \|(K_{\tau}f) - f_{\sigma}\|_{L_2} = \|K_{\tau} * f^* - f_{\sigma}\|_{L_2} \leq \sigma \cdot \|f_{\sigma}\|. \quad (23)$$

The solution (recovery result) is a function $\tilde{f}_{\tau^*(\sigma)}(x) = (K_{\tau^*} * f^*)(x)$, belonging to the set of continuous functions $C(\Omega)$. If the function $f(x)$ goes beyond the scope of this functional class, then the indicated solution should be understood as its certain regular (continuous) component $f_H(x)$ [8]. Despite the fact that, by definition, σ -approximation is just a measurable function, the stated approach is consistent due to Luzin's theorem: if $f(x)$ is finite and measurable on Ω ($|\Omega| < \infty$), then for any ε there exists a measurable set $A \subset \Omega$ such that $|\Omega - A| < \varepsilon$ and $f(x)$ is continuous on A . This means that every measurable function has a continuous component almost everywhere on Ω . Outside the set A , there are those points $x \in \Omega$ at which $f(x)$ loses continuity. For the set of these functions, Lebesgue's theorem holds: any function of bounded variation on $[a, b]$ has at most a countable set of discontinuity points. This class of functions is closely related to those problems of the theory of approximation of functions, which can be represented by vectors of their values $\vec{f} = \{f(x_k)\}_{k=0}^m$. In this case, optimization problem (23) can be represented in matrix form, namely

$$\min_{\tau \in (0,1)} \min_{\vec{f} \in R_m} \left\| \hat{G}_{\tau} \vec{f} - \vec{f}_{\sigma} \right\|_{l_2(R_n)} \leq \sigma \cdot \left\| \vec{f}_{\sigma} \right\|_{l_2(R_n)}, \quad (24)$$

where the matrix \hat{G}_{τ} transforms a vector \vec{f} of dimension m into a vector \vec{f}_{τ} of dimension n , in accordance with the dimension of the original vector \vec{f}_{σ} (empirical vector, if we

are talking about processing empirical data) [9, 10]. Let's make a few remarks about the approaches for constructing the matrix \hat{G}_τ . It is clear, that in the general case we are talking about the algebraization of the integral equation $K_\tau f = f_\sigma$, corresponding to expression (22), and its subsequent solution by generalized inversion methods. Recall that the construction of a summation analogue for an integral operator is based on the choice of a certain covering $X = \bigcup_{l=1}^m X_l$, where $X_l = [x_{l-1}, x_l]$, $l = 1, \dots, m$, $x_0 = a$, $x_m = b$. Next, we need to make an assumption about the nature of the analytical behavior of the desired function within the intervals X_l . One of the simple options for calculating integrals, widely used in practice, are the assumptions about the linear nature of the behavior $f(x)$ within $[x_{l-1}, x_l]$. In this case, integrals of type (22) over the domain X_l take the form

$$I_l(x) = \int_{x_{l-1}}^{x_l} K_\tau(x, x') \left(f_{l-1} + \frac{f_l - f_{l-1}}{x_l - x_{l-1}} (x' - x_{l-1}) \right) dx'. \quad (25)$$

Considering the position of the point x on $[a, b]$, we will assume that $K_\tau(x, x') = K_\tau(x - x') = K_\tau(h)$ if $|x - x'| \leq h = \frac{b-a}{m}$. Omitting the subsequent construction of the quadrature formula, we restrict ourselves to the final calculation of integrals of the type (25).

$$\begin{aligned} I_l(x) &= f_{l-1} N_{\tau,l}(x) + f_l M_{\tau,l}(x), \\ N_{\tau,l}(x) &= x_l A_{\tau,l}(x) + B_{\tau,l}(x), \quad M_{\tau,l}(x) = B_{\tau,l}(x) + x_{l-1} A_{\tau,l}(x), \\ A_{\tau,l}(x) &= \frac{1}{x_l - x_{l-1}} \int_{x_{l-1}}^{x_l} K_\tau(x, x') dx', \quad B_{\tau,l}(x) = \frac{1}{x_l - x_{l-1}} \int_{x_{l-1}}^{x_l} K_\tau(x, x') x' dx'. \end{aligned}$$

Finally

$$(K_\tau f)(x) \sim \sum_{k=0}^m f_k G_{m,k}(x),$$

$$G_{m,k}(x) = M_{\tau,k}(x) + N_{\tau,k+1}(x), \quad k = 0, 1, \dots, m, \quad M_{\tau,k=0}(x) \equiv 0, \quad N_{\tau,m+1}(x) \equiv 0.$$

Thus, if a vector \vec{f}_σ , whose components are associated with the points x_i of the segment $[a, b]$, is given then the functions $G_{m,k}(x)$ for $x = x_i$ ($i = 1, \dots, n$) form a matrix of dimension $m \times n$ of the optimization problem (24) with elements $G_{m,k_i}(x) = G_{m,k}(x_i)$.

5 Conclusion

The main goal of this article, providing the restoring of the real-valued functions $f = \{f(x_0), f(x_1), \dots, f(x_m)\}$, associated with discrete sequence of points $\{x_k\}_{k=0}^m$ on $[a, b]$, using generalized kernel $(K_n f)(x)$, is achieved. Theoretical calculations imply the restoring of the unit approximation, which is able to increase the accuracy of approximations. The above algorithm for the problem of recovering an approximate unit is formed as an optimization problem and requires further implementation in a computational experiment.

Acknowledgments. The authors would like to thank the North-Caucasus Federal University for supporting in the contest of projects competition of scientific groups and individual scientists of the North-Caucasus Federal University. The work is supported by North-Caucasus Federal Center for Mathematical Research under agreement №075-02-2021-1749 with the Ministry of Science and Higher Education of the Russian Federation.

References

1. Mokrousova, T.A., Andreeva, D.V., Goltyaev, Yu.A.: Direct work and conversion of generalized functions. *Colloquium – J.* **2**, 33–36 (2018)
2. Aïmar, H., Gómez, I.: Boundedness and concentration of random singular integrals defined by wavelet summability kernels. *J. Math. Anal. Appl.* **514**, 126315 (2022)
3. Lanczos, C.: *Applied Analysis*. Prentice Hall, Upper Saddle River (1961)
4. Natanson, I.P.: *Constructive Function Theory*. Frederick Ungar Publishing, New York (1965)
5. Chen, X., Tan, J., Liu, Z., Xie, J.: Approximation of functions by a new family of generalized Bernstein operators. *J. Math. Anal. Appl.* **450**, 244–261 (2017)
6. Lerner, A.K., Ombrosi, S., Rivera-Ríos, I.P.: Commutators of singular integrals revisited. *Bull. Lond. Math. Soc.* **51**, 107–119 (2018)
7. Lebedev, V.I.: *Functional analysis and computational mathematics*. FIZMATLIT (2000)
8. Shoukalla, E.S., Markos, M.A.: The economized monic Chebyshev polynomials for solving weakly singular Fredholm integral equations of the first kind. *Asian-Eur. J. Math.* **13**, 2050030 (2020)
9. Matysik, O.V.: Implicit method for solving a self-adjoint ill-posed problem with approximately operator and a posteriori choice of the regularization parameter. *Vesnik Hrodzenskaha Dziarzhaunaha Universiteta Imia Ianki Kupaly* **3**, 75–82 (2015)
10. Naats, V.I., Yartseva, E.P., Andrukhiv, L.V.: Computational model for a differential equation with approximate initial data based on the Volterra integral equation of the second kind. *News Kabardino-Balkarian Sci. Center RAS* **4**, 5–16 (2021)



RNS Reverse Conversion Algorithm and Parity Detection for the Wide Arbitrary Moduli Set

Vitaly Slobodskoy^(✉), Elizaveta Martirosyan, Valeria Ryabchikova,
and Roman Kurmaev

Huawei, Nizhny Novgorod Research Center, Computing, Nizhny Novgorod, Russia
{vitaly.slobodskoy, ryabchikova.valeria.georgievna,
kurmaev.roman1}@huawei.com,
martirosyan.elizaveta@huawei-partners.com

Abstract. Due to absence of long carry propagation logic in its modular adders and multipliers Residue Number System (RNS) provides advantages for performing arithmetic computations in the hardware accelerators targeting multiplications and additions only. However, a conversion from RNS to Binary Number System (BNS) is a well-known complex operation requiring larger hardware area and consuming more power than modular multipliers which significantly limits applicability of RNS. Many non-modular operation (e.g. comparison, division) implementations are also based on reverse conversion. In this article we demonstrate a variation of Chinese Remainder Theorem (CRT) based algorithm for RNS to BNS reverse conversion for an arbitrary moduli set. We propose an approximate method to overcome the heaviest $\text{mod } M$ operation via the rank of number calculation and demonstrate its correctness on the covered number range. We show that algorithm parameters selection has no restrictions on moduli set and needed parameters can always be found assuming that covered maximum value is less than $M - 1$. Hardware implementation of the new RNS reverse conversion algorithm based on CRT is more than 30% faster and consumes up to 80% less power than implementations of CRT based reverse conversion algorithms using other known ways to compute the final reduction operation.

Keywords: Residue Number System (RNS) · Binary Number System (BNS) · Chinese Remainder Theorem (CRT) · reverse conversion · Mixed Radix Conversion (MRC) · RNS rank of number

1 Introduction

Residue Number System (RNS) is widely exposed in the hardware related to digital signal processing (FIR, DFT, FFT, etc.) [1], communication systems [2], cryptography [3, 4] and recently in neural networks acceleration [5–8] in the areas where arithmetic computations contain multiplications and additions only. In RNS addition and multiplication operations can be decomposed into smaller

carry-free modulo operations for parallel processing. While RNS is very hardware and power efficient for such computations, it has limited applicability in the cases when additional (non-modular) operations are involved (e.g. comparison, division). Many implementations of such operations rely on reverse conversion (RNS to Binary Number System (BNS)).

There are two main directions in RNS reverse conversion - methods based on Chinese Remainder Theorem (CRT) and on Mixed-Radix Conversion (MRC) [9]. Reverse conversion based on MRC can be implemented via modulo multipliers enabling ability to re-use RNS arithmetic logic unit hardware using multi-cycle approach. However, this method is iterative and needs $O(n^2)$ mathematical operations (n - number of moduli). As result, these methods demonstrate quite poor performance, but require relatively small hardware area. In case $2^n \pm 1$ moduli MRC based reverse conversion can be implemented using just bitwise rotation [10–12].

The CRT based algorithms can also be efficiently used in case of the three and four moduli sets of special forms [13–15]. For arbitrary moduli set, CRT-based reverse conversion needs complex final modulo reduction and large multipliers. The final reduction can be simplified by approximate *rank of number* computation [16] or using redundant modulo [17]. However, the proposed approximation in [16] adds requirements on moduli set and can not be applied for a wide range of moduli sets where product of moduli is not significantly higher than the biggest covered binary value. Attempts to significantly increase product of moduli or using redundant modulo inevitably increases overall number of bits in moduli causing loss of hardware area and power consumption efficiency for the main arithmetic operations. For example, in case of integer matrix multiplication the number of RNS multiplications is order of magnitude higher than the number of RNS reverse conversions needed to perform the whole matrix multiplication of integer values originally represented in BNS.

In this paper we focus on hardware and power consumption efficiency of RNS reverse conversion unit hardware implementation for the arbitrary moduli set with a big number of moduli (more than 10 moduli). We propose a modification of approximation described in [16] which doesn't have limitations on moduli set, however, depending on moduli set its hardware and power efficiency varies.

2 Residue Number System

The Residue Number System is a non-positional numeral system where numbers are represented as remainders of division by the special set of n coprime numbers m_1, m_2, \dots, m_n called *moduli set*. The product of all moduli $M = \prod_{i=1}^n m_i$ is a *dynamic range*. Any integer number $X \in [0, M - 1]$ can be uniquely represented as $X = (x_1, x_2, \dots, x_n)$, where $x_i = X \bmod m_i = |X|_{m_i}$, $i \in [1, n]$ are *residues* of the number X modulo m_i .

The binary number X corresponding to the given residuals (x_1, x_2, \dots, x_n) in RNS can be obtained using Chinese Remainder Theorem (CRT) [9] as:

$$|X|_M = \left| \sum_{i=1}^n |x_i|_{m_i} |M_i|_{m_i}^{-1} M_i \right|_M, \quad (1)$$

where $x_i = |X|_{m_i}$, $M_i = \frac{M}{m_i}$, $|M_i|_{m_i}^{-1}$, is a multiplicative inverse of M_i by modulo m_i , $i \in [1, n]$.

The main disadvantage of CRT based methods is the final modulo operation by big number M significantly increasing hardware area and lowering performance of the reverse conversion unit. In order to overcome this issue, CRT can be represented the following way:

$$X = \sum_{i=1}^n |x_i|_{m_i} |M_i|_{m_i}^{-1} M_i - rM, \quad (2)$$

where r is a non-negative integer called *rank* of number X , $0 \leq r < n$. Ability to compute r efficiently determines the efficiency of methods like RNS reverse conversion, base extension or parity detection.

2.1 Rank Calculation Using Redundant Modulo

A method to calculate rank of number using redundant modulo was suggested in [17]. Let x_{n+1} is a residue of X by redundant modulo $m_{n+1} \geq n$:

$$|X|_{m_{n+1}} = x_{n+1} = \left| \left| \sum_{i=1}^n |x_i|_{m_i} |M_i|_{m_i}^{-1} M_i \right|_{m_{n+1}} - |rM|_{m_{n+1}} \right|_{m_{n+1}} \quad (3)$$

$$|rM|_{m_{n+1}} = \left| \left| \sum_{i=1}^n |x_i|_{m_i} |M_i|_{m_i}^{-1} M_i \right|_{m_{n+1}} - x_{n+1} \right|_{m_{n+1}} \quad (4)$$

Multiplying (4) by $|M^{-1}|_{m_{n+1}}$ - the inverse of M modulo m_{n+1} :

$$|r|_{m_{n+1}} = \left| |M^{-1}|_{m_{n+1}} \left(\left| \sum_{i=1}^n |x_i|_{m_i} |M_i|_{m_i}^{-1} M_i \right|_{m_{n+1}} - x_{n+1} \right) \right|_{m_{n+1}} \quad (5)$$

Since $0 \leq r < n$ and $m_{n+1} \geq n$:

$$r = |r|_{m_{n+1}} = \left| |M^{-1}|_{m_{n+1}} \left(\left| \sum_{i=1}^n |x_i|_{m_i} |M_i|_{m_i}^{-1} M_i \right|_{m_{n+1}} - x_{n+1} \right) \right|_{m_{n+1}} \quad (6)$$

Obviously, this method has quite serious drawbacks: firstly, it needs to maintain residue by the redundant modulo extending the cost of all the arithmetic operations in RNS, secondly, it requires a large number of multipliers and adders by redundant modulo, namely $n + 1$ multipliers and n adders by modulo.

2.2 Approximate Rank Calculation

Let $\xi_i = |x_i M_i^{-1}|_{m_i}$, $i \in [1, n]$, then:

$$X = \sum_{i=1}^n \xi_i M_i - rM \quad (7)$$

Dividing both sides by M :

$$\sum_{i=1}^n \frac{\xi_i}{m_i} = r + \frac{X}{M} \quad (8)$$

$$r = \left\lfloor \sum_{i=1}^n \frac{\xi_i}{m_i} \right\rfloor \quad (9)$$

In [16] (where rank is called *reduction factor*) the following approximations are introduced:

1. m_i is replaced by 2^w , where $2^{w-1} < m_i \leq 2^w$, $i \in [1, n]$. This implies that $w = \lceil \log_2 m_i \rceil$, $i \in [1, n]$, such moduli sets are called *balanced* - all moduli have the same number of bit. In [18] there is even stronger requirement for the moduli: $m_i = 2^w - \mu_i$, where μ_i is an integer in the range $0 \leq \mu_i < 2^{\lfloor w/2 \rfloor}$, $i \in [1, n]$.
2. ξ_i is replaced by $\text{trunc}(\xi_i) = \xi_i \& (\overbrace{1\dots 1}^q \overbrace{0\dots 0}^{w-q})_{(2)}$ where $\&$ means a bitwise AND operation.

Then approximation \hat{r} of r is defined as:

$$\hat{r} = \left\lfloor \sum_{i=1}^n \frac{\text{trunc}(\xi_i)}{2^w} + \alpha \right\rfloor \quad (10)$$

where $0 < \alpha \leq 1$ is an adjusting parameter designed to compensate for the approximation error. To evaluate the effect of approximation ϵ and δ are defined as:

$$\epsilon_i = \frac{2^w - m_i}{2^w}, \quad \delta_i = \frac{\xi_i - \text{trunc}(\xi_i)}{m_i}, \quad i \in [1, n] \quad (11)$$

$$\epsilon = \max_{i \in [1, n]} (\epsilon_i), \quad \delta = \max_{i \in [1, n]} (\delta_i) \quad (12)$$

The paper proofs theorem claiming that when $0 \leq n(\epsilon + \delta) \leq \alpha < 1$, $\hat{k} = k$ for all $0 \leq X < (1 - \alpha)M$.

In order to represent a n -bit binary word in RNS, it is necessary to choose a moduli-set which leads to $M > 2^n$. Using rank calculation method it is possible to choose the basis and parameters in such way that $(1 - \alpha)M > 2^n$. This method will only produce the wrong result outside the $[0; 2^n - 1]$ range of the RNS dynamic range.

However, this method can not be applied for any modulo set. Consider the need to cover $[0; 2^{106} - 1]$ binary range with 12×9 -bit moduli. From (11–12) we can derive:

$$\epsilon = \frac{2^w - \min_{i \in [1, n]}(m_i)}{2^w} \quad (13)$$

so ϵ depends on the minimal modulo in the balanced moduli set. Moduli set with the largest minimal modulo and the largest product of moduli satisfying conditions above is (512, 511, 509, 505, 503, 501, 499, 493, 491, 487, 481, 479), so minimal modulo is 479. In order to cover $[0; 2^{106} - 1]$ binary range without approximation error: $0 \leq X < 2^{106} = (1 - \alpha)M$, $\alpha = 1 - \frac{2^{106}}{M} \approx 0.65$. However $n\epsilon = 12 \frac{512-479}{512} \approx 0.77 \leq n(\epsilon + \delta) \leq \alpha$. As result, there is no way to apply this method under the requirements to the moduli set above.

3 Proposed Method

Let introduce integer values d_i , $i \in [1, n]$ and s so that:

$$\frac{1}{m_i} \geq \frac{d_i}{2^s} \quad (14)$$

Then

$$\frac{\xi_i}{m_i} \geq \frac{\xi_i d_i}{2^s} \quad (15)$$

$$\epsilon_i = \frac{\xi_i}{m_i} - \frac{\xi_i d_i}{2^s}, \epsilon = \max_{i \in [1, n]} \epsilon_i \quad (16)$$

$$\sum_{i=1}^n \frac{\xi_i d_i}{2^s} > \sum_{i=1}^n \frac{\xi_i}{m_i} - n\epsilon \quad (17)$$

$$\sum_{i=1}^n \frac{\xi_i}{m_i} - n\epsilon < \sum_{i=1}^n \frac{\xi_i d_i}{2^s} \leq \sum_{i=1}^n \frac{\xi_i}{m_i} \quad (18)$$

Substituting (8), multiplying each sides to 2^s and adding α as in [16]:

$$2^s \left(r + \frac{X}{M} \right) - 2^s n\epsilon + \alpha < \sum_{i=1}^n \xi_i d_i + \alpha \leq 2^s \left(r + \frac{X}{M} \right) + \alpha \quad (19)$$

If the following conditions are true:

$$0 \leq 2^s n\epsilon \leq \alpha < 2^s \quad (20)$$

$$0 \leq 2^s X < (2^s - \alpha)M \quad (21)$$

then

$$r < \frac{\sum_{i=1}^n \xi_i d_i + \alpha}{2^s} < r + 1 \quad (22)$$

so

$$\hat{r} = \left\lfloor \frac{\sum_{i=1}^n \xi_i d_i + \alpha}{2^s} \right\rfloor = r \quad (23)$$

3.1 Selection of Parameters s , d_i and α

It is worth to note that all the parameters s , d_i and α can be of integer type, the proposed method does not involve any complex arithmetic - n additional multipliers are needed, while final division by 2^s can be replaced by lightweight right shift operation allowing various optimizations on the hardware level.

According to (21):

$$0 \leq \alpha < \frac{2^s(M - X)}{M} \quad (24)$$

Consider the need to cover binary range $[0; 2^p - 1]$, $0 < p < \log_2 M$, then:

$$0 \leq \alpha < 2^s - \frac{2^{s+p}}{M} \quad (25)$$

Continuing (16):

$$\begin{aligned} \epsilon &= \max_{i \in [1, n]} \left(\frac{\xi_i}{m_i} - \frac{\xi_i d_i}{2^s} \right) = \frac{\xi_i}{m_i} \max_{i \in [1, n]} \left(1 - \frac{d_i m_i}{2^s} \right) < \\ &< \max_{i \in [1, n]} \left(\frac{2^s - d_i m_i}{2^s} \right) = 1 - \frac{\min_{i \in [1, n]}(d_i m_i)}{2^s} \end{aligned} \quad (26)$$

Consider the following condition is true:

$$1 - \frac{\min_{i \in [1, n]}(m_i d_i)}{2^s} \leq \frac{\alpha}{2^{sn}} \quad (27)$$

then (20) is true as well. Combining with (25):

$$n(2^s - \min_{i \in [1, n]}(m_i d_i)) < 2^s - \frac{2^{s+p}}{M} \quad (28)$$

Parameter s defines overall complexity of the proposed method, it defines the magnitude of d_i and α , smaller s means smaller multipliers and adders. So, we need to find the lowest possible integer s satisfying condition (28) where:

$$d_i = \left\lceil \frac{2^s}{m_i} \right\rceil \quad (29)$$

Since $\lim_{s \rightarrow \infty} m_i \left\lceil \frac{2^s}{m_i} \right\rceil = 2^s$, $\forall i \in [1, n]$, so $\lim_{s \rightarrow \infty} (n(2^s - \min_{i \in [1, n]}(m_i d_i))) = 0$ and $\lim_{s \rightarrow \infty} (2^s - \frac{2^{s+p}}{M}) = \infty$, there is always exists s that satisfies (28), so the proposed method doesn't have any limitations on the moduli set.

3.2 Base Extension and Parity Detection

Consider the need to find residue of X given in RNS by new modulo q (RNS base extension):

$$|X|_q = \left| \sum_{i=1}^n \xi_i M_i - rM \right|_q = \left| \sum_{i=1}^n \xi_i M_i \right|_q - |rM|_q \quad (30)$$

Table 1. CRT Reverse Design Area, Latency, Power

Moduli	Algorithm	Area(μm^2)	Latency(ns)	Diff	Power(W)	Diff
1	Proposed, $s = 14$	700698	24.74		4.17	
	Traditional mod	698115	35.32	30.0%	18.5	77.5%
	Comps and subs	683604	36.35	29.2%	12.5	66.7%
	Barrett Reduction	677621	35.52	30.3%	13.8	69.8%
2	Proposed, $s = 15$	779818	24.16		3.34	
	Traditional mod	774004	35.86	32.6%	6.72	50.3%
	Comps and subs	769767	35.88	32.7%	9.75	65.7%
	Barrett Reduction	779805	34.38	29.7%	17.6	81.0%
3	Proposed, $s = 13$	1073831	27.13		13.0	
	Traditional mod	1095977	42.81	36.6%	15.4	15.6%
	Comps and subs	1068485	39.13	30.7%	18.7	30.5%
	Barrett Reduction	1062649	33.85	19.9%	13.9	6.5%

Knowing the rank of number r using (23), it is only needed to perform all the parties of (30) using modulo q operations including final modulo q accumulation.

Quite often when $|M|_2 = 1$ (all the moduli are odd), it is needed to determine the parity of X , perform base extension for modulo 2. In this case (30) can be simplified to:

$$|X|_2 = \left| \sum_{i=1}^n |\xi_i|_2 - |r|_2 \right|_2 \quad (31)$$

Note: (30) and (31) are true using (23) only for those X when (21) is true.

4 Results

The proposed algorithm of RNS reverse conversion has been compared with CRT-based algorithms using different ways of $\text{mod } M$ operation (the rest of operations except the final reduction is the same):

1. Proposed method
2. Using traditional mod operation based on division
3. Via $\lceil \log_2 n \rceil$ comparisons and subtractions
4. Barrett Reduction based [19]

Moduli sets without ability to apply approximate algorithm proposed in [16] have been used in comparison:

1. (256, 255, 253, 251, 247, 241, 239, 233, 223, 217, 191, 127, 107, 59) covering $[0; 2^{106} - 1]$ binary range, $s = 14$
2. (255, 253, 251, 247, 241, 239, 233, 229, 227, 199, 128, 127, 103, 97) covering $[0; 2^{106} - 1]$ binary range, $s = 15$
3. (1024, 511, 257, 255, 253, 251, 247, 241, 239, 233, 229, 227, 223, 211, 199, 191) covering $[0; 2^{128} - 1]$ binary range, $s = 13$

Hardware implementation was done in Verilog using OpenLane design flow with SkyWater SKY130 PDK. Table 1 shows hardware area, latency (as result of timing analysis) and power consumption of 4 methods applied for 3 moduli sets. Difference columns show an improvement (in %) of the proposed method vs the one in the row ($\frac{row-proposed}{row}$). Proposed modification of the CRT Reverse conversion algorithm is generally more than 30% faster than similar algorithms with traditional implementations of the final $mod M$ reduction and consumes up to 81% less power, however, occupies slightly larger hardware area.

Power consumption savings are primarily because rank calculation based method is much faster. Rank value can be computed in parallel with $\sum_{i=1}^n \xi_i M_i$ after ξ_i values have been computed. The final reduction requires n -multiplexor choosing the right constant based on rank value and single subtraction operation only.

5 Conclusion

We proposed a modification of approximation scheme suggested in [16] and proved its validity when parameters have been properly chosen. Moreover, we demonstrated that our modification has no restrictions on the moduli set and needed parameters can always be found assuming that maximum covered value is less than $M - 1$. However, the hardware efficiency depends on the parameters and large s value might cause big multipliers for the rank calculation and neglect all the advantages.

Comparing hardware implementations proposed algorithm of RNS reverse conversion based on CRT is more than 30% faster than similar CRT-based algorithms using other known ways to compute the final $mod M$ operation.

References

1. Mohan, P.: Residue number systems: theory and applications. Birkhäuser Basel (2016)
2. Omondi, A., Premkumar, B.: Residue number systems: theory and implementation. Imperial College Press (2007)
3. Sousa, L., Antao, S., Martins, P.: Combining residue arithmetic to design efficient cryptographic circuits and systems. *IEEE Circ. Syst. Mag.* **16**(4), 6–32 (2016)
4. Bajard, J., Eynard, J., Merkiche, N.: Montgomery reduction within the context of residue number system arithmetic. *J. Cryptogr. Eng.* **8**(3), 189–200 (2018)
5. Salamat, S., Shubhi, S., Khaleghi, B., Rosing, T.: Residue-net: multiplication-free neural network by in-situ no-loss migration to residue number systems. In: 2021 26th Asia And South Pacific Design Automation Conference (ASP-DAC), pp. 222–228 (2021)
6. Lin, L., Schroff, J., Lin, T., Huang, T.: Residue number system design automation for neural network acceleration. In: 2020 IEEE International Conference On Consumer Electronics - Taiwan (ICCE-Taiwan), pp. 1–2 (2020)
7. Chervyakov, N., et al.: Area-efficient FPGA implementation of minimalistic convolutional neural network using residue number system. In: 2018 23rd Conference Of Open Innovations Association (FRUCT), pp. 112–118 (2018)

8. Samimi, N., Kamal, M., Afzali-Kusha, A., Pedram, M.: Res-DNN: a residue number system-based DNN accelerator unit. *IEEE Trans. Circ. Syst. I: Regular Papers* **67**(2), 658–671 (2020)
9. Szabó, N., Tanaka, R.: *Residue Arithmetic and Its Applications to Computer Technology*. McGraw-Hill, New-York (1967)
10. Wesolowski, M., Patronik, P., Berezowski, K., Biernat, J.: Design of a novel flexible 4-moduli RNS and reverse converter, pp. 1–6 (2012)
11. Sousa, L., Antao, S., Chaves, R.: On the design of RNS reverse converters for the four-moduli set $\{2^n + 1, 2^n - 1, 2^n, 2^{n+1} + 1\}$. In: *IEEE Transactions On Very Large Scale Integration (VLSI) Systems*, vol. 21, no. 10, pp. 1945–1949 (2013)
12. Sousa, L., Antao, S.: MRC-based RNS reverse converters for the four-moduli sets $\{2^n + 1, 2^n - 1, 2^n, 2^{2n+1} - 1\}$ and $\{2^n + 1, 2^n - 1, 2^{2n}, 2^{2n+1} - 1\}$. *IEEE Trans. Circ. Syst. II: Express Briefs* **59**(4), 244–248 (2012)
13. Hiasat, A., Sweidan, A.: Residue number system to binary converter for the moduli set $(2^{n-1}, 2^n - 1, 2^n + 1)$. *J. Syst. Architect.* **49**, 53–58 (2003)
14. Ahmadifar, H., Jaberipur, G.: A new residue number system with 5-Moduli set: $\{2^{2^q}, 2^q \pm 3, 2^q \pm 1\}$. *Comput. J.* **58**(7), 1548–1565 (2015)
15. Gbolagade, K., Chaves, R., Sousa, L., Cotofana, S.: An improved RNS reverse converter for the $\{2^{2n+1} - 1, 2^n, 2^n - 1\}$ moduli set. In: *Proceedings Of 2010 IEEE International Symposium On Circuits And Systems*, pp. 2103–2106 (2010)
16. Kawamura, S., Koike, M., Sano, F., Shimbo, A.: Cox-rower architecture for fast parallel montgomery multiplication. In: Preneel, B. (ed.) *EUROCRYPT 2000*. LNCS, vol. 1807, pp. 523–538. Springer, Heidelberg (2000). https://doi.org/10.1007/3-540-45539-6_37
17. Shenoy, A., Kumaresan, R.: Fast base extension using a redundant modulus in RNS. *IEEE Trans. Comput.* **38**(2), 292–297 (1989)
18. Kawamura, S., et al.: Efficient algorithms for sign detection in RNS using approximate reciprocals. In: *IEICE Transactions On Fundamentals Of Electronics, Communications And Computer Sciences*, vol. E104.A, pp. 121–134 (2021)
19. Barrett, P.: Implementing the Rivest Shamir and Adleman public key encryption algorithm on a standard digital signal processor. In: *Proceedings on Advances in Cryptology-CRYPTO 1986*, pp. 311–323 (1987)



Anomalous Solute Transport in an Inhomogeneous Porous Medium Taking into Account Mass Transfer

T. O. Dzhiyanov^(✉), Sh Mamatov, and M. S. Zokirov

Samarkand State University Republic of Uzbekistan, St. University, 15, Samarkand 703100,
Uzbekistan
t.dzhiyanov@mail.ru

Abstract. The paper considers non-isothermal anomalous solute transport in an inhomogeneous porous medium consisting of two zones with different filtration-capacitive properties and characteristics of the solute transport. In contrast to the approaches that are used in the theory of interpenetrating continua for the problems of filtration, solute transport, one medium is considered here, the solute transport to another medium in the corresponding equations is taken into account through the source (sink) term in the form of a fractional time derivative of the concentration of the solute. In the study, we investigate replaced bicontinuum medium by multicontinuum medium.

Keywords: anomalous solute transport · fractional derivatives · solute concentration · porous medium · numerical solution

1 Introduction

Underground reservoirs containing various liquids, in particular, oil and gas fields, aquifers, as a rule, have heterogeneous reservoir properties. The latter determines the nature of the movement of liquids in them and the transfer of various substances along with liquids. In inhomogeneous media, as a rule, the laws of Darcy and Fick are violated [1, 2]. A typical example of an inhomogeneous medium is fractured - porous media, where the movement of fluid and the solute transport occurs with the manifestation of a delay due to fluid exchange [3, 4].

Recently, to simulate the effects of delay in the transfer of liquid and solute in inhomogeneous media, an approach has been used that takes into account fractional derivatives both in time and in space variable in the equations of filtration and transfer of substances [5–7].

In [6, 8], a model of substance transfer in a fractured porous medium is presented, where a zone changed under the influence of the substance is formed around the cracks, the diffusion characteristics of which differ from the unchanged zone of blocks.

The advanced propagation of substances in a porous medium can be the result of many factors. Therefore, there are certain difficulties in mathematical modeling of this

phenomenon. Some models in this direction were presented in [10, 12–14, 17]. Mass transfer between the zones is modeled by a first-order kinetic equation [9, 15]. Another approach, combining kinetic and linear mass transfer between zones, was proposed in [11].

In [20], an inhomogeneous two-zone medium is considered as a single-zone medium with some source (sink). The second zone is modeled through the source (sink). This approach is fundamentally new, because, in fact, a bicontinuum medium is represented as a monocontinuum. The validity of this approach is substantiated by approximation of the results based on the monocontinuum approach with the corresponding results of the bicontinuum approach.

Unlike [21], the paper proposes a new model [20], where the presence of the second zone of an inhomogeneous medium is taken into account as a sink (source) term in the transport equation written for the first zone. The sink term is presented as a fractional time derivative of the solute concentration in the first zone with a certain coefficient. Thus, [20] approach is monocontinuous, while in [21] the bicontinual approach is used.

Here, an inhomogeneous porous medium is considered, consisting of well-permeable (transit) and poorly permeable (stagnant) zones, where the process of solute transport occurs in the transit zone, and mass transfer occurs between the zones due to the resulting concentration gradients. The solute transport into the medium is allowed when it is permeable. Without specifying the geometric characteristics of the stagnant zone, the mass transfer between the zones, as well as the diffusion flow of solute in the equation for the solute transport, is modeled by the source term in the form of a fractional derivative of the concentration of solute with respect to time. This makes it possible to take into account anomalous phenomena in the process of solute transport. On the basis of the numerical solution of the problem of solute transport, profiles of the concentration of solute are constructed for various orders of fractional derivatives included in the mathematical model and characterizing the anomalous nature of the transfer, as well as the nature of mass transfer between media. Note that some problems of the solute transport in such media with and without specifying the geometric characteristics of the zones without taking into account anomalous phenomena were considered in [16–19].

2 Mathematical Model

In accordance with the formulated conditions, the solute transport in the one-dimensional case can be described by the equation [6, 7]

$$m \frac{\partial \tilde{c}}{\partial t} + v \frac{\partial \tilde{c}}{\partial x} + a_2 \frac{\partial^\gamma \tilde{c}}{\partial t^\gamma} + a_3 \frac{\partial^\beta \tilde{c}}{\partial t^\beta} = \frac{\partial}{\partial x}(mJ), \quad (1)$$

$$J = D \left(p \frac{\partial^\alpha \tilde{c}}{\partial x^\alpha} + (1-p) \frac{\partial^\alpha \tilde{c}}{\partial (-x)^\alpha} \right), \quad (2)$$

where \tilde{c} - volumetric concentration of the substance, t - time, s , J - relative diffusion mass flow, M/c , m - porosity of the transition zone, D - coefficient of effective diffusion, $M^{1+\alpha}/c$, a_3 - retardation factor due to the transfer of a substance into the medium, $c^{\beta-1}$, a_2 - retardation factor related to the mass transfer between the two zones, $c^{\gamma-1}$, v - filtration

rate m/c , α , β , γ – order of derivatives ($0 < \alpha \leq 1$, $0 < \beta < 1$, $0,5 \leq \gamma \leq 1$), p ($0 \leq p \leq 1$) characterizes the proportion of advanced and lagging dispersion from the central symmetry. If $p < \frac{1}{2}$, the dispersion has a character lagging behind the symmetry with the formation of a tail. If $p > \frac{1}{2}$ the dispersion has a forward slope with respect to symmetry with the formation of a fast front and a relatively short tail [5].

Let a semi-infinite porous medium be filled with a liquid with a volume concentration \tilde{c}_0 , from the moment $t > 0$ it begins to receive a liquid with a volume concentration of the substance \tilde{c}_1 . At infinity, the original concentration is preserved \tilde{c}_0 . Then the initial and boundary conditions have the form

$$\tilde{c}(0, x) = \tilde{c}_0, \quad 0 \leq x < \infty, \quad (3)$$

$$\tilde{c}(t, 0) = \tilde{c}_1, \quad \tilde{c}(t, \infty) = \tilde{c}_0. \quad (4)$$

Equations (1), (2) are reduced to a dimensionless form. To do this, it is necessary to introduce characteristic scales. Dimensionless variables can be introduced as follows:

$$\begin{aligned} X &= \frac{x}{x_m}, \quad C = \frac{\tilde{c}}{\tilde{c}_m}, \quad \bar{t} = \frac{t}{t_m}, \quad Pe = \frac{x_m^{1+\alpha}}{t_m D}, \\ b_3 &= \frac{a_3}{m t_m^{\beta-1}}, \quad b_2 = \frac{a_2}{m t_m^{\gamma-1}}, \quad V = \frac{v t_m}{m x_m}, \end{aligned} \quad (5)$$

where t_m is the characteristic time for the transfer of a substance, x_m is the characteristic length for the processes of the transfer of a substance, \tilde{c}_m is the characteristic concentration.

The dimensionless equation of solute transport corresponding to (1) has the form

$$\frac{\partial C}{\partial \bar{t}} + V \frac{\partial C}{\partial X} + b_2 \frac{\partial^\gamma C}{\partial \bar{t}^\gamma} + b_3 \frac{\partial^\beta C}{\partial \bar{t}^\beta} = \frac{1}{Pe} \frac{\partial}{\partial X} \left(p \frac{\partial^\alpha C}{\partial X^\alpha} + (1-p) \frac{\partial^\alpha C}{\partial (-X)^\alpha} \right). \quad (6)$$

The initial and boundary conditions (3), (4) are also reduced to the dimensionless form

$$C(0, X) = C_0, \quad 0 \leq \infty, \quad C_0 = \frac{\tilde{c}_0}{\tilde{c}_m}, \quad (7)$$

$$C(\bar{t}, 0) = C_1, \quad C(\bar{t}, \infty) = C_0, \quad C_1 = \frac{\tilde{c}_1}{\tilde{c}_m}. \quad (8)$$

In the case $\alpha = 1$, $p = 1$ of Eq. (6) goes to

$$\frac{\partial C}{\partial \bar{t}} + V \frac{\partial C}{\partial X} + b_2 \frac{\partial^\gamma C}{\partial \bar{t}^\gamma} + b_3 \frac{\partial^\beta C}{\partial \bar{t}^\beta} = \frac{1}{Pe} \frac{\partial^2 C}{\partial X^2}. \quad (9)$$

When $p = 1$ from (6) we have

$$\frac{\partial C}{\partial \bar{t}} + V \frac{\partial C}{\partial X} + b_2 \frac{\partial^\gamma C}{\partial \bar{t}^\gamma} + b_3 \frac{\partial^\beta C}{\partial \bar{t}^\beta} = \frac{1}{Pe} \frac{\partial^{1+\alpha} C}{\partial X^{1+\alpha}}. \quad (10)$$

Equation (6) takes into account the effects of anomalous mass transfer between two media, the outflow of matter into the medium, and the anomalous diffusion solute transport. In Eq. (9), as a special case of (6), the diffusion transfer of matter has no anomaly. In (10), in contrast to (6), only the leading diffusion flow is taken into account, but with allowance for the anomaly.

3 Numerical Solution

To solve Eqs. (9), (10) with conditions (7), (8), we use the finite difference method [22]. In the area $D = \{0 \leq X < \infty, 0 \leq \bar{t} \leq \bar{t}_{\max}\}$ where \bar{t}_{\max} is the maximum time during which the process is studied, we introduce a grid with a step h in direction and τ in time. As a result, we have a grid: $\omega_{h\tau} = \{(X_i, \bar{t}_j), i = 0, 1, 2, \dots, j = 0, 1, \dots, J; X_i = ih; \bar{t}_j = j\tau; \tau = T/K\}$.

Equation (9) and we approximate on a grid $\omega_{h\tau}$ using an implicit difference scheme

$$\begin{aligned} & \frac{C_i^{j+1} - C_i^j}{\tau} + V \frac{C_i^{j+1} - C_{i-1}^{j+1}}{h} + \frac{b_2 \tau^{1-\gamma}}{\Gamma(2-\gamma)} \left[\sum_{k=0}^{j-1} \frac{C_i^{k+1} - C_i^k}{\tau} ((j-k+1)^{1-\gamma} - (j-k)^{1-\gamma}) + \right. \\ & \left. + \frac{(C_i^{j+1} - C_i^j)}{\tau} \right] \\ & + \frac{b_3 \tau^{1-\beta}}{\Gamma(2-\beta)} \left[\sum_{k=0}^{j-1} \frac{C_i^{k+1} - C_i^k}{\tau} ((j-k+1)^{1-\beta} - (j-k)^{1-\beta}) + \frac{(C_i^{j+1} - C_i^j)}{\tau} \right] \\ & = \frac{1}{Pe} \frac{C_{i+1}^{j+1} - 2C_i^{j+1} + C_{i-1}^{j+1}}{h^2}, \quad i = \overline{1, I-1}, j = \overline{0, J}, \end{aligned} \quad (11)$$

where C_i^j is the grid function corresponding to C , $\Gamma(\cdot)$ - Gamma function.

The initial and boundary conditions are approximated as

$$C_i^0 = C_0, \quad i = 0, 1, \dots, I, \quad (12)$$

$$C_0^j = C_1, \quad C_I^j = C_0, \quad j = 0, 1, \dots, J, \quad (13)$$

where I is a sufficiently large number for which the second of conditions (13) is approximately satisfied.

Difference scheme (11) is reduced to a system of linear equations

$$AC_{i-1}^{j+1} - BC_i^{j+1} + EC_{i+1}^{j+1} = -F_i^j, \quad i = \overline{1, I-1}, j = \overline{0, J}, \quad (14)$$

where

$$\begin{aligned} A &= \frac{V\tau}{h} + \frac{\tau}{Peh^2}, \quad B = 1 + \frac{V\tau}{h} + \frac{b_2 \tau^{1-\gamma}}{\Gamma(2-\gamma)} + \frac{b_3 \tau^{1-\beta}}{\Gamma(2-\beta)} + \frac{2\tau}{Peh^2}, \quad E = \frac{\tau}{Peh^2} \\ F_i^j &= \left(1 + \frac{b_2 \tau^{1-\gamma}}{\Gamma(2-\gamma)} + \frac{b_3 \tau^{1-\beta}}{\Gamma(2-\beta)} \right) C_i^j - \frac{b_2 \tau^{1-\gamma}}{\Gamma(2-\gamma)} \left[\sum_{k=0}^{j-1} ((j-k+1)^{1-\gamma} - (j-k)^{1-\gamma}) C_i^{k+1} - \right. \\ & \left. - ((j-k+1)^{1-\gamma} - (j-k)^{1-\gamma}) C_i^k \right] \\ & - \frac{b_3 \tau^{1-\beta}}{\Gamma(2-\beta)} \left[\sum_{k=0}^{j-1} ((j-k+1)^{1-\beta} - (j-k)^{1-\beta}) C_i^{k+1} - ((j-k+1)^{1-\beta} - (j-k)^{1-\beta}) C_i^k \right]. \end{aligned}$$

We solve system (14) with respect to C_i^{j+1} the sweep method for known C_i^j . Equation (10) and we approximate on a grid $\omega_{h\tau}$ using an explicit difference scheme

$$\begin{aligned} & \frac{C_i^{j+1} - C_i^j}{\tau} + V \frac{C_i^j - C_{i-1}^j}{h} + \frac{b_2}{\Gamma(2-\gamma)} \left[\sum_{k=0}^{j-1} \frac{C_i^{k+1} - C_i^k}{\tau} ((j-k+1)^{1-\gamma} - (j-k)^{1-\gamma}) + \right. \\ & \left. + \frac{(C_i^{j+1} - C_i^j)\tau^{1-\gamma}}{\tau} \right] + \\ & + \frac{b_3}{\Gamma(2-\beta)} \left[\sum_{k=0}^{j-1} \frac{C_i^{k+1} - C_i^k}{\tau} ((j-k+1)^{1-\beta} - (j-k)^{1-\beta}) + \frac{(C_i^{j+1} - C_i^j)\tau^{1-\beta}}{\tau} \right] \\ & = \frac{1}{Pe} \frac{h^{-\alpha}}{(3-\alpha)} \sum_{k=0}^{i-1} [C_{i-k+1}^j - 2C_{i-k}^j + C_{i-k-1}^j] [(k+1)^{2-\alpha} - k^{2-\alpha}]. \end{aligned} \quad (15)$$

We introduce the following notation

$$\begin{aligned} Q_\alpha^k &= \frac{h^{-\alpha}}{(3-\alpha)} \sum_{k=0}^{i-1} [C_{i-k+1}^j - 2C_{i-k}^j + C_{i-k-1}^j] [(k+1)^{2-\alpha} - k^{2-\alpha}], \\ Q_\beta^k &= \frac{b_3}{\Gamma(2-\beta)} \left[\sum_{k=0}^{j-1} \frac{C_i^{k+1} - C_i^k}{\tau} ((j-k+1)^{1-\beta} - (j-k)^{1-\beta}) \right], \\ Q_\gamma^k &= \frac{b_2}{\Gamma(2-\gamma)} \left[\sum_{k=0}^{j-1} \frac{C_i^{k+1} - C_i^k}{\tau} ((j-k+1)^{1-\gamma} - (j-k)^{1-\gamma}) \right], \\ Q_\alpha &= \frac{h^{-\alpha}}{(3-\alpha)}, \quad Q_\beta = \frac{b_3}{\Gamma(2-\beta)}, \quad Q_\gamma = \frac{b_2}{\Gamma(2-\gamma)}. \end{aligned} \quad (16)$$

Taking into account these notations, we write the solution of Eq. (15) in the form

$$\begin{aligned} C_i^{j+1} &= Q_\alpha^k / (1/\tau + Q_\beta + Q_\gamma) - ((V/h - 1/\tau - Q_\beta - Q_\gamma) / (1/\tau + Q_\beta + Q_\gamma)) C_i^j \\ &+ V / (h(1/\tau + Q_\beta + Q_\gamma)) C_{i-1}^j - Q_\gamma^k / (1/\tau + Q_\beta + Q_\gamma) - Q_\beta^k / (1/\tau + Q_\beta + Q_\gamma). \end{aligned} \quad (17)$$

4 Numerical Calculations

The following values of the initial parameters were used in the calculations: $C_0 = 0$, $C_1 = 1$, $V = 0,01$, $Pe = 10^4$, $\alpha = 1$ and various β , γ , b_2 , b_3 .

Some representative results are shown in Figs. 1–3. As can be seen from the results, a decrease in the indicator β from one leads to a lag in the development of profiles (Fig. 1a). Accounting for mass transfer between transit and stagnant zones, which is observed at $b_2 \neq 0$, leads to an even greater lag in the development of profiles (Fig. 1b). It should be noted that when the solute transport into the environment and mass transfer between zones are taken into account, the influence of the parameter β weakens. In Fig. 1a, the effect of β on the concentration profiles is more significant than in Fig. 1b. Obviously,

at relatively long times, \bar{t} is the transient process ends and the influence of the terms $b_3 \frac{\partial^\beta C}{\partial \bar{t}^\beta}$ and $b_2 \frac{\partial^\gamma C}{\partial \bar{t}^\gamma}$ on the concentration profiles disappears.

Similarly, the influence of the index γ on the concentration profiles was studied (Fig. 2). Like β , a decrease in the parameter values γ from unity leads to a delayed development of concentration profiles. An increase in values b_2 leads to an increase in the solute transport from the transit zone to the stagnant one. Therefore, there is a slow development of C profiles. From a physical point of view, this can be explained as the absorption of matter into stagnant zones.

Figure 3 shows the profiles C for various b_2 and b_3 at fixed β and γ . As expected, an increase in b_1 and b_3 enhances the role of the substance transfer to the environment and mass transfer to the stagnant zone. Due to this, there is a lag in the development of substance concentration profiles in the transit zone. One can notice a decrease in the zone of change in the concentrations of a substance both with an increase b_3 in and b_2 . In this regard, their effect on the characteristics of the solute transport is the same.

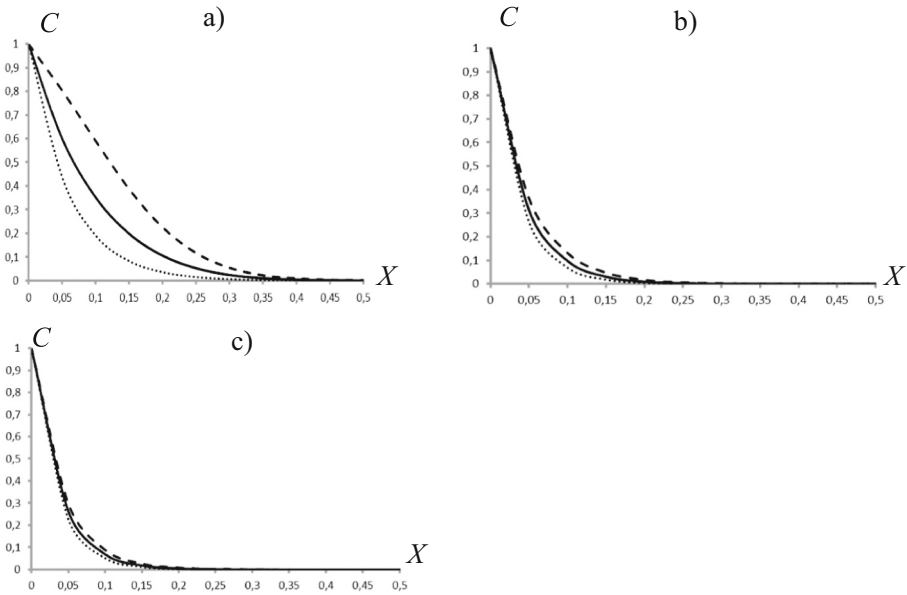


Fig. 1. Concentration profiles at $b_3 = 0, 4, \alpha = 1, b_2 = 0$ (a); 0,5 (b); 0,7 (c); $\beta=0,5$ (.....), 0,7(——), 0,9(---).

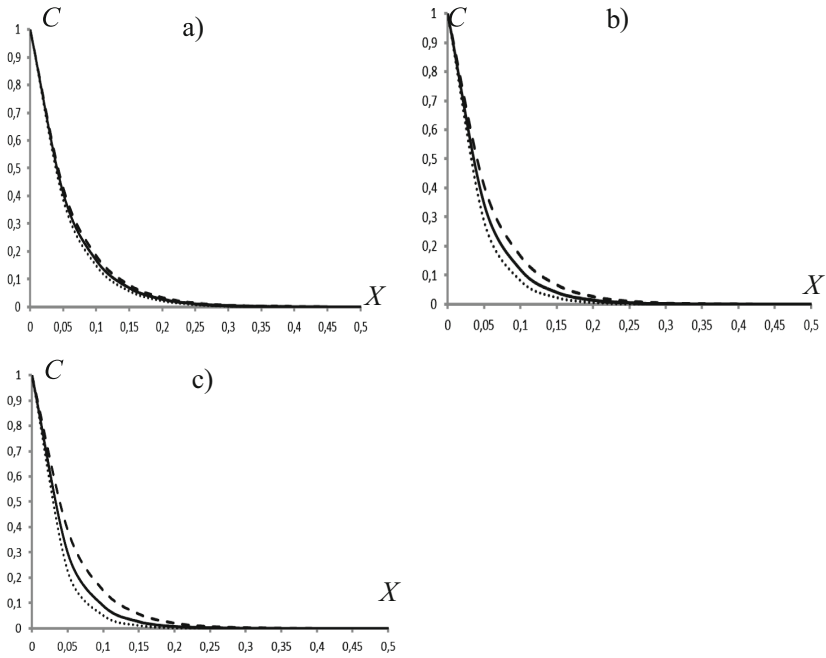


Fig. 2. Concentration profiles at $b_3 = 0, 1$ (a); 0,4 (b); 0,7 (c); $\gamma = 0,5$ (.....), 0,7(——), 0,9(— — —).

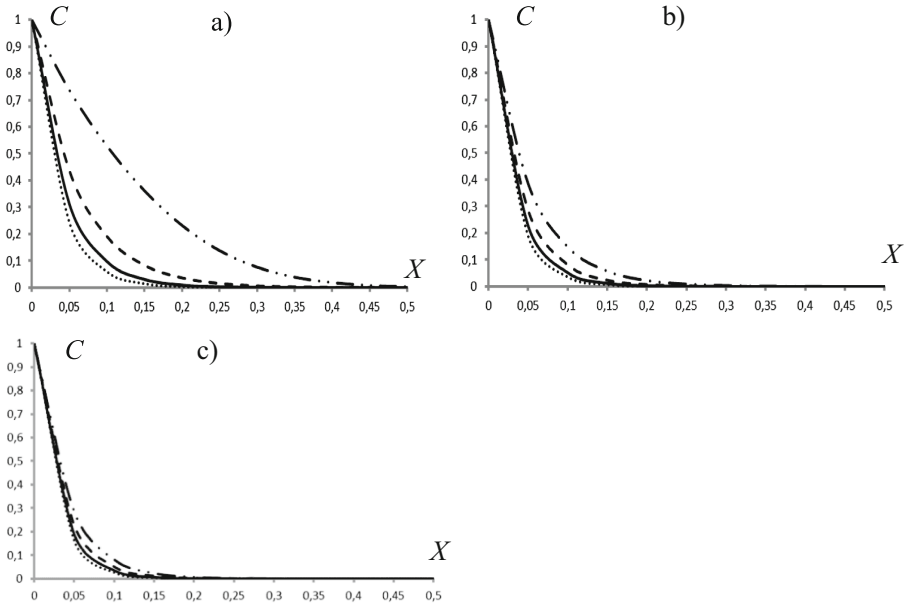


Fig. 3. Concentration profiles at $\beta = 0, 5$, $\gamma = 0, 5$, $b_2 = 0$ (a); 0,4 (b); 0,7 (c); $b_3 = 0,1$ (— · — · —), 0,4(— — —), 0,7(——), 1,0(.....).

5 Conclusion

An analysis of the anomalous solute transport in a two-zone inhomogeneous medium shows that the absorption of matter into stagnant zones and outflow into the medium significantly slows down the process of the spread of matter in the transit zone. The anomalous solute transport to stagnant zones and the environment can be effectively modeled by the sink terms in the corresponding transport equations as fractional derivatives of the substance concentration versus time. Depending on the orders of these derivatives and retardation factors, one can obtain various effects of substance absorption, and as a result, various characteristics of transport in cracks.

The obtained results show the previously shown [20] fundamental possibility of replacing a bicontinuum with a monocontinuum by considering the second continuum as a sink (or source) for the first continuum.

Acknowledgement. The work was supported by the Ministry of Innovative Development Republic of Uzbekistan (Grant OT-F4–64).

References

1. Fetter, C.W.: Contaminant Hydrogeology, pp. 58–70. Prentice-Hall, Inc.: Upper Saddle River, NJ, USA (1999)
2. Fetter, C.W.: Applied Hydrogeology, 3rd edn. Upper Saddle River, New Jersey: Prentice Hall (2001)
3. Barenblatt, G.I., Entov, V.M., Ryzhik, V.M.: Theory of Fluid Flow Through Natural Rocks. Kluwer Academic, Dordrecht, The Netherlands (1990)
4. Chen, Z.-X.: Transient flow of slightly compressible fluids through double-porosity, double-permeability systems – A state-of-the-Art Review. *Transport in Porous Media* **4**, 147–184 (1989)
5. Benson, D.A., Wheatcraft, S.W., Meerschaert, M.M.: Application of a fractional advection–dispersion equation. *Water Resour. Res.* **36**, 1403–1412 (2000)
6. Fomin, S.A., Chugunov, V.A., Hashida, T.: Non-Fickian mass transport in fractured porous media. *Advances in Water Resources*. **34**(2), 205–214 (2011)
7. Suzuki, A., Horne, R.N., Makita, H., Niibori, Y., Fomin, S.A., Chugunov, V.A., Hashida, T.: Development of fractional derivative - based mass and heat transport model. In: *Proceedings, Thirty-Eighth Workshop on Geothermal Reservoir Engineering* Stanford University, Stanford, California, February 11–13. 2013. SGP-TR-198
8. Fomin, S.A., Chugunov, V.A.: Hashida T. Application of Fractional Differential Equations for Modeling the Anomalous Diffusion of Contaminant from Fracture into Porous Rock Matrix with Bordering Alteration Zone. *Transport in Porous Media*. **81**, 187–205 (2010)
9. Coats, K.H., Smith, B.D.: Dead-end pore volume and dispersion in porous media. *Soc. Pet. Eng. J.* **4**, 73–84 (1964)
10. Gerke, H.H., van Genuchten, M.T.: Macroscopic representation of structural geometry for simulating water and solute movement in dualporosity media. *Adv. Water Resources*. **19**, 343–357 (1996)
11. Leij, F.J., Bradford, S.A.: Combined physical and chemical nonequilibrium transport model: analytical solution, moments, and application to colloids. *J. Contaminant Hydrol.* **110**, 87–99 (2009)

12. Selim, H.M., Ma, L.: *Physical Nonequilibrium in Soils: Modeling and Applications*. Ann Arbor Press, Chelsea, MI (1998)
13. Simunek, J., van Genuchten, M.Th.: Modeling nonequilibrium flow and transport processes using HYDRUS. *Vadose Zone J.* **7**, 782–797 (2008)
14. Toride, N., Leij, F.J., van Genuchten, M.Th.: The CXTFIT code for estimating transport parameters from laboratory or field tracer experiments. Version 2.0. Res. Rep. 137. US Salinity Lab, Riverside, CA. 1995
15. Van Genuchten, M.Th., Wierenga, P.J.: Mass Transfer Studies in Sorbing Porous media. I. Analytical Solution. *Soil Sci. Soc. Am. J.* **40**(4), 473–480 (1976)
16. Khuzhayorov, B.Kh., Makhmudov, J.M.: Flow of Suspensions in 2D Porous Media with Mobile and Immobile Liquid Zones. *J. Porous Media* **13**(5), 423–437 (2010)
17. Khuzhayorov, B.Kh.: *Filtration of inhomogeneous liquids in porous media*. Tashkent: Fan, 280 p. (2012)
18. Khuzhayorov, B.Kh., Makhmudov, J.M.: Mathematical models of filtration of inhomogeneous liquids in porous media. Tashkent: Fan, 20 14. – 280 s
19. Khuzhayorov, B., Makhmudov, J.M., Zikiryayev, S.: Solute transport in a porous medium saturated with mobile and immobile liquid. *Eng. Phys. J.* **83**(2), 248–254 (2010)
20. Khuzhayorov, B.Kh., Jiyanov, T.O.: Solute transport with nonequilibrium adsorption in a non-uniform porous medium. *Problems of Computational and Applied Mathematics*, No. 3(9). 63–70 (2017)
21. Leij, F.L., Bradford, S.A.: Colloid transport in dual-permeability media. *J. Contaminant Hydrology* **150**, 65–76 (2013)
22. Samarsky, A.A.: *Theory of difference schemes*. M. Science, 656 p. (1977)



Determination of Relaxation and Flow Coefficients During Filtration of a Homogeneous Liquid in Fractured-Porous Media

Erkin Kholiyarov¹(✉) and Mirzohid Ernazarov²

¹ Termez University of Economics and Service, Termez, Uzbekistan
e.kholiyarov@mail.ru

² Termez State University, Termez, Uzbekistan

Abstract. In this paper, we posed and numerically solved the inverse problem of determining the relaxation coefficient and the flow coefficient in a simplified model of relaxation filtration of a homogeneous fluid in fractured-porous media. To solve the problem, a regularizing identification method was used. It has been established that these coefficients, at various initial approximations with unperturbed initial data, are restored in almost ten to twenty iterations. At a more remote initial approximation from the equilibrium point, the iterative procedure converges to the shifted values of the desired parameters. The optimal range of regularization parameter values is determined.

Keywords: filtration · fluid · homogeneous fractured-porous medium · inverse problem · permeability · porosity · pressure · regularization · relaxation

1 Introduction

The most productive oil fields in the world consist mainly of carbonate reservoirs [1, 2]. Typically, carbonate reservoirs are fractured, fractured-porous and fractured-cavernous types [1, 2]. Geological justifications for the appearance of a crack in reservoirs are given in [2].

The theoretical foundations for the filtration of homogeneous liquids in fractured-porous media were developed in [3, 4]. During unsteady filtration in fractured-porous media [3, 4], fluid is exchanged between porous blocks and fractures. The generalized filtration equation for homogeneous liquids in fractured-porous media has the form [3, 4]

$$\begin{cases} \beta_1^* \frac{\partial p_1}{\partial t} = \frac{k_1}{\mu} \frac{\partial^2 p_1}{\partial x^2} + \frac{\alpha_0}{\mu} (p_2 - p_1), \\ \beta_2^* \frac{\partial p_2}{\partial t} = \frac{k_2}{\mu} \frac{\partial^2 p_2}{\partial x^2} - \frac{\alpha_0}{\mu} (p_2 - p_1), \end{cases} \quad (1)$$

where k_l is permeability, m^2 ; p_l is pressure, MPa; t is time, s; x is coordinate, m; α_0 is flow coefficient, which characterizes the exchange of fluid between fractures and porous

blocks; $\beta_l^* = \beta_{cl} + m_{0l}\beta_f$, $m_l = m_{0l} + \beta_{cl}(p_l - p_0)$; m_l is porosity; m_{0l} is porosity at $p_l = p_0$; μ is the viscosity of the liquid, MPa s; the index $l = 1$ corresponds to cracks, $l = 2$ porous blocks.

Under the conditions $m_1 \ll m_2$, $\beta_{c1} \ll \beta_{c2}$, $k_2 \ll k_1$ from system (1) a simplified system of equations can be obtained

$$\begin{cases} \frac{k_1}{\mu\beta_2^*} \frac{\partial^2 p_1}{\partial x^2} + \frac{\alpha_0}{\mu\beta_2^*} (p_2 - p_1) = 0, \\ \beta_2^* \frac{\partial p_2}{\partial t} + \frac{\alpha_0}{\mu} (p_2 - p_1) = 0. \end{cases} \quad (2)$$

If the compressibility of cracks is taken into account, but there is no fluid movement in porous blocks, from (1) the following system of equations is obtained [5]

$$\begin{cases} \beta_1^* \frac{\partial p_1}{\partial t} = \frac{k_1}{\mu} \frac{\partial^2 p_1}{\partial x^2} + \frac{\alpha_0}{\mu} (p_2 - p_1), \\ \beta_2^* \frac{\partial p_2}{\partial t} + \frac{\alpha_0}{\mu} (p_2 - p_1) = 0. \end{cases} \quad (3)$$

In the literature, this system of equations is called Warren-Root or “truncated”.

During the filtration of heavy oils and their mixtures with gases, relaxation phenomena are observed due to the nonequilibrium of the filtration flow [6]. Simulation of relaxation filtration of homogeneous liquids is reflected in the works [7–12]. Some problems of relaxation fluid filtration are considered in [13–15].

In [16, 17], Eqs. (1), (2), (3) are generalized taking into account Darcy relaxation laws in cracks and porous blocks. We present these equations below. The generalized system of Eqs. (1) for the case of pressure gradient relaxation has the form [16, 17]

$$\begin{cases} \frac{\partial p_1}{\partial t} = \frac{k_1}{\mu\beta_1^*} \left(\frac{\partial^2 p_1}{\partial x^2} + \lambda_1 \frac{\partial^3 p_1}{\partial x^2 \partial t} \right) + \frac{\alpha_0}{\mu\beta_1^*} (p_2 - p_1), \\ \frac{\partial p_2}{\partial t} = \frac{k_2}{\mu\beta_2^*} \left(\frac{\partial^2 p_2}{\partial x^2} + \lambda_2 \frac{\partial^3 p_2}{\partial x^2 \partial t} \right) - \frac{\alpha_0}{\mu\beta_2^*} (p_2 - p_1), \end{cases} \quad (4)$$

where λ_1 , λ_2 are relaxation of the pressure gradient, in cracks and porous blocks respectively.

Assuming $m_1 \ll m_2$, $k_2 \ll k_1$ from (4) one can obtain a “truncated” system [16, 17]

$$\begin{cases} \frac{\partial p_1}{\partial t} = \frac{k_1}{\mu\beta_1^*} \left(\frac{\partial^2 p_1}{\partial x^2} + \lambda_1 \frac{\partial^3 p_1}{\partial x^2 \partial t} \right) + \frac{\alpha_0}{\mu\beta_1^*} (p_2 - p_1), \\ \frac{\partial p_2}{\partial t} + \frac{\alpha_0}{\mu\beta_2^*} (p_2 - p_1) = 0. \end{cases} \quad (5)$$

Under the condition, $\beta_1^* \ll \beta_2^*$ from the “truncated” system (5) one can obtain a simplified system [16, 17]

$$\begin{cases} \frac{k_1}{\mu} \left(\frac{\partial^2 p_1}{\partial x^2} + \lambda_1 \frac{\partial^3 p_1}{\partial x^2 \partial t} \right) + \frac{\alpha_0}{\mu} (p_2 - p_1) = 0, \\ \beta_2^* \frac{\partial p_2}{\partial t} + \frac{\alpha_0}{\mu} (p_2 - p_1) = 0, \quad 0 < x < L, \quad 0 < t \leq T. \end{cases} \quad (6)$$

Efficient numerical methods for solving inverse problems for partial differential equations are given in [18–23]. Some inverse problems of fluid filtration are considered in [24–29]. Note that for the model [16, 17] the inverse problems have not yet been solved.

In this paper, based on (6), we consider the inverse problem of determining the pressure gradient relaxation coefficient λ_1 and the flow coefficient α_0 . First, the formulation of the problem and then its solution algorithm is given. To minimize the residual functional, an iterative procedure with unknown coefficients is used. As a criterion for stopping the iterative procedure, both the minimization of the value of the functional and the convergence of the iterative values of the coefficients to the limit value are used. The initial data for solving the inverse problem were prepared from the solution of the corresponding direct problem, i.e. a “quasi-real” computational experiment is carried out. Solutions of the inverse problem are also obtained for artificially perturbed initial data.

2 Statement of the Inverse Problem

An inverse problem is considered to determine the relaxation coefficients of the pressure gradient λ_1 and flow α_0 in the case when the filtration process is described by a simplified system of Eqs. (6) with initial

$$p_1(0, x) = p_2(0, x) = p_0, \quad p_0 = \text{const}, \quad 0 \leq x \leq L, \quad (7)$$

and boundary conditions

$$-\frac{k_1}{\mu} \left(\frac{\partial p_1}{\partial x} + \lambda_1 \frac{\partial^2 p_1}{\partial t \partial x} \right) \Big|_{x=0} = v_0 = \text{const}, \quad p_1(t, L) = p_0, \quad 0 < t \leq T. \quad (8)$$

Additionally, we know the pressure change in the well:

$$p_1(t, 0) = z(t). \quad (9)$$

Solution of the inverse problem (6)–(9) determine the coefficients of pressure relaxation λ_1 and flow α_0 , leads to minimizing the discrepancy functional

$$J(\boldsymbol{\gamma}) = \int_0^T [p_1(t, 0) - z(t)]^2 dt, \quad \boldsymbol{\gamma} = (\lambda_1, \alpha_0), \quad (10)$$

where $z(t)$ is the observed values of the bottom hole pressure, $p_1(t, 0)$ is the calculated values of the bottom hole pressure.

3 Solution of the Inverse Problem

Iterative methods often used to solve inverse problems become divergent at certain parameter values. Therefore, to solve inverse problems, it is advisable to use regularized iterative methods. For this reason, on each iterative layer, instead of functional (10), the following modified discrepancy functional is used [18]:

$$J_M(\boldsymbol{\gamma}^{s+1}) = J(\boldsymbol{\gamma}^s) + \alpha(\boldsymbol{\gamma}^{s+1} - \boldsymbol{\gamma}^s)^2, \quad (11)$$

α is parameter regularization.

Stationarity conditions for the functional (11) has the form

$$\frac{dJ_M(\boldsymbol{\gamma}^{s+1})}{d\boldsymbol{\gamma}} = 2 \int_0^T \left[\overset{s}{p}_1(t, 0) - z(t) \right] \mathbf{w}^s(t, 0) dt + 2\alpha(\boldsymbol{\gamma}^{s+1} - \boldsymbol{\gamma}^s) = 0. \quad (12)$$

Function p_1 in the neighborhood $\boldsymbol{\gamma}^s$ in a series up to terms of the second order [18]

$$p_1^{s+1}(t, x) \approx p_1^s(t, x) + (\boldsymbol{\gamma}^{s+1} - \boldsymbol{\gamma}^s) \mathbf{w}^s(t, x), \quad (13)$$

where \mathbf{w} is column vector, $\boldsymbol{\gamma}$ is row vector

$$\frac{\partial \overset{s}{p}_1}{\partial \boldsymbol{\gamma}} = \mathbf{w}^s = \left(\overset{s}{w}_{11}, \overset{s}{w}_{12} \right)^T = \left(\frac{\partial \overset{s}{p}_1}{\partial \lambda_1^s}, \frac{\partial \overset{s}{p}_1}{\partial \alpha_0^s} \right)^T, \quad \boldsymbol{\gamma}^s = (\lambda_1^s, \alpha_0^s), \quad (14)$$

where $w_{11} = \frac{\partial p_1}{\partial \lambda_1}$, $w_{12} = \frac{\partial p_1}{\partial \alpha_0}$ are the sensitivity functions [18, 19] with respect to the coefficients λ_1 and α_0 respectively.

Substituting expansion (13) into relation (12), we obtain the following system of linear algebraic equations with respect to λ_1^{s+1} , α_0^{s+1} :

$$\begin{cases} a_{11}\lambda_1^{s+1} + a_{12}\alpha_0^{s+1} = b_1, \\ a_{21}\lambda_1^{s+1} + a_{22}\alpha_0^{s+1} = b_2, \end{cases} \quad (15)$$

where

$$\begin{aligned}
a_{11} &= \int_0^T w_{11}^2{}^s(t, 0)dt + \alpha, \quad a_{22} = \int_0^T w_{12}^2{}^s(t, 0)dt + \alpha, \quad a_{12} = a_{21} = \int_0^T w_{11}{}^s(t, 0)w_{12}{}^s(t, 0)dt, \\
b_1 &= \int_0^T [w_{11}{}^s(t, 0)\lambda_1^s + w_{12}{}^s(t, 0)\alpha_0^s - p_1{}^s(t, 0) + z(t)]w_{11}{}^s(t, 0)dt + \alpha\lambda_1^s, \\
b_2 &= \int_0^T [w_{11}{}^s(t, 0)\lambda_1^s + w_{12}{}^s(t, 0)\alpha_0^s - p_1{}^s(t, 0) + z(t)]w_{12}{}^s(t, 0)dt + \alpha\alpha_0^s.
\end{aligned} \tag{16}$$

The next approximations λ_1^{s+1} , α_0^{s+1} can be easily determined from system (15) using the well-known Cramer formulas in the form

$$\lambda_1^{s+1} = \frac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}, \quad \alpha_0^{s+1} = \frac{\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}}{\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}}. \tag{17}$$

Let us first differentiate problem (6)–(8) with respect to λ_1 and obtain the following problem

$$\begin{cases} \frac{k_1}{\mu} \left(\frac{\partial^2 w_{11}}{\partial x^2} + \lambda_1 \frac{\partial^3 w_{11}}{\partial x^2 \partial t} + \frac{\partial^3 p_1}{\partial x^2 \partial t} \right) + \frac{\alpha_0}{\mu} (w_{21} - w_{11}) = 0, \\ \beta_2^* \frac{\partial w_{21}}{\partial t} + \frac{\alpha_0}{\mu} (w_{21} - w_{11}) = 0, \quad 0 < x < L, \quad 0 < t \leq T, \end{cases} \tag{18}$$

$$w_{11}(0, x) = w_{21}(0, x) = 0, \quad 0 \leq x \leq L, \tag{19}$$

$$-\frac{k_1}{\mu} \left(\frac{\partial w_{11}}{\partial x} + \lambda_1 \frac{\partial^2 w_{11}}{\partial t \partial x} + \frac{\partial^2 p_1}{\partial t \partial x} \right) \Big|_{x=0} = 0, \quad w_{11}(t, L) = 0, \quad 0 < t \leq T, \tag{20}$$

where $w_{21} = \frac{\partial p_2}{\partial \lambda_1}$ is the sensitivity function [18, 19] with respect to the coefficient λ_1 .

Similarly, we differentiate problem (6)–(8) with respect to the coefficient α_0 and obtain the following problem

$$\begin{cases} \frac{k_1}{\mu} \left(\frac{\partial^2 w_{12}}{\partial x^2} + \lambda_1 \frac{\partial^3 w_{12}}{\partial x^2 \partial t} \right) + \frac{\alpha_0}{\mu} (w_{22} - w_{12}) + \frac{1}{\mu} (p_2 - p_1) = 0, \\ \beta_2^* \frac{\partial w_{22}}{\partial t} + \frac{\alpha_0}{\mu} (w_{22} w_{12}) + \frac{1}{\mu} (p_2 - p_1) = 0, \quad 0 < x < L, \quad 0 < t \leq T, \end{cases} \tag{21}$$

$$w_{12}(0, x) = w_{22}(0, x) = 0, \quad 0 \leq x \leq L, \tag{22}$$

$$-\frac{k_1}{\mu} \left(\frac{\partial w_{12}}{\partial x} + \lambda_1 \frac{\partial^2 w_{12}}{\partial t \partial x} \right) \Big|_{x=0} = 0, \quad w_{12}(t, L) = 0, \quad 0 < t \leq T, \quad (23)$$

where $w_{22} = \frac{\partial p_2}{\partial \alpha_0}$ is the sensitivity function [18, 19] with respect to the coefficient α_0 .

Algorithm for determining the coefficients λ_1 and α_0 can be built like this:

1. We set some initial approximations $\lambda_1 = \lambda_1^0$ and $\alpha_0 = \alpha_0^0$ (we assume $s = 0$);
2. We solve problems (7)–(9), (19)–(21) and (22)–(24) from $t = 0$ to $t = T$ and calculate the functions $p_1^s(t, x)$, $p_2^s(t, x)$ and $w_{1k}^s(t, x)$, $w_{2k}^s(t, x)$, $k = 1, 2$.
3. We solve the system of Eqs. (16)–(18).
4. We $s = s + 1$ believe $\lambda_1 = \lambda_1^{s+1}$, $\alpha_0 = \alpha_0^{s+1}$;
5. Repeat steps 2, 3, 4 until the condition is met

$$\frac{|J^{s+1} - J^s|}{J^s} \leq \varepsilon, \quad \frac{|\lambda_1^{s+1} - \lambda_1^s|}{|\lambda_1^s|} \leq \varepsilon_1, \quad \frac{|\alpha_0^{s+1} - \alpha_0^s|}{|\alpha_0^s|} \leq \varepsilon_2,$$

where ε , ε_1 , ε_2 are fairly acceptable errors.

4 Difference Problems

Within the framework of the quasi-real quasi-experiment [20], the direct problem (7)–(9) with known values $\lambda_1 = \lambda_1^{\text{exact}} = 100$ s, $\alpha_0 = \alpha_0^{\text{exact}} = 3,6 \cdot 10^{-16}$ is first considered, which is solved by the finite difference method [30]. In the area $D = \{0 \leq x \leq L, 0 \leq t \leq T\}$, we introduce a uniform grid

$$\Omega_{ht} = \{(x_i, t_j), x_i = ih, i = 0, 1, \dots, N, h = L/N, t_j = j\tau, j = 0, 1, \dots, M, \tau = T/M\}$$

where h is the grid step in coordinate x , where τ is the grid step in time t . We introduce the notation $p1_i^j = p_1(t_j, x_i)$, $p2_i^j = p_2(t_j, x_i)$.

The difference approximation of the problem (7)–(9) has the form:

$$\begin{aligned} \frac{k_1}{\mu} \Delta p1_i^{j+1} + \frac{\lambda_1 k_1}{\mu \tau} [\Delta p1_i^{j+1} - \Delta p1_i^j] + \frac{\alpha_0}{\mu} (p2_i^j - p1_i^{j+1}) &= 0, \\ i = 1, 2, \dots, N-1, j = 0, 1, \dots, M-1, \\ \frac{p2_i^{j+1} - p2_i^j}{\tau} + \frac{\alpha_0}{\mu \beta_2^*} (p2_i^{j+1} - p1_i^{j+1}) &= 0, \quad i = 0, 1, \dots, N, j = 0, 1, \dots, M-1, \\ p1_i^0 &= p_0, \quad p2_i^0 = p_0, \quad i = 0, 1, \dots, N, \\ -\frac{k_1}{\mu} \left[\frac{p1_0^{j+1} - p1_1^{j+1}}{h} + \frac{\lambda_1}{\tau} \left(\frac{p1_0^{j+1} - p1_1^{j+1}}{h} - \frac{p1_0^j - p1_1^j}{h} \right) \right] &= v_0, \\ p1_N^{j+1} &= p_0, \quad j = 0, 1, \dots, M-1. \end{aligned} \quad (24)$$

where $\Delta p1_i^j = \frac{1}{h^2} (p1_{i-1}^j - 2p1_i^j + p1_{i+1}^j)$.

Difference Eqs. (24) are reduced to the form

$$Ap1_{i-1}^{j+1} - Cp1_i^{j+1} + Bp1_{i+1}^{j+1} = -F_i, \quad i = 1, 2, \dots, N-1, \quad j = 0, 1, \dots, M-1, \quad (25)$$

$$p2_i^{j+1} = \frac{p2_i^j + \frac{\alpha_0^s \tau}{\mu \beta_2^*} p2_i^{j+1}}{1 + \frac{\alpha_0^s \tau}{\mu \beta_2^*}}, \quad i = 1, 2, \dots, N-1, \quad j = 0, 1, \dots, M-1, \quad (26)$$

where

$$A = B = \frac{k_1}{\mu h^2} \left(1 + \frac{\lambda_1}{\tau} \right), \quad C = 2A + \frac{\alpha_0}{\mu}, \quad F_i = \frac{\alpha_0}{\mu} p2_i^j - \frac{\lambda_1 k_1}{\mu \tau h^2} \Lambda p1_i^j.$$

To solve problem (25), we use the Thomas algorithm [30]:

$$p1_i^{j+1} = \alpha_{i+1}^{(1)} p1_{i+1}^{j+1} + \beta_{i+1}^{(1)}, \quad i = N-1, \dots, 1, 0, \quad j = 0, 1, \dots, M-1, \quad p1_N^{j+1} = p_0, \quad (27)$$

$$\alpha_{i+1}^{(1)} = \frac{B}{C - A\alpha_i^{(1)}}, \quad \beta_{i+1}^{(1)} = \frac{A\beta_i^{(1)} + F_i}{C - A\alpha_i^{(1)}}, \quad i = 1, 2, \dots, N-1, \quad (28)$$

$$p1_0^{j+1} = \alpha_1^{(1)} p1_1^{j+1} + \beta_1^{(1)}, \quad p1_0^{j+1} = p1_1^{j+1} + \left(\frac{\lambda_1}{\tau} (p1_0^j - p1_1^j) - \frac{v_0 h \mu}{k_1} \right) / \left(1 + \frac{\lambda_1}{\tau} \right),$$

$$\alpha_1^{(1)} = 1, \quad \beta_1^{(1)} = \left(\frac{\lambda_1}{\tau} (p1_0^j - p1_1^j) - \frac{v_0 h \mu}{k_1} \right) / \left(1 + \frac{\lambda_1}{\tau} \right). \quad (29)$$

For the numerical solution of problem (6)–(9), the following values of the initial data were used: $T = 2000$ s, $L = 60$ m, $k_1 = 1 \cdot 10^{-12}$ m², $p_0 = 10$ MPa, $\mu = 2, 5 \cdot 10^{-8}$ MPa·s, $\beta_2^* = 1 \cdot 10^{-5}$ MPa⁻¹, $v_0 = 2 \cdot 10^{-6}$ m/s.

According to the results of numerical calculations, the grid function is set $z^j = z(t_j)$, $j = 0, 1, \dots, M$. Function $z(t_j)$ here serve as input data for solving the inverse problem. In real situations, this function is usually determined experimentally or based on a hydrodynamic study of wells, so it has certain errors. To simulate these errors, the function $z(t)$ is perturbed by random errors [20] as follows:

$$z_\delta^j = z^j + 2\delta \left(\sigma^j - \frac{1}{2} \right), \quad (30)$$

where σ^j is a random function uniformly distributed over the interval $[0, 1]$, δ is the level of error. Graphs $z_\delta(t)$ are shown in Fig. 1.

Problems (6)–(8), (18)–(20) and (21)–(23) at $\lambda_1 = \lambda_1^s$, $\alpha_0 = \alpha_0^s$ are also solved numerically using the finite difference method [30]. We introduce the notation $w11_i^j = w_{11}(t_j, x_i)$, $w21_i^j = w_{21}(t_j, x_i)$, $w12_i^j = w_{12}(t_j, x_i)$, $w22_i^j = w_{22}(t_j, x_i)$.

The residual functional (11) is calculated by the following integral sums

$$J = \sum_{j=0}^{M-1} \tau \left(p1_0^{j+1} - z^{j+1} \right)^2, \quad z^j = z(t_j).$$

Integrals (16) are calculated as follows:

$$a_{11} = \sum_{j=0}^{M-1} \tau \left[w11_0^{j+1} \right]^2 + \alpha, \quad a_{22} = \sum_{j=0}^{M-1} \tau \left[w12_0^{j+1} \right]^2 + \alpha, \quad a_{12} = a_{21} = \sum_{j=0}^{M-1} \tau w11_0^{j+1} w12_0^{j+1},$$

$$b_1 = \sum_{j=0}^{M-1} \tau \left[w11_0^{j+1} \lambda_1^s + w12_0^{j+1} \alpha_0^s - p1_0^{j+1} + z^{j+1} \right] w11_0^{j+1} + \alpha \lambda_1^s,$$

$$b_2 = \sum_{j=0}^{M-1} \tau \left[w11_0^{j+1} \lambda_1^s + w12_0^{j+1} \alpha_0^s - p1_0^{j+1} + z^{j+1} \right] w12_0^{j+1} + \alpha \alpha_0^s.$$

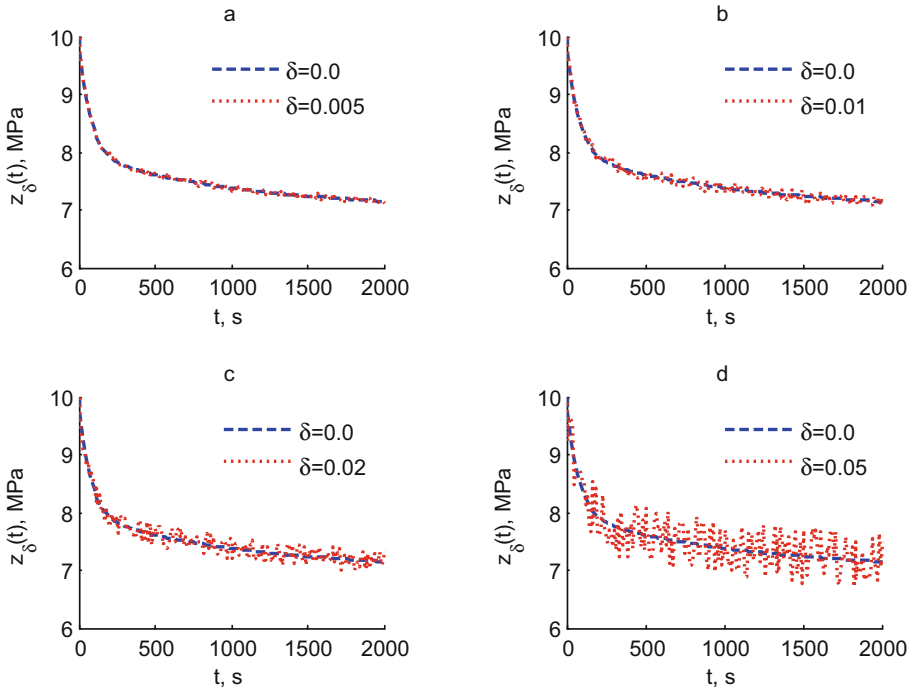


Fig. 1. Perturbed function $z_\delta(t)$

5 Results of Numerical Calculations

The grid divided the coordinate segment $[0, 60]$ into 120 intervals, the time segment $[0, 2000]$ into 4000 intervals. “Measurement z_{δ}^j data” (21) prepared on the basis of this decision at 200 “time” points. The calculation results are shown in Figs. 2, 3, 4, 5, 6 and Table 1.

Figures 2, 3 show the results of calculations to determine the coefficients λ_1, α_0 with accuracy $\varepsilon = \varepsilon_1 = \varepsilon_2 = 1 \cdot 10^{-6}$ at various initial approximations λ_1^0, α_0^0 with unperturbed initial data ($\delta = 0, 0$). The calculations were carried out with the regularization parameter $\alpha = 0$ (Fig. 2, Fig. 3). From Fig. 2 it can be seen that near the equilibrium point for various initial approximations λ_1^0, α_0^0 (up to 1.8 times more or up to 5 times less) coefficients λ_1, α_0 is restored in almost twelve to sixteen iterations. Figure 3 presents the results of calculations for cases where the initial approximations are far from the equilibrium point. From Fig. 3 it can be seen that the iterative process converges in almost ten to twenty iterations.

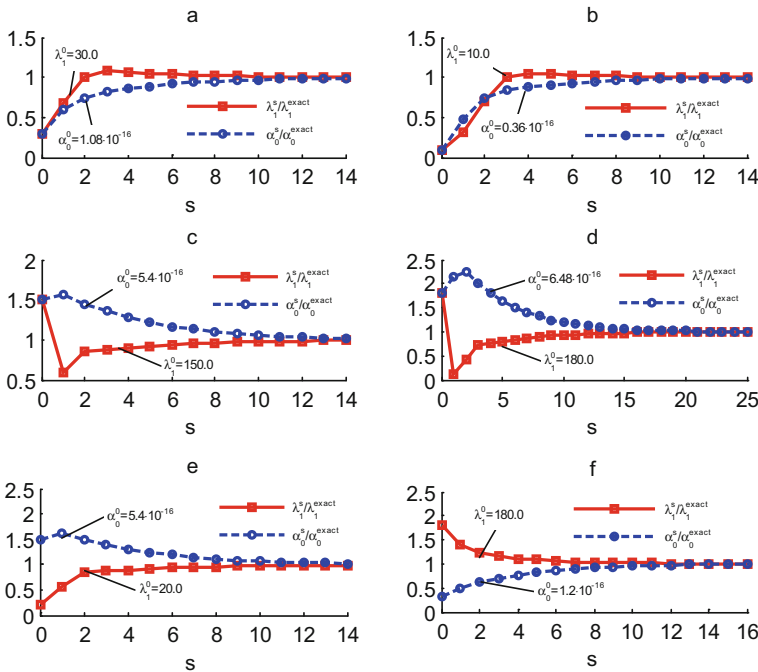


Fig. 2. Recovery of coefficients λ_1 and α_0 for initial approximations near the equilibrium point with unperturbed initial data (at $\delta = 0$), $\lambda_1^{\text{exact}}, \alpha_0^{\text{exact}}$ are the values of the parameters used in preparing the initial data, respectively λ_1 and α_0

From Fig. 4a it can be seen that the initial approximations are twice as large as the exact values of the sought coefficients, the iterative process diverges at the parameter $\alpha = 0$. Therefore, to obtain a satisfactory result, it is necessary to apply the regularization parameter ($\alpha \neq 0$). On Figs. 4 b–d shown the results of calculations for various values of the regularization parameter α . It can be seen here that the maximum number of

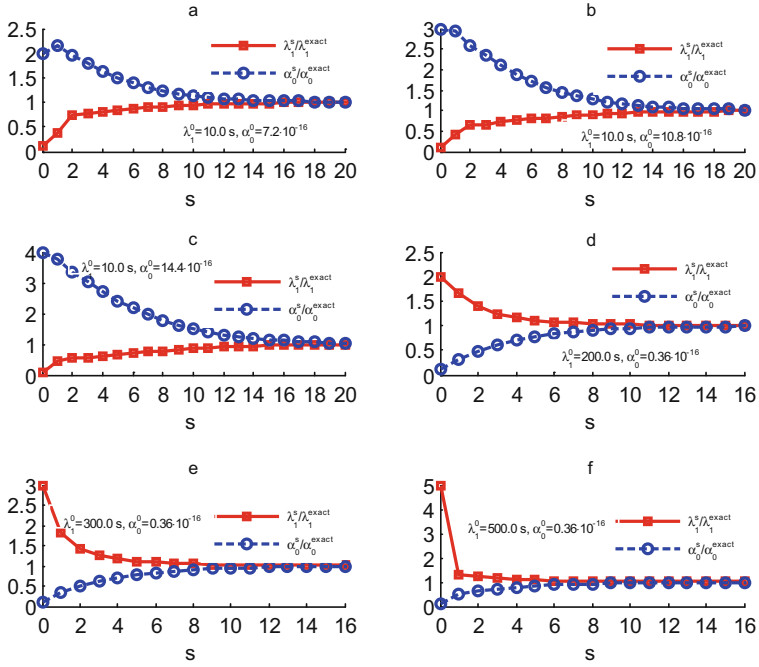


Fig. 3. Recovery of coefficients λ_1 and α_0 for initial approximations far from the equilibrium point with unperturbed initial data (at $\delta = 0$), $\lambda_1^{\text{exact}}, \alpha_0^{\text{exact}}$ – as in Fig. 2.

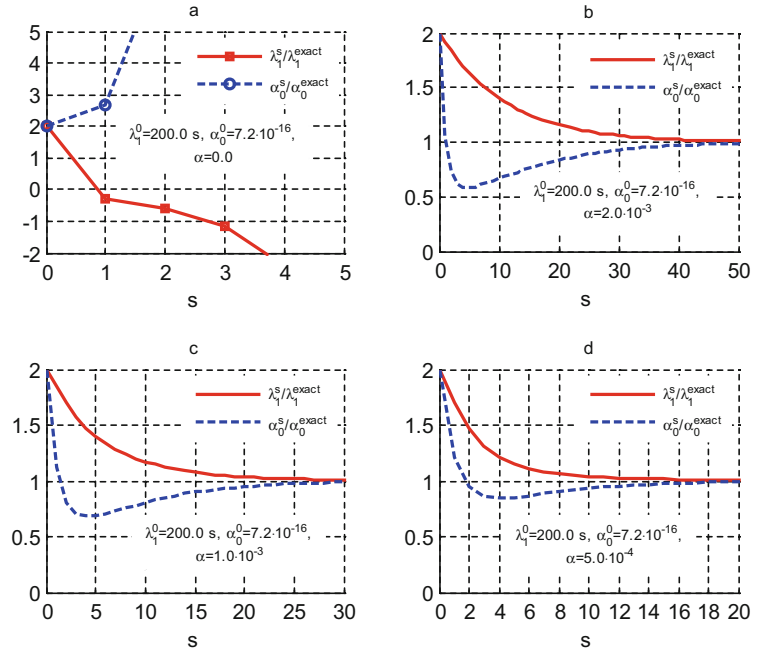


Fig. 4. Recovery of coefficients λ_1 and α_0 with unperturbed initial data (at $\delta = 0$) using regularization, $\lambda_1^{\text{exact}}, \alpha_0^{\text{exact}}$ – as in Fig. 2.

iterations varies depending on the value of the α in the range from twenty to fifty (Fig. 4, b–d).

Also Fig. 5, 6 show the results of calculations for determining the coefficients λ_1 and α_0 for different initial approximations λ_1^0 and α_0^0 depending on the regularization parameter α . As can be seen from Fig. 5, 6, with remote initial approximations λ_1^0 , α_0^0 from the equilibrium point for different values of the regularization parameter, α the coefficient λ_1 and α_0 is restored with sufficient accuracy. Depending on the initial approximations and the regularization parameter α maximum number of iterations varies from fourteen to sixty (Fig. 5, a–d; Fig. 6, a - d). For this series of calculations, the optimal value of the regularization parameter is $\alpha = 2 \cdot 10^{-4}$. With this value of α the iterative process converges faster than other values of the parameter α . The calculation results show that if α changes in the range, $5 \cdot 10^{-5} < \alpha < 2 \cdot 10^{-4}$ number of iterations increase, If the value of the $\alpha \leq 5 \cdot 10^{-5}$ iteration process diverges. Therefore, in calculations with perturbed initial data, this value of the regularization parameter was used α .

Some results of calculations with perturbed initial data are given in Table 1. Numerical calculations are reduced by initial approximations λ_1^0 and α_0^0 , which are quite far from the equilibrium point. The relative recovery errors of the coefficient λ_1 varies from 0.001651% to 7.087416%, and the relative recovery errors α_0 vary from 0.0035% to 23.2870%. As can be seen from the presented results of Table 1, the relative errors in determining λ_1 and α_0 increase with the increase in the error of the initial data. In general, the use of the regularizing procedure makes it possible to obtain acceptable estimates of the desired parameters λ_1 , α_0 with perturbed initial data and with the choice of the initial approximation of the iterative procedure at a considerable distance from the equilibrium point.

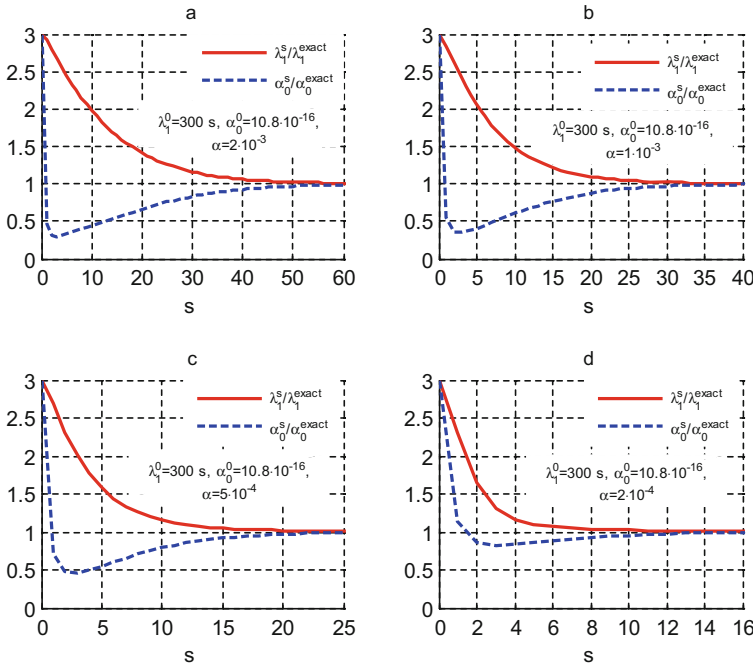


Fig. 5. Recovery of coefficients λ_1 and α_0 with unperturbed initial data (at $\delta = 0$) with change in the regularization parameters, λ_1^{exact} , α_0^{exact} – as in Fig. 2.

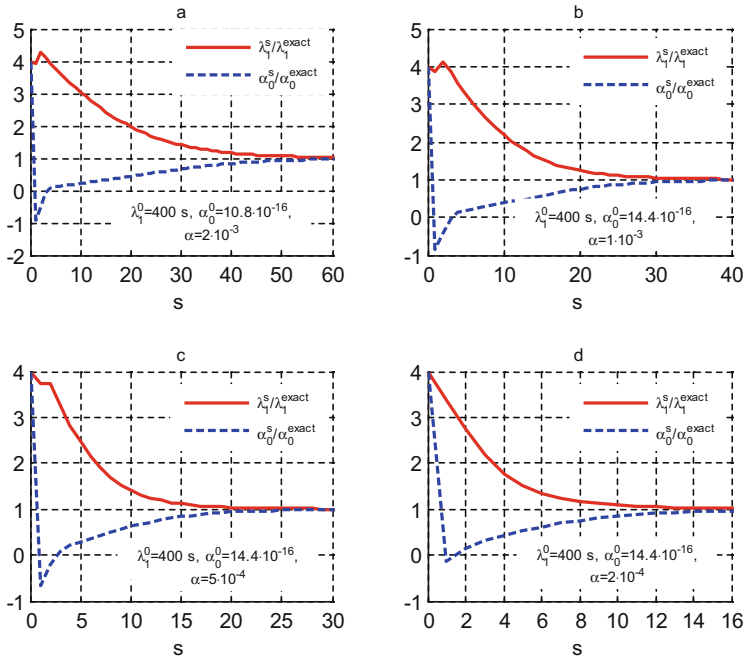


Fig. 6. Recovery of coefficients λ_1 and α_0 in more remote initial approximation from the equilibrium point with unperturbed initial data (at $\delta = 0$) with change in the regularization parameters, $\lambda_1^{\text{exact}}, \alpha_0^{\text{exact}}$ – as in Fig. 2.

Table 1. Recovery of coefficients λ_1 and α_0 with perturbed initial data with regularization

δ	$\alpha = 2 \cdot 10^{-4}$				
	$\lambda_1^0 = 300, 0 \text{ c}$			$\alpha_0^0 = 10.8 \cdot 10^{-16}$	
	s	$\lambda_1^s, \text{ s}$	Relative error $\frac{ \lambda_1 - \lambda_1^{\text{exact}} }{\lambda_1^{\text{exact}}} \cdot 100, \%$	α_0^s	Relative error $\frac{ \alpha_0 - \alpha_0^{\text{exact}} }{\alpha_0^{\text{exact}}} \cdot 100, \%$
0.0	40	100,001651	0.001651	$3,599874 \cdot 10^{-16}$	0.003500
0.005	41	100.105216	0.105216	$3.545263 \cdot 10^{-16}$	1.520472
0.01	47	102.056326	2.056326	$3.407356 \cdot 10^{-16}$	5.351222
0.02	43	106.115035	6.115035	$3.19095 \cdot 10^{-16}$	11.362250
0.05	37	107.087416	7.087416	$2.761668 \cdot 10^{-16}$	23.287000

6 Conclusion

In this paper, we consider the identification of the pressure gradient relaxation coefficient and the overflow coefficient from the solution of the inverse coefficient problem for a simplified model of relaxation filtration of a homogeneous fluid in fractured-porous media.

In order to prepare additional information for solving the inverse problem, the corresponding direct problem was considered with known values of the relaxation time and flow coefficient.

Thus, the “initial data” for solving the inverse problem is prepared. Calculations were also carried out with perturbed initial data, which were prepared by perturbing data with random errors.

The solution of the inverse problem is found by minimizing the residual functional. The minimum of the functional is found from the condition of stationarity with respect to the desired parameter. The results of the calculations show that if the initial approximations in the iterative procedure are close to the equilibrium point, the coefficient of relaxation and flow is restored rather quickly. But, with remote initial approximations, the desired coefficient is determined with a significant deviation from the real values, i.e. the iterative process diverges. In this case, the modified method with regularization gives a satisfactory results. From the numerical experiments, the optimal interval and the optimal value of the regularization parameter were determined. For certain values of the regularization parameter, the relaxation coefficient and the overflow coefficient are restored with sufficient accuracy. As the initial iteration approximation moves away from the equilibrium point, the required number of iterations increases. In the case of perturbed initial data, the use of the method with regularization also gives satisfactory results.

References

1. Shaimuratov, R.V.: Hydrodynamics of an oil fractured reservoir. Moscow: Nedra, p. 223 (1980). [in Russian]
2. Van Golf-Racht, T.D.: Fundamentals of fractured reservoir engineering. Amsterdam: Elsevier Scientific Publishing Company, p. 710 (1982)
3. Barenblatt, G.I., Zheltov, Yu.P.: On the basic equations of filtration of homogeneous fluids in fractured rocks. In: DAN SSSR, vol. 132, No. 3. pp. 545–548 (1960). [in Russian]
4. Barenblatt, G.I., Zheltov, Y.P., Kochina, I.N.: On the main ideas of the theory of filtration of homogeneous fluids in fractured rocks. In: PMM, vol. 24, no. 5. pp. 852–864 (1960). [in Russian]
5. Warren, J.E., Root, P.J.: The behavior of naturally fractured reservoirs. Soc. Petrol. Eng. J., 245–255 (1963)
6. Ametov, I.M., et al.: Production of Heavy And High-Viscosity Oils, p. 205 (1985). [in Russian]
7. Barenblatt, G.I.: Non-equilibrium effects in the filtration of viscoelastic fluids Izv. Acad. Sci. USSR, Mech. Liquid Gas. 5, 76–83 (1973). [in Russian]
8. Alishaev, M.G., Mirzajanzade, A.Kh.: On taking into account the phenomena of delay in the theory of filtration. In: Ed. universities. Oil and gas, no. 6, pp. 71–74 (1975). [in Russian]
9. Molokovich, Y.M., et al.: Relaxation filtering. Kazan: KGU (1980). [in Russian]

10. Molokovich, Y.M.: On the theory of linear filtering with allowance for relaxation effects *Izv. Univ. Maths.* (5), 66–73 (1977). [in Russian]
11. Molokovich, Y.M.: Fundamentals of relaxation filtering. In: *Nauka, M., Problems of Filtration Theory and Mechanics of Enhanced Oil Recovery Processes*, pp. 142–153 (1987). [in Russian]
12. Molokovich, Y.M.: Peculiarities of one-dimensional non-stationary nonlinear filtering with allowance for double relaxation. In: *On Sat “Numerical solution of problems of filtration of a multiphase incompressible fluid”*. Novosibirsk, pp. 152–157 (1977). [in Russian]
13. Iktisanov, V.A., Baigushev, A.V., Garipova, L.I.: Accounting for the finite velocity of propagation of perturbations during unsteady liquid filtration. In: *Neftyanoe khozyaystvo*, no. 2, pp. 78–80 (2010). [in Russian]
14. Iktisanov, V.A., Bilalov, M. Kh., Garipova, L.I.: The use of relaxation models in the interpretation of the results of hydrodynamic studies. In: *Neftpromyslovoe delo*, no. 10, pp.13–16 (2010). [in Russian]
15. Iktisanov, V.A., Bilalov, M.Kh.: Studying the features of relaxation fluid filtration *Oilfield business*, no. 11, pp. 14–18 (2010). [in Russian]
16. Khuzhayorov, B. Kh., Bobokulov, E., Khudoerov, S.C.: Relaxation filtration of homogeneous fluids in with racked-porous media. *J. Eng. Phys. Thermophys.* 74(5), 1073–1082 (2001). <https://doi.org/10.1023/A:1012947326197>
17. Khuzhayorov, B.: Filtration equations for relaxing fluids in fractured-porous media. *Rep. Acad. Sci. UzR.* 7–8, 13–16 (1995). [in Russian]
18. Babe, G.D., Bondarev, E.A., Voevodin, A.F., Kanibolotsky, M.A.: Identification of hydraulic models. *Nauka, Novosibirsk* (1980). [in Russian]
19. Alifanov, O.M., Artyukhin, E.A., Rumyantsev, S.V.: Extreme methods for solving ill-posed problems. *Nauka, Moscow* (1988). [in Russian]
20. Samarsky, A.A., Vabishchevich, P.N.: Numerical methods for solving inverse problems of mathematical physics. *de Gruyter, Berlin* (2007)
21. Alifanov, O.M.: Inverse problems of heat transfer. *Mashinostroenie, Moscow* (1988). [in Russian]
22. Hao, D.N.: Methods for inverse heat conduction problems. *Lang. pub. Inc., Peter* (1998)
23. Beck, J.V., Blackwell, B., Clair, C.R.: Inverse heat conduction: ill-posed problems. *Wiley* (1985)
24. Khairullin, M.Kh., Khisamov, R.S., Shamsiev, M.N., Farkhullin, R.G.: Interpretation of the results of hydrodynamic studies of wells by regularization methods. *Regular and Chaotic Dynamics, Izhevsk Institute for Computer Research* (2006). [in Russian]
25. Khairullin, M.H., Abdullin, A.I, Morozov, P.E, Shamsiev, M.N.: The numerical solution of the inverse problem for the deformable porous fractured reservoir. *Matem. mod.* 20(11), 35–40 (2008). [in Russian]
26. Khairullin, M.H., et al.: Thermohydrodynamic studies of vertical wells with hydraulic fracturing of a reservoir. *High Temp.* 49(5), 769–772 (2011). [in Russian]
27. Khuzhayorov, B., Kholiyarov, E.: Inverse problems of elastoplastic filtration of liquid in a porous medium. *J. Eng. Phys. Thermophys.* 8(3), 517–525 (2007). [in Russian]
28. Khuzhayorov, B., Ali, M., Sulaymonov, F., Kholiyarov, E.: Inverse coefficient problem for mass transfer in two-zone cylindrical porous medium. *AIP Conf. Proc.* 1739, 020028 (2016)
29. Narmuradov, Ch.B., Kholiyarov, E.Ch., Gulomkodiroy, K.A.: Numerical simulation of the inverse problem of relaxation filtration of a homogeneous liquid in a porous medium. *Probl. Comput. Appl. Math.* (2), 12–19 (2017). [in Russian]
30. Samarskii, A.A.: Theory of Difference Schemes. *Nauka, Moscow* (1989). [in Russian]



Solution of the Anomalous Filtration Problem in Two-Dimensional Porous Media

Jamol Makhmudov^(✉), Azizbek Usmonov, and Jakhongir Kuljonov

Samarkand State University, Samarkand, Uzbekistan

j.makhmudov@inbox.ru

Abstract. The problem of anomalous filtration and solute transport in a two-dimensional formulation is posed and numerically solved. It is considered that the medium has a fractal structure. The fractional order piezoconductivity equation based on anomalous Darcy's law, porosity and fluid density equations of state and continuity equation are proposed. The initial and boundary value problem for the system of equations consisting of the balance equation expressed in relation to the concentration of solids in suspension, fractional order piezoconductivity equation and Darcy's anomalous law is solved by finite difference method. The effect of the order of fractional derivative in the anomalous Darcy law on the filtration characteristics of the medium is evaluated. The fields of alteration in concentration, pressure and filtration velocity are determined.

Keywords: anomalous Darcy's law · fractional derivative · solute transport · filtration · porous medium

1 Introduction

When filtering and transporting solute in nonlinear media, as well as in the flow of rheologically complex media, the characteristics often show scale invariance (fractality) both in space and time. This circumstance makes it possible to develop some general methods for modeling complex media and, in some cases, facilitates the description of the processes occurring in them [1].

The paper [2] presents a fractional-differential modification of the model constructed using a time-fractional generalization of Darcy's law, and based on this model, it is proposed to build a hydrodynamic simulator of filtration flows in oil and gas reservoirs.

In [3], a filtration model of flow through a porous medium was proposed. A porous medium is a fractal object, the structure of which is determined by the gap between mating surfaces, which consists of pores and contact areas of wavy and rough mating surfaces. Methods for determining the fractal dimensions of tortuosity and porosity of a medium are given. Dependences of the leakage on the parameters of the porous and compacted medium, as well as the fractal dimension of the tortuosity and porosity of the compacted medium are obtained.

In [4], the derivation of an equation describing the process of fluid filtration in a porous medium with fractal properties is given. It is shown that the structure of the

equation is identical to the structure of the known equations describing the processes of diffusion and random walks. The physical interpretation of the parameters included in the equation and the method of their experimental determination are given. We also note that a number of works are devoted to the problem of modeling and solute transport in media with a fractal structure, as well as to the numerical analysis of equations with fractional derivatives [5–10]. Works [11–13] are devoted to the problems of modeling the processes of pollution and geomigration of groundwater. The processes of radon transfer in media with a fractal structure were analyzed in [14].

This article studies the filtration and solute transport in a two-dimensional medium of a fractal structure.

2 Statement and Numerical Solution of the Problem

Let us assume that the convective and diffusion properties of the medium are different in different directions. However, we consider the medium to be homogeneous in the filtration sense, i.e. the permeability of the medium is the same in different directions. In a more general formulation, of course, it is necessary to consider an inhomogeneous medium in terms of filtration parameters.

Let the study area of the problem consist of $R\{0 \leq x < \infty, 0 \leq y \leq h\}$. Initially, the area R is filled with liquid without particles.

The upper and lower boundaries of the region R are impervious to liquid and particles. The fluid moves in directions x and y in the area R . The process of solute transport, taking into account anomalous effects, can be described by the following equation

$$\frac{\partial c}{\partial t} = D_x \frac{\partial^{\beta_1} c}{\partial x^{\beta_1}} + D_y \frac{\partial^{\beta_2} c}{\partial y^{\beta_2}} - \frac{\partial(v_x c)}{\partial x} - \frac{\partial(v_y c)}{\partial y}, \quad (1)$$

where c is the concentration of suspension, v_x, v_y are the components of the filtration rate, D_x, D_y are the longitudinal and transverse diffusion coefficients, β_1, β_2 are the derivative orders, t is time.

Filtration rate components are defined as [2]

$$v_x = -\frac{k}{\mu} \frac{\partial^{\gamma_1} p}{\partial x^{\gamma_1}}, \quad v_y = -\frac{k}{\mu} \frac{\partial^{\gamma_2} p}{\partial y^{\gamma_2}}. \quad (2)$$

where p is the pressure, μ is the viscosity coefficient of the suspension, k is the permeability coefficient, and γ_1, γ_2 are the orders of the derivative.

The continuity equation for the flow of a compressible fluid through a porous medium can be written as [15]

$$\frac{\partial(\rho m)}{\partial t} + \operatorname{div}(\rho \vec{v}) = 0, \quad (3)$$

where m is the porosity coefficient, ρ is the density of the liquid.

We use the state equations for an elastic fluid and an elastic porous medium [15].

$$\rho = \rho_0(1 + \beta_l(p - p_0)), \quad m = m_0 + \beta_m(p - p_0), \quad (4)$$

where β_l is the coefficient of volumetric compression of the liquid, β_m is the coefficient of elasticity of the medium, ρ_0 is the initial density of the liquid, and p_0 is the initial pressure.

Substituting (2), (4) into (3), we can obtain the piezoconductivity equation with a fractional derivative

$$\frac{\partial p}{\partial t} = \chi \left(\frac{\partial^{\gamma_1+1} p}{\partial x^{\gamma_1+1}} + \frac{\partial^{\gamma_2+1} p}{\partial y^{\gamma_2+1}} \right), \quad (5)$$

where $\chi = \frac{k}{\mu\beta^*}$ - coefficient of piezoconductivity, β^* - is the elastic compressibility coefficient of the medium.

So, we get a system of equations for the solute transport, consisting of the balance Eq. (1), Darcy's law (2) and the piezoconductivity Eq. (3)

$$\begin{aligned} \frac{\partial c}{\partial t} &= D_x \frac{\partial^{\beta_1} c}{\partial x^{\beta_1}} + D_y \frac{\partial^{\beta_2} c}{\partial y^{\beta_2}} - \frac{\partial(v_x c)}{\partial x} - \frac{\partial(v_y c)}{\partial y}, \\ v_x &= -\frac{k}{\mu} \frac{\partial^{\gamma_1} p}{\partial x^{\gamma_1}}, \quad v_y = -\frac{k}{\mu} \frac{\partial^{\gamma_2} p}{\partial y^{\gamma_2}}, \\ \frac{\partial p}{\partial t} &= \chi \left(\frac{\partial^{\gamma_1+1} p}{\partial x^{\gamma_1+1}} + \frac{\partial^{\gamma_2+1} p}{\partial y^{\gamma_2+1}} \right). \end{aligned} \quad (6)$$

The initial and boundary conditions of the problem have the form

$$c(0, x, y) = 0, \quad (7)$$

$$c(t, 0, y) = c_0, \quad c_0 = \text{const}, \quad y = h/2, \quad (8)$$

$$\frac{\partial c}{\partial y}(t, x, 0) = 0, \quad 0 \leq x < \infty, \quad (9)$$

$$\frac{\partial c}{\partial y}(t, x, h) = 0, \quad 0 \leq x < \infty, \quad (10)$$

$$\frac{\partial c}{\partial x}(t, 0, y) = 0, \quad y \neq h/2, \quad 0 \leq y \leq h, \quad (11)$$

$$\frac{\partial c}{\partial x}(t, \infty, y) = 0, \quad 0 \leq y \leq h, \quad (12)$$

$$p(0, x, y) = p_0, \quad p_0 = \text{const}, \quad (13)$$

$$p(t, 0, y) = p_c, \quad p_c > p_0, \quad p_c = \text{const}, \quad y = h/2, \quad (14)$$

$$\frac{\partial p}{\partial y}(t, x, 0) = 0, \quad 0 \leq x < \infty, \quad (15)$$

$$\frac{\partial p}{\partial y}(t, x, h) = 0, \quad 0 \leq x < \infty, \tag{16}$$

$$\frac{\partial p}{\partial x}(t, 0, y) = 0, \quad y \neq h/2, \quad 0 \leq y \leq h, \tag{17}$$

$$\frac{\partial p}{\partial x}(t, \infty, y) = 0, \quad 0 \leq y \leq h. \tag{18}$$

To solve the problem (6)–(18) we use the method of finite differences. To do this, we construct a grid in the region of R in the form

$$\omega_{h_1 h_2 \tau} = \{(t_k, x_i, y_j), t_k = \tau k, x_i = ih_1, y_j = jh_2, k = \overline{0, K}, i = 0, 1, \dots, j = 0, 1, \dots, J^+, \tau = T/K, h_2 = h/J\},$$

where h_1 is the grid step in the direction of x , h_2 is the grid step in the direction of y , τ is the grid step in time, T is the maximum time during which the process is studied, K is the number of grid intervals of t , J , is the number of grid intervals of y in R .

Instead of the functions $c(t, x, y)$, $v(t, x, y)$ and $p(t, x, y)$, we will consider network functions whose values at the nodes (t_k, x_i, y_j) , respectively, will be denoted c_{ij}^k by, v_{ij}^k and p_{ij}^k .

On the grid $\omega_{h_1 h_2 \tau}$, we approximate the first equation of the system (6) as follows [16–20]

$$\begin{aligned} & \frac{c_{ij}^{k+1/2} - c_{ij}^k}{0,5\tau} \\ &= \frac{D_x}{\Gamma(3 - \beta_1)h_1^{\beta_1}} \sum_{l=0}^{i-1} \left(c_{i-(l-1),j}^k - 2c_{i-l,j}^k + c_{i-(l+1),j}^k \right) \left((l+1)^{2-\beta_1} - l^{2-\beta_1} \right) \\ &+ \frac{D_y}{\Gamma(3 - \beta_2)h_2^{\beta_2}} \sum_{l=0}^{j-1} \left(c_{i,j-(l-1)}^k - 2c_{i,j-l}^k + c_{i,j-(l+1)}^k \right) \left((l+1)^{2-\beta_2} - l^{2-\beta_2} \right) \\ &- \frac{(v_x)_{ij}^k c_{ij}^k - (v_x)_{i-1,j}^k c_{i-1,j}^k}{h_1} - \frac{(v_y)_{ij}^k c_{ij}^k - (v_y)_{i,j-1}^k c_{i,j-1}^k}{h_2}, \end{aligned} \tag{19}$$

$$i = \overline{1, I-1}, \quad j = \overline{1, J-1}, \quad k = \overline{0, K-1},$$

$$\begin{aligned} & \frac{c_{ij}^{k+1} - c_{ij}^{k+1/2}}{0,5\tau} \\ &= \frac{D_x}{\Gamma(3 - \beta_1)h_1^{\beta_1}} \sum_{l=0}^{i-1} \left(c_{i-(l-1),j}^{k+1/2} - 2c_{i-l,j}^{k+1/2} + c_{i-(l+1),j}^{k+1/2} \right) \left((l+1)^{2-\beta_1} - l^{2-\beta_1} \right) \\ &+ \frac{D_y}{\Gamma(3 - \beta_2)h_2^{\beta_2}} \sum_{l=0}^{j-1} \left(c_{i,j-(l-1)}^{k+1/2} - 2c_{i,j-l}^{k+1/2} + c_{i,j-(l+1)}^{k+1/2} \right) \left((l+1)^{2-\beta_2} - l^{2-\beta_2} \right) \end{aligned}$$

$$-\frac{(v_x)_{ij}^{k+1/2} c_{ij}^{k+1/2} - (v_x)_{i-1,j}^{k+1/2} c_{i-1,j}^{k+1/2}}{h_1} - \frac{(v_y)_{ij}^{k+1/2} c_{ij}^{k+1/2} - (v_y)_{i,j-1}^{k+1/2} c_{i,j-1}^{k+1/2}}{h_2}, \quad (20)$$

$$i = \overline{1, I-1}, \quad j = \overline{1, J-1}, \quad k = \overline{0, K-1},$$

where $\Gamma(\bullet)$ is gamma function.

For the filtration velocity component, we use the following schemes

$$(v_x)_{ij}^{k+1/2} = -\frac{k p_{i+1,j}^{k+1/2} - \gamma_1 p_{i,j}^{k+1/2}}{\mu \Gamma(2 - \gamma_1) h_1^{\gamma_1}}, \quad i = \overline{0, I-1}, \quad j = \overline{0, J}, \quad k = \overline{0, K-1}, \quad (21)$$

$$(v_y)_{ij}^k = -\frac{k p_{i,j+1}^k - \gamma_2 p_{i,j}^k}{\mu \Gamma(2 - \gamma_2) h_2^{\gamma_2}}, \quad i = \overline{0, I}, \quad j = \overline{0, J-1}, \quad k = \overline{0, K-1}, \quad (22)$$

$$(v_y)_{ij}^{k+1} = -\frac{k p_{i,j+1}^{k+1} - \gamma_2 p_{i,j}^{k+1}}{\mu \Gamma(2 - \gamma_2) h_2^{\gamma_2}}, \quad i = \overline{0, I}, \quad j = \overline{0, J-1}, \quad k = \overline{0, K-1}. \quad (23)$$

The third equation of system (6) is approximated as

$$\begin{aligned} & \frac{p_{ij}^{k+1/2} - p_{ij}^k}{0,5\tau} \\ &= \frac{\chi}{\Gamma(3 - \gamma_1) h_1^{\gamma_1}} \sum_{l=0}^{i-1} \left(p_{i-(l-1),j}^k - 2p_{i-l,j}^k + p_{i-(l+1),j}^k \right) \left((l+1)^{2-\gamma_1} - l^{2-\gamma_1} \right) \\ &+ \frac{\chi}{\Gamma(3 - \gamma_2) h_2^{\gamma_2}} \sum_{l=0}^{j-1} \left(p_{i,j-(l-1)}^k - 2p_{i,j-l}^k + p_{i,j-(l+1)}^k \right) \left((l+1)^{2-\gamma_2} - l^{2-\gamma_2} \right), \end{aligned} \quad (24)$$

$$i = \overline{1, I-1}, \quad j = \overline{1, J-1}, \quad k = \overline{0, K-1},$$

$$\begin{aligned} & \frac{p_{ij}^k - p_{ij}^{k+1/2}}{0,5\tau} \\ &= \frac{\chi}{\Gamma(3 - \gamma_1) h_1^{\gamma_1}} \sum_{l=0}^{i-1} \left(p_{i-(l-1),j}^{k+1/2} - 2p_{i-l,j}^{k+1/2} + p_{i-(l+1),j}^{k+1/2} \right) \left((l+1)^{2-\gamma_1} - l^{2-\gamma_1} \right) \\ &+ \frac{\chi}{\Gamma(3 - \gamma_2) h_2^{\gamma_2}} \sum_{l=0}^{j-1} \left(p_{i,j-(l-1)}^{k+1/2} - 2p_{i,j-l}^{k+1/2} + p_{i,j-(l+1)}^{k+1/2} \right) \left((l+1)^{2-\gamma_2} - l^{2-\gamma_2} \right), \end{aligned} \quad (25)$$

$$i = \overline{1, I-1}, \quad j = \overline{1, J^+ - 1}, \quad k = \overline{0, K-1}.$$

The initial and boundary conditions are approximated as

$$c_{i,j}^k = 0, \quad i = \overline{0, I}, \quad j = \overline{0, J}, \quad k = 0, \quad (26)$$

$$c_{i,j}^k = c_0, \quad i = 0, \quad j = J/2, \quad k = \overline{0, \overline{K}}, \quad (27)$$

$$\frac{c_{i,j+1}^k - c_{i,j}^k}{h_2} = 0, \quad i = \overline{0, I}, \quad j = 0, \quad k = \overline{0, \overline{K}}, \quad (28)$$

$$\frac{c_{i,j}^k - c_{i,j-1}^k}{h_2} = 0, \quad i = \overline{0, I}, \quad j = J, \quad k = \overline{0, \overline{K}}, \quad (29)$$

$$\frac{c_{i+1,j}^k - c_{i,j}^k}{h_1} = 0, \quad i = 0, \quad j \neq J/2, \quad j = \overline{0, \overline{J}}, \quad k = \overline{0, \overline{K}}, \quad (30)$$

$$\frac{c_{i,j}^k - c_{i-1,j}^k}{h_1} = 0, \quad i = I, \quad j = \overline{0, \overline{J}}, \quad k = \overline{0, \overline{K}}, \quad (31)$$

$$p_{i,j}^k = p_0 = \text{const}, \quad i = \overline{0, I}, \quad j = \overline{0, \overline{J}}, \quad k = 0 \quad (32)$$

$$p_{i,j}^k = p_c, \quad i = 0, \quad j = J/2, \quad k = \overline{0, \overline{K}}, \quad (33)$$

$$\frac{p_{i,j+1}^k - p_{i,j}^k}{h_2} = 0, \quad i = \overline{0, I}, \quad j = 0, \quad k = \overline{0, \overline{K}}, \quad (34)$$

$$\frac{p_{i,j}^k - p_{i,j-1}^k}{h_2} = 0, \quad i = \overline{0, I}, \quad j = J, \quad k = \overline{0, \overline{K}}, \quad (35)$$

$$\frac{p_{i+1,j}^k - p_{i,j}^k}{h_1} = 0, \quad i = 0, \quad j \neq J/2, \quad j = \overline{0, \overline{J}}, \quad k = \overline{0, \overline{K}}, \quad (36)$$

$$\frac{p_{i,j}^k - p_{i-1,j}^k}{h_1} = 0, \quad i = I, \quad j = \overline{0, \overline{J}}, \quad k = \overline{0, \overline{K}}, \quad (37)$$

where I is a sufficiently large number for which the equation approximately holds $c_{Ij}^k = 0$.

The calculation sequence is as follows: first p_{ij}^k , are determined from the difference scheme (24) on the $(k + 1/2)$ -layer, then from (21), (23) the components of the filtration rate are calculated, then it c_{ij}^k is determined on the $(k + 1/2)$ -layer from the difference Eqs. (19). Then, on the $(k + 1)$ -layer, it is determined p_{ij}^k from the difference scheme (25), from (22) the filtration velocity components are calculated, after that it c_{ij}^k is determined on the $(k + 1)$ -layer from the difference Eqs. (20).

For Eqs. (17), (18) we introduce the following notations

$$E_1^k = \frac{(v_x)_{ij}^k c_{ij}^k - (v_x)_{i-1,j}^k c_{i-1,j}^k}{h_1}, \quad E_2^k = \frac{(v_y)_{ij}^k c_{ij}^k - (v_y)_{i,j-1}^k c_{i,j-1}^k}{h_2},$$

$$G_1 = \frac{0, 5\tau D_x}{\Gamma(3 - \beta_1)h_1^{\beta_1}}, \quad G_2 = \frac{0, 5\tau D_y}{\Gamma(3 - \beta_2)h_2^{\beta_2}},$$

$$S_1^k = \sum_{l=0}^{i-1} \left(c_{i-(l-1),j}^k - 2c_{i-l,j}^k + c_{i-(l+1),j}^k \right) \left((l+1)^{2-\beta_1} - l^{2-\beta_1} \right),$$

$$S_2^k = \sum_{l=0}^{j-1} \left(c_{i,j-(l-1)}^k - 2c_{i,j-l}^k + c_{i,j-(l+1)}^k \right) \left((l+1)^{2-\beta_2} - l^{2-\beta_2} \right).$$

Taking into account these notations, Eqs. (19) and (20), respectively, have the form

$$c_{ij}^{k+1/2} = c_{ij}^k + G_1 S_1 + G_2 S_2 - 0, 5\tau \left(E_1^k + E_2^k \right), \quad (38)$$

$$c_{ij}^{k+1} = c_{ij}^{k+1/2} + G_1 S_1 + G_2 S_2 - 0, 5\tau \left(E_1^{k+1/2} + E_2^{k+1/2} \right). \quad (39)$$

The concentration field is determined step by step from (38), (39).

3 Results and Discussion

The following values of the initial parameters were used in the calculations: $k = 10^{-13} m^{1+\gamma_1}$, $\mu = 10^{-3} Pa \cdot s$, $\beta^* = 3 \cdot 10^{-8} Pa^{-1}$, $p_c = 5 \cdot 10^5 Pa$, $p_0 = 10^5 Pa$, $c_0 = 0, 01$ and $D_x = 5 \cdot 10^{-5} m^{\beta_1} / s$, $D_y = 10^{-5} m^{\beta_2} / s$.

Some of the results are shown in Fig. 1–4. In Fig. 1–2 results show, a decrease in the values of γ_1 and γ_2 from 1 leads to an increase in the concentration field and a wider spread. On Fig. 3–4 shows the changes in the pressure and velocity fields at different values of γ_1 and γ_2 . Obtained results show that as γ_1 and γ_2 decreases from 1, a wider distributions of the pressure and velocity fields are observe.

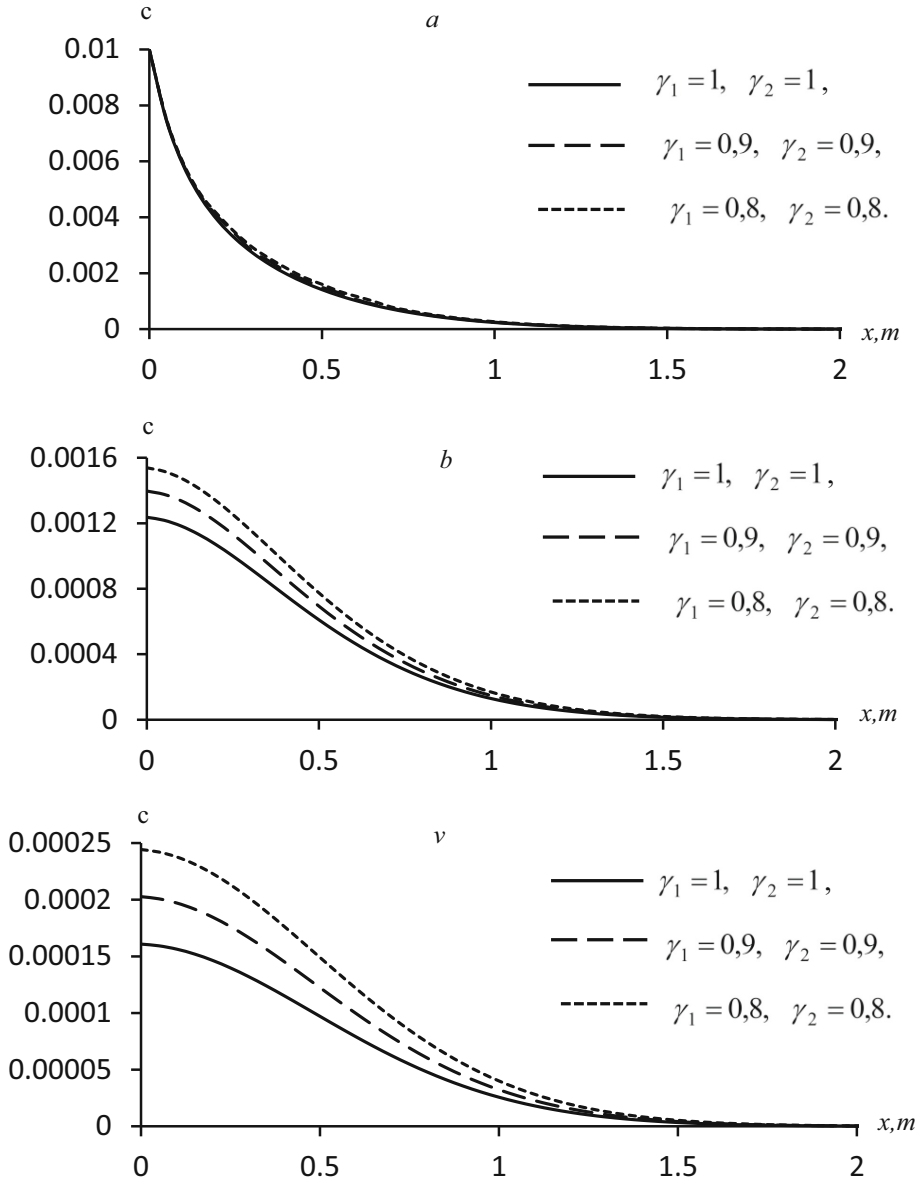


Fig. 1. Distribution of concentration in sections $y = 0,75$ m (a), $1,0$ m (b), $1,25$ m (v) at $k = 10^{-13} m^{1+\gamma_1}$, $\mu = 10^{-3} Pa \cdot s$, $\beta^* = 3 \cdot 10^{-8} Pa^{-1}$, $p_c = 5 \cdot 10^5 Pa$, $p_0 = 10^5 Pa$, $c_0 = 0,01$ and $D_x = 5 \cdot 10^{-5} m^{\beta_1} / s$, $D_y = 10^{-5} m^{\beta_2} / s$, $\beta_1 = 2$, $\beta_2 = 2$, $t = 3600$ s.

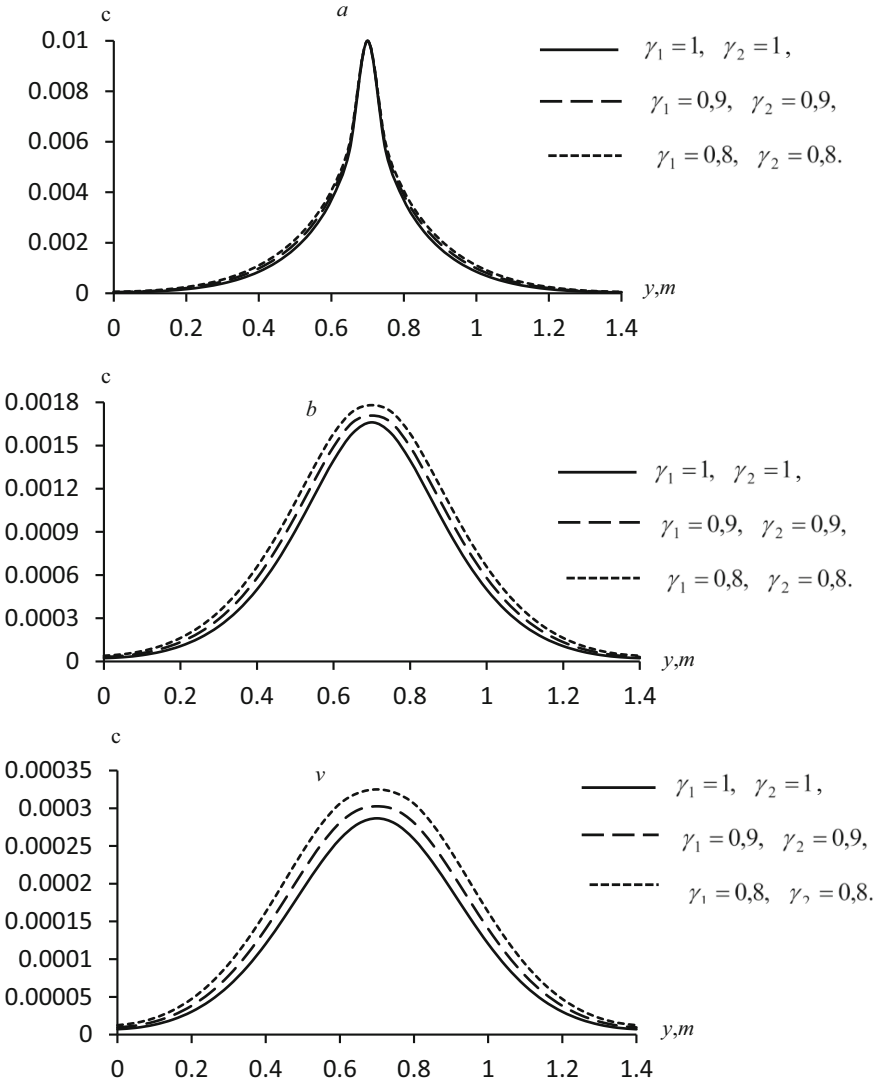


Fig. 2. Concentration distribution in sections $x = 0(a), 0,5 m (b), 1,0 m (v)$ at various, $k = 10^{-13} m^{1+\gamma_1} \mu = 10^{-3} Pa \cdot s, \beta^* = 3 \cdot 10^{-8} Pa^{-1}, p_c = 5 \cdot 10^5 Pa, p_0 = 10^5 Pa, c_0 = 0,01$ and $D_x = 5 \cdot 10^{-5} m^{\beta_1} / s, D_y = 10^{-5} m^{\beta_2} / s, \beta_1 = 2, \beta_2 = 2, t = 3600 s$.

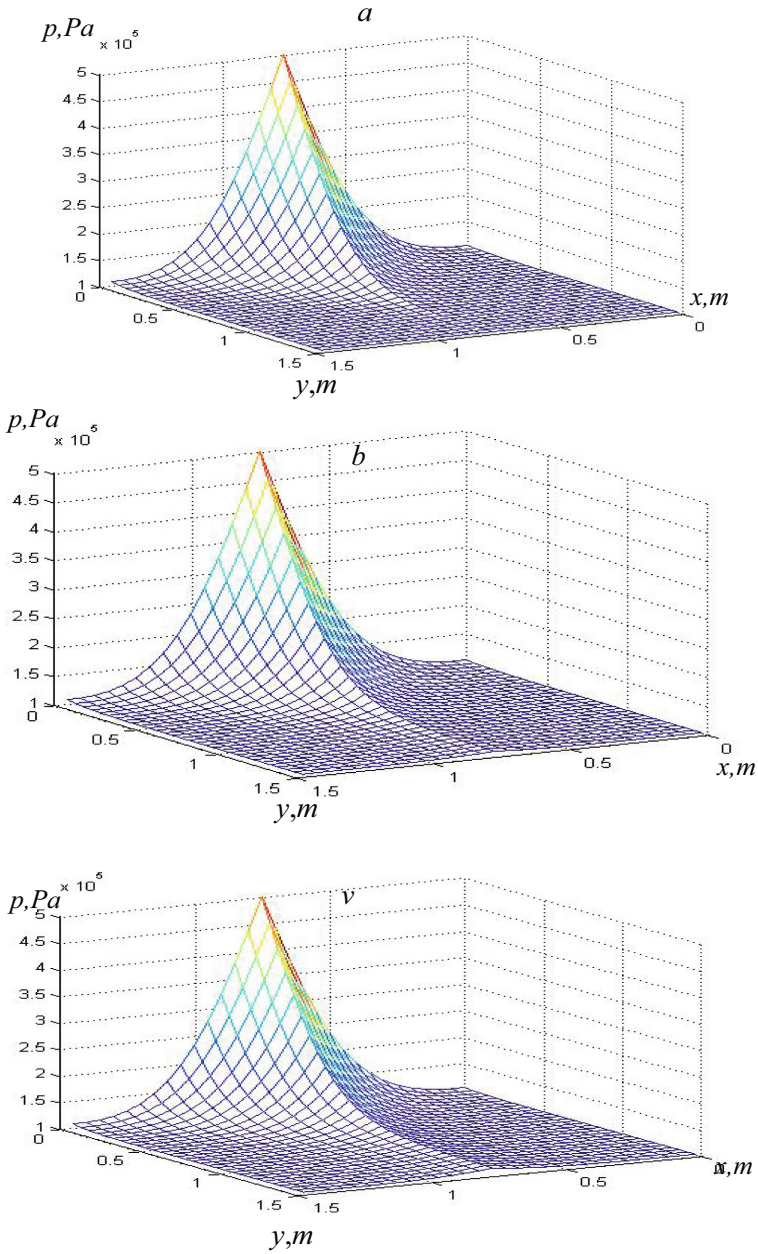


Fig. 3. Surfaces p at $k = 10^{-13} m^{1+\gamma_1}$, $\mu = 10^{-3} Pa \cdot s$, $\beta^* = 3 \cdot 10^{-8} Pa^{-1}$, $p_c = 5 \cdot 10^5 Pa$, $p_0 = 10^5 Pa$, $c_0 = 0,01$ and $D_x = 5 \cdot 10^{-5} m^{\beta_1} / s$, $D_y = 10^{-5} m^{\beta_2} / s$, $\beta_1 = 2$, $\beta_2 = 2$, $t = 1000 s$, $\gamma_1 = 1$, $\gamma_2 = 1$ (a); $\gamma_1 = 0,9$, $\gamma_2 = 0,9$ (b); $\gamma_1 = 0,8$, $\gamma_2 = 0,8$ (v).

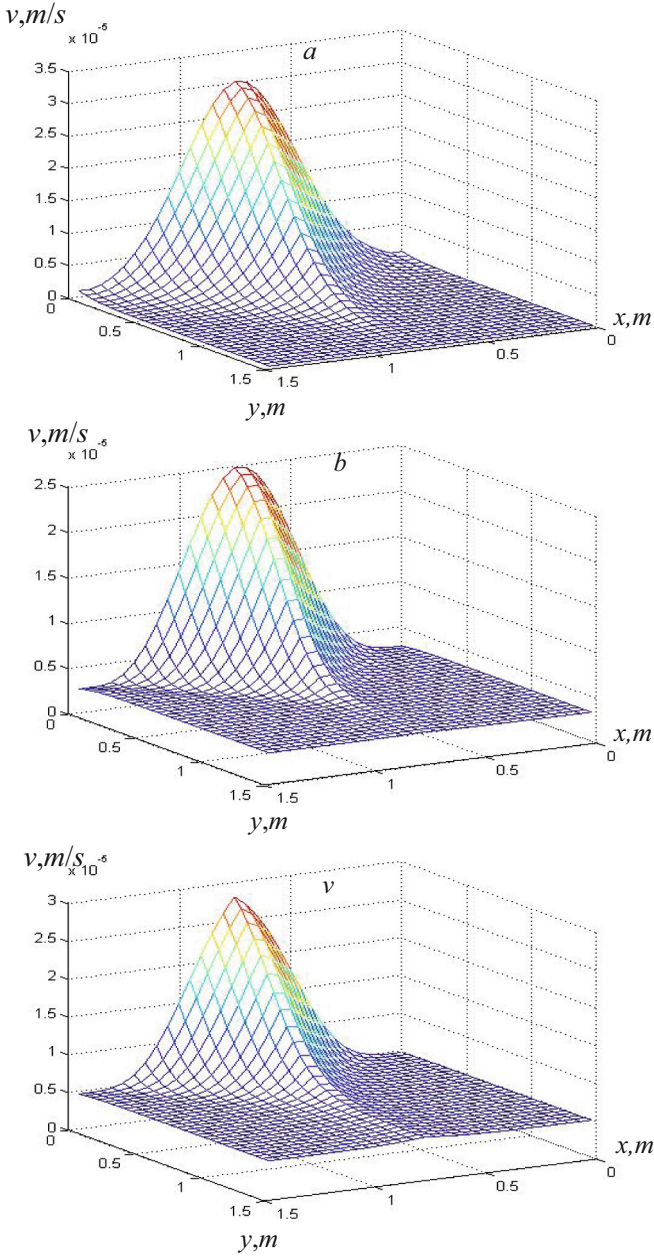


Fig. 4. Surfaces v at $k = 10^{-13} m^{1+\gamma_1}$, $\mu = 10^{-3} Pa \cdot s$, $\beta^* = 3 \cdot 10^{-8} Pa^{-1}$, $p_c = 5 \cdot 10^5 Pa$, $p_0 = 10^5 Pa$, $c_0 = 0, 01$ and $D_x = 5 \cdot 10^{-5} m^{\beta_1} / s$, $D_y = 10^{-5} m^{\beta_2} / s$ $\beta_1 = 2$, $\beta_2 = 2$, $t = 1000 s$, $\gamma_1 = 1$, $\gamma_2 = 1$ (a); $\gamma_1 = 0, 9$, $\gamma_2 = 0, 9$ (b); $\gamma_1 = 0, 8$, $\gamma_2 = 0, 8$ (v).

4 Conclusion

The problem of filtration and transport of solute in a two-dimensional porous medium with a fractal structure is considered. The transport of a solute in such media is described by an equation with fractional derivatives with respect to the spatial coordinates. The calculation results show that a decrease in the order of the derivative in the filtration equation leads to an increase in the pressure field and filtration velocity. Based on the numerical solution of the equation under the appropriate initial and boundary conditions, it is shown that a decrease in the order of the fractional derivative with respect to the spatial coordinate from 1 leads to a greater distribution of the total mass. At the same time, with a decrease in the order of the derivative in the anomalous filtration equation and in the diffusion term of the solute transfer equation, a wider distribution of concentration profiles was observed.

References

1. Khasanov, M.M., Bulgakova, G.T.: Nonlinear and nonequilibrium effects in rheologically complex media. - Moscow-Izhevsk: Institute for Computer Research (2003). 288 p. (in Russian)
2. Belevtsov N.S.: On one fractional-differential modification of the non-volatile oil model. *Mathematics and mathematical modeling* (06), 13–27 (2020). 10.24108/mathm.0620.0000228. (in Russian)
3. Izmerov M.A., Tikhomirov V.P.: Filtration model of flow through a fractal porous medium. *Fundamental Appl. Prob. Eng. Technol.* (3), 7–14 (2014). (in Russian)
4. Bagmanov, V.Kh., Baikov, V.A., Latypov, A.R., Vasiliev, I.B.: Method of interpretation and determination of the parameters of the filtration equation in a porous medium with fractal properties. *Bulletin of the Ufa State Aviation University* (2006). pp.146–149. (in Russian)
5. Bazaev, A.K.: Local-one-dimensional scheme for the diffusion equation of fractional order with a fractional derivative in lower terms with boundary conditions of the first kind. *Vladikavkaz. math. magazine* **16**(2), 3–13 (2014). (in Russian)
6. Bazaev, A.K., Tsopanov, I.D.: Locally one-dimensional difference schemes for a fractional-order diffusion equation with a fractional derivative in lower terms. *Sibirsk. electron. math. Izv.* **12**, 80–91 (2015). (in Russian)
7. Bazaev, A.K., Shkhanukov-Lafishev, M.: Local-one-dimensional scheme for the fractional-order diffusion equation with boundary conditions of the third kind. *Zh. Vychisl. math. and mat. physical* **50**(7), 1200–1208 (2010). (in Russian)
8. Beybalaev, V.D.: Mathematical model of heat transfer in media with a fractal structure. *Matem. modeling* **21**(5), 55–62 (2009). (in Russian)
9. Beybalaev, V.D., Yakubov, A.Z.: Analysis of the difference scheme of an analogue of the wave equation with a fractional differentiation operator. *Vestn. Sam. State. Tech. University Ser. Phys.-Math. sci.* 1(34). 125–133 (2014). (in Russian)
10. Beibalaev, V.D., Shabanova, M.R.: Numerical method for solving the initial-boundary problem for a two-dimensional heat equation with fractional derivatives. *Vestn. Myself. state tech. university Ser. Phys.-Math. Sci.* **5**(21), 244–251 (2010). (in Russian)
11. Afonin, A.A.: Linear two-dimensional models of geofiltration in porous media with a fractal structure // *Izvestiya SFU. Technical science. Section II. Mathematical modeling of ecosystems*, pp.150–154. (in Russian)

12. Bulavatsky, V.M.: Fractional-differential mathematical models of the dynamics of non-equilibrium geomigration processes and problems with non-local boundary conditions. National Academy of Sciences of Ukraine. *Comput. Sci. Cybernet.* (12), 31–40 (2012). (in Russian)
13. Serbina, L.I., Vendina, A.A.: An asymptotic method for solving a fractional equation for the migration of groundwater pollution. *SamGU. Nat. Sci. Ser.* **5**(86), 104–108 (2011). (in Russian)
14. Parovik, R.I., Shevtsov, B.M.: Processes of radon transfer in media with a fractal structure. *Math. Mod.* **21**(8), 30–36 (2009). (in Russian)
15. Barenblatt, G.I., Entov, V.M., Ryzhik, V.M.: Theory of non-stationary filtration of liquid and gas. *Nedra* (1972). – 288 p. (in Russian)
16. Khuzhayorov, B.K., Djiyanov, T.O., Yuldashev, T.R.: Anomalous nonisothermal transfer of a substance in an inhomogeneous porous medium. *J. Eng. Phys. Thermophys.* **92**, 104–113 (2019)
17. Khuzhayorov, B.K., Djiyanov, T.O., Eshdavlatov, Z.: Numerical investigation of solute transport in a non-homogeneous porous medium using nonlinear kinetics. *Int. J. Mech. Eng. Robot Res.* **11**(2), 79–85 (2022)
18. Khuzhayorov, B., Usmonov, A., Nik Long, N.M.A., Fayziev, B.: Anomalous solute transport in a cylindrical two-zone medium with fractal structure. *Appl. Sci. (Switzerland)* **10**, 5349 (2020)
19. Shen, Ch., Phanikumar, M.S.: An efficient space-fractional dispersion approximation for stream solute transport modeling. *Adv. Water Resour.* **32**, 1482–1494 (2009)
20. Xia, Y., Wu, J., Zhou, L.: Numerical solutions of time-space fractional advection–dispersion equations. *ICCES* **9**(2), 117–126 (2009)



On Calculating the Hyperbolic Parameter of a Two-Dimensional Lattice of Linear Comparison Solutions

N. N. Dobrovol'skii^{1,2(✉)}, N. M. Dobrovol'skii¹, I. Yu. Rebrova¹,
and E. D. Rebrov¹

¹ Tula State Lev Tolstoy Pedagogical University, 125 Lenin Avenue,
Tula 300026, Russia

nikolai.dobrovolsky@gmail.com

² Tula State University, Tula 92 Lenin Avenue, 300012, Russia

http://www.mathnet.ru/php/person.phtml?&personid=77189&option_lang=eng

Abstract. The paper is devoted to the issues of efficient calculation of important characteristics of parallelepipedal nets with optimal coefficients. On the class E_s^α of periodic functions, the norm of the linear error functional of approximate integration using quadrature formulas with parallelepipedal nets with optimal coefficients is expressed in terms of the hyperbolic zeta function of the corresponding lattice of linear comparison solutions. For the hyperbolic zeta function of lattices of linear comparison solutions, important estimates of N. S. Bakhvalov from above and from below through the hyperbolic lattice parameter are well known. The paper discusses an efficient algorithm for calculating the hyperbolic parameter of a two-dimensional lattice of linear comparison solutions that requires $O(\ln N)$ arithmetic operations. The basis of the algorithms under consideration are Euler brackets, which L. Euler used back in the XVIII century. The term itself was proposed by K. F. Gauss in his famous "Arithmetic studies". It was actively used by P. G. Lejeune Dirichlet. Currently, the term continuant is often used. Such use violates historical justice. Currently, in the two-dimensional case, most algorithmic problems have been solved in the optimal coefficients method, although dense sequences of parallelepipedal nets with increasing sequences of hyperbolic parameter values and limited values of Sharygin constants have not yet been constructed.

Keywords: Euler brackets · continued fractions · lattice · hyperbolic lattice parameter · quadrature formula · parallelepipedal nets

1 Introduction

In 1959, Professor N. M. Korobov in [6] proposed parallelepipedal nets for integrating periodic functions from the class E_s^α :

$$\iint_{G_s} f\left(\left\{\frac{a_1 k}{p}\right\}, \dots, \left\{\frac{a_s k}{p}\right\}\right) dx = \frac{1}{N} \sum_{k=0}^{N-1} f(\mathbf{x}_k) - R_N(f), \quad \lim_{N \rightarrow \infty} R_N(f) = 0.$$

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

A. Alikhanov et al. (Eds.): APAMCS 2022, LNNS 702, pp. 81–86, 2023.

https://doi.org/10.1007/978-3-031-34127-4_8

The class of periodic functions E_s^α consists of periodic functions of variables x_ν ($\nu = 1, 2, \dots, s$)

$$f(\mathbf{x}) = \sum_{m_1, \dots, m_s = -\infty}^{\infty} C(\mathbf{m}) e^{2\pi i(\mathbf{m}, \mathbf{x})},$$

which satisfy the conditions¹

$$\sup_{\mathbf{m} \in Z^s} |C(\mathbf{m})| (\overline{m}_1 \dots \overline{m}_s)^\alpha = \|f(\mathbf{x})\|_{E_s^\alpha} < \infty. \tag{1}$$

The norm of the linear error functional of approximate integration is expressed in terms of the hyperbolic zeta function of the lattice

$$A = A(a_1, \dots, a_s; N)$$

solutions of linear comparison:

$$\|R_N(f)\|_{E_s^\alpha} = \zeta_H(A|\alpha) = \sum'_{m_1, \dots, m_s = -\infty}^{+\infty} \frac{\delta_N(a_1 \cdot m_1 + \dots + a_s \cdot m_s)}{(\overline{m}_1 \dots \overline{m}_s)^\alpha},$$

where the Korobov symbol $\delta_N(m)$ is defined by the equalities

$$\delta_N(m) = \begin{cases} 1 & \text{at } m \equiv 0 \pmod{N}, \\ 0 & \text{at } m \not\equiv 0 \pmod{N}, \end{cases}$$

and $(a_j, N) = 1$ ($j = 1, 2, \dots, s$). Hereafter \sum' means that summation is carried out over all nonzero points.

N. S. Bakhvalov ([2], 1959) proved a convenient estimate (see [2], p.126, theorem 19)

$$\frac{1}{q(A)^\alpha} < \zeta_H(A|\alpha) \leq 4\alpha \left(\frac{3\alpha^2}{\alpha - 1}\right)^s \frac{(\ln q(A) + 1)^{s-1}}{q(A)^\alpha}, \tag{2}$$

where $q(A)$ — is the hyperbolic lattice parameter A of linear comparison solutions

$$a_1 \cdot m_1 + \dots + a_s \cdot m_s \equiv 0 \pmod{N}.$$

The hyperbolic parameter of the lattice A is the maximum value of the parameter T of the hyperbolic cross $K(T)$, given by the equality

$$K(T) = \{(x_1, \dots, x_s) | \overline{x}_1 \dots \overline{x}_s \leq T\},$$

which does not contain nonzero lattice points inside itself.

¹ Here and further for real m we assume $\overline{m} = \max(1, |m|)$. Thus, the value \overline{m} can be called the truncated norm of the number m , which is consistent with the concept of the truncated norm of the vector, which will be discussed further.

It follows from the formula (2) that numerical integration using parallelepipedal nets belongs to the number of unheard algorithms (see [1]). It follows from this that it is necessary to be able to construct sequences of optimal coefficients for which the corresponding hyperbolic lattice parameter is of great importance.

The purpose of this paper is to consider the calculation of the hyperbolic lattice parameter $\Lambda((a_1, a_2), N)$. In particular, a natural question arises about when the equality

$$q(\Lambda((a_1, a_2), N)) = |a_1| \cdot |a_2|?$$

2 Formulas for the Hyperbolic Parameter

In [5] a formula is found for calculating the hyperbolic parameter of a two-dimensional lattice $\Lambda(a, N)$ solutions of linear comparison $x + ay \equiv 0 \pmod{N}$, $(a, N) = 1$, $1 \leq a < N$. According to this formula, equality is valid

$$q(\Lambda(a, N)) = \min_{0 \leq m \leq n-1} [q_{m+2}, \dots, q_n]_{(n-m-1)} \cdot Q_m,$$

where

$$\frac{a}{N} = \{q_0; q_1, \dots, q_n\} = q_0 + \frac{1}{q_1 + \frac{1}{q_2 + \frac{1}{q_3 + \dots + \frac{1}{q_{n-1} + \frac{1}{q_n}}}}}$$

- decomposition of the fraction $\frac{a}{N}$ into a chain fraction, $\frac{P_m}{Q_m}$ ($m = 0, \dots, n$)
- suitable fractions and Euler brackets $[b_1, \dots, b_n]_{(n)}$ of the order of n , defined recursively

$$[]_{(-1)} = 0, \quad []_{(0)} = 1,$$

$$[b_1, \dots, b_n]_{(n)} = b_n [b_1, \dots, b_{n-1}]_{(n-1)} + [b_1, \dots, b_{n-2}]_{(n-2)} \quad (n \geq 1).$$

Denote by $G(N) = [-N_1, N_2]^2$ the square where $N_1 = \lceil \frac{N-1}{2} \rceil$, $N_2 = \lfloor \frac{N}{2} \rfloor$, which contains N^2 integer points (x, y) . It is not difficult to see that $|G(N) \cap \Lambda(a, N)| = N$. For convenience, we put

$$GA^*(a, N) = G(N) \cap \Lambda(a, N) \setminus \{(0, 0)\}.$$

Consider two segments on the coordinate axes:

$$I_1 = \{(x, 0) \mid -N_1 \leq x \leq N_2\}, \quad I_2 = \{(0, y) \mid -N_1 \leq y \leq N_2\}.$$

It is not difficult to see that for any lattice of $\Lambda(a, N)$ solutions of linear comparison, the equalities are valid $I_1 \cap \Lambda(a, N) = I_2 \cap \Lambda(a, N) = \{(0, 0)\}$. If we put $G^*(N) = G(N) \setminus (I_1 \cup I_2)$, then the set $G^*(N)$ contains exactly $(N - 1)^2$ integer points (x, y) . For convenience, we put

$$GA^{**}(a, N) = G^*(N) \cap \Lambda(a, N).$$

It is not difficult to verify that $|GA^{**}(a, N)| = |G^*(N) \cap \Lambda(a, N)| = N - 1$.

Lemma 1. *Let $N = p$ — an odd prime number, then for any natural $1 \leq a < b \leq p - 1$ the equality holds*

$$GA^*(a, p) \cap GA^*(b, p) = \emptyset.$$

Proof. Indeed, if the point

$$(x, y) \in GA^*(a, p) \cap GA^*(b, p),$$

then

$$x + ay \equiv x + by \equiv 0 \pmod{N}.$$

It follows that $(a - b)y \equiv 0 \pmod{N}$, but since $(y, p) = 1$, then $a = b$. The resulting contradiction proves the statement of the lemma.

If the point $(x, y) \in GA^{**}(a, N)$ at $N = p$ — is a prime number, then from the comparison

$$x + ay \equiv 0 \pmod{N}$$

it follows that $a = p \left\{ -\frac{xy^{-1}}{p} \right\}$. Here y^{-1} means an integer y_1 such that

$$y \cdot y_1 \equiv 1 \pmod{p}, \quad 1 \leq y_1 \leq p - 1.$$

Thus, for $N = p$ — a prime number, each lattice can be given either traditionally $\Lambda(a, N)$, or through an arbitrary partial solution (x, y) .

Indeed, if (x_1, y_1) — is another particular solution, then from the comparisons of $x + ay \equiv 0 \pmod{N}$ and $x_1 + ay_1 \equiv 0 \pmod{N}$ it follows that $xy_1 - yx_1 \equiv 0 \pmod{N}$. In this case, we will write $\Lambda((x, -y), N)$ and the equality will be fulfilled $\Lambda((-y, x), N) = \Lambda\left(p \left\{ -\frac{xy^{-1}}{p} \right\}, N\right)$.

3 Relations for the Hyperbolic Parameter

Analyzing the work of [5], we see that in the set

$$GA^{**}(a, N) = G^*(N) \cap \Lambda(a, N)$$

the most important role is played by a subset of $B(a, N)$, which in this paper is called the Bykovsky set. It is established about him that $B(a, N) = B^*(a, N) \cup -B^*(a, N)$ and

$$B^*(a, N) = \{((-1)^m [q_{m+2}, \dots, q_n]_{(n-m-1)}, Q_m) | m = 0, \dots, n-1\}.$$

Let the integer m_0 be defined from the condition that

$$q(\Lambda(a, N)) = [q_{m_0+2}, \dots, q_n]_{(n-m_0-1)} Q_{m_0}$$

and $x = (-1)^m [q_{m_0+2}, \dots, q_n]_{(n-m_0-1)}$, $y = Q_{m_0}$, then $q(A(a, N)) = |x| \cdot y$ and $a = p \left\{ -\frac{xy^{-1}}{p} \right\}$.

Hence the natural problem arises of determining the value of Q_m^{-1} for any $m = 0, \dots, n - 1$. For $m = 0$ this problem is trivial, since $Q_0 = 1$ and $Q_0^{-1} = 1$. For $m = n - 1$ the identity $P_n Q_{n-1} - P_{n-1} Q_n = (-1)^{n-1}$ is well known, from which it follows that $Q_{n-1}^{-1} \equiv P_n (-1)^{n-1} \pmod{N}$. So $Q_{n-1}^{-1} = a$, if n is an odd number and $Q_{n-1}^{-1} = N - a$, if n is an even number.

In general, it is necessary to use an algorithmic approach. Namely, we decompose the fraction $\frac{x}{p}$ to a continued fraction:

$$\frac{x}{p} = \{b_0; b_1, \dots, b_k\} = b_0 + \frac{1}{b_1 + \frac{1}{\dots + \frac{1}{b_{k-1} + \frac{1}{b_k}}}}$$

Calculate its $k - 1$ -th suitable fraction $\frac{S_{k-1}}{T_{k-1}}$ and find

$$x^{-1} = N \cdot \frac{1 - (-1)^{k-1}}{2} + N \left\{ (-1)^{k-1} \frac{S_{k-1}}{N} \right\}.$$

The second article in the same collection is devoted to the implementation of these and other algorithms.

4 Conclusion

As shown in 1963 by I. F. Sharygin [10], for any choice of nets for quadrature formulas on the class E_s^α for the norm $\|R_N(f)\|_{E_s^\alpha}$ the linear error functional is estimated

$$\|R_N(f)\|_{E_s^\alpha} \gg \frac{\ln^{s-1} N}{N^\alpha}.$$

This estimate is achievable on the class of quadrature formulas with algebraic nets, which are used in the method of K. K. Frolov [9]. For parallelepipedal nets, the question of the achievability of Sharygin estimates remains open. Nevertheless, the result of I. F. Sharygin allows us to introduce a convenient scale for comparing the quality of various nets on the class E_s^α . Let's call the Sharygin constant the value C_{Sh} , given by the equality

$$C_{Sh} = \frac{\|R_N(f)\|_{E_s^\alpha} N^\alpha}{\ln^{s-1} N}.$$

In our opinion, for practical purposes of using quadrature formulas with various nets, it is important to have tables with parameters defining these nets, as well as with values of such quantities as Sharygin constants and hyperbolic lattice parameters corresponding to parallelepipedal nets.



Acknowledgement. The work was prepared under the RFBR grant No 19-41-710004_r_a and with the financial support of a grant from the Government of the Tula region under the Agreement DS/294, 16.11.2021.

References

1. Babenko, K.I.: *Osnovy chislennogo analiza* [Fundamentals of numerical analysis]. Nauka, Moscow, Russia (1986)
2. Bakhvalov, N.S.: On approximate computation of multiple integrals. *Vestnik Moskovskogo universiteta*, **4**, 3–18 (1959)
3. Gauss, K.F.: *Proceedings on the theory of numbers*. Translation of B. B. Demyanov, general edition I. M. Vinogradova, comments B. N. Delone. Publishing House of the USSR Academy of Sciences, Moscow (1959). 978 p
4. Dobrovol'skaya, L.P., Dobrovol'skii, M.N., Dobrovol'skii, N.M., Dobrovol'skii, N.N.: The hyperbolic Zeta function of grids and lattices, and calculation of optimal coefficients. *Chebyshevskij sbornik* **13**(4(44)), 4–107 (2012)
5. Kormacheva, A.N., Dobrovol'skii, N.N., Rebrova, I.Y., Dobrovol'skii, N.M.: On the hyperbolic parameter of a two-dimensional lattice of comparisons. *Chebyshevskii sbornik* **22**(4), 168–182 (2021)
6. Korobov, N.M.: The evaluation of multiple integrals by method of optimal coefficients. *Vestnik Moskovskogo universiteta* **4**, 19–25 (1959)
7. Korobov, N.M.: *Teoretiko-chislovye metody v priblizhennom analize* [Number-theoretic methods in approximate analysis]. Fizmat-giz, Moscow, Russia (1963)
8. Lying Dirichlet, P.G.: *Chessel's theory lessons*. M.– L. ONTI USSR (1936)
9. Frolov, K.K.: Upper bounds on the error of quadrature formulas on classes of functions. *Doklady Akademii nauk SSSR* **231**(4), 818–821 (1976)
10. Sharygin, I.F.: Lower bounds for the error of quadrature formulas on function classes. *Zhurnal vychislitel'noj matematiki i matematicheskoy fiziki* **7**(4), 784–802 (1963)
11. Dobrovol'skaya, L.P., Dobrovol'skii, M.N., Dobrovol'skii, N.M., Dobrovol'skii, N.N.: On Hyperbolic Zeta Function of Lattices. *Continuous and Distributed Systems. Solid Mech. Appl.* 211, 23–62 (2014). https://doi.org/10.1007/978-3-319-03146-0_2
12. Euler, L.: *De fractinibus continuis*. *Commun. Acad. Sci. Imper. Petropol.* 9 (1737)
13. Euler, L.: *De relatione inter ternas pluresve quantitates instituenda*. *Petersburger Akademie Notiz. Exhib.* August 14, 1775 // *Commentationes arithmeticae collectae*. V. II. St. Petersburg, pp. 99–104 (1849)



Inverse Problem of Contaminant Transport in Porous Media

B. H. Khuzhayorov¹, E. Ch. Kholiyarov² , and O. Sh. Khaydarov¹  

¹ Samarkand State University, Samarkand, Uzbekistan
khaydarovodiljon1981@gmail.com

² Pedagogical Institute of Termez State University, Termiz, Uzbekistan

Abstract. In this paper, we consider the inverse problem of identifying the retardation coefficient of a parabolic differential equation describing the solute transport in a porous medium, taking into account equilibrium adsorption, advection-diffusion, convection and decomposition (decay) of solute transport. Concentration time curves at given three points of the medium are used as initial data. The retardation coefficient is determined by minimizing the quadratic discrepancy functional. On the basis of a quasi-real experiment, it is shown that the discrepancy parameter can be restored with sufficient accuracy.

Keywords: approximation · functional minimum · regularization · difference schemes · mathematical model · solute transport · porous medium · solute decomposition (decomposition)

1 Introduction

The quality of groundwater, which is a source of drinking water and used in industrial activities, may deteriorate due to the particles of pollutants into aquifers from various sources. In most cases, the characteristics of pollution sources and parameters of its transport are unknown. Many works are devoted to direct problems of contaminant transport in porous media, including the following [1–4].

However, inverse problems of solute transport have not been studied enough. There are only a few works on the estimation of some parameters of mathematical models based on a priori information about the solution of the direct problem in some parts of the research area.

In particular, some theoretical estimates for determining the parameters of the filtration function are given in [5].

Efficient numerical methods for solving inverse problems of mathematical physics are given in [5–11]. Some coefficient inverse problems of the elastic regime of fluid filtration in porous and fractured-porous media are considered in [12–16], the technique for solving inverse problems in which can be used to solve inverse problems of the solute transfer in porous media. Some inverse coefficient problems of the solute transfer in porous media were solved in the solute transfer in a homogeneous isotropic porous medium [17, 18].

In this paper, we consider the inverse problem of determining the retardation coefficient based on the model of solute transport in homogenous isotropic porous medium [4]. The model takes into account such physical phenomena as convective transport, diffusion (or hydrodynamic dispersion), adsorption and solute transport radioactive decay (or chemical reaction). From the general equation in the case of linear equilibrium adsorption, a solute transport equation containing the retardation coefficient is obtained. Here the inverse problem is posed to determine this coefficient on the basis of a priori information about the solution of the corresponding direct problem at given points of the area under consideration. The problem is solved in a variational formulation. A quasi-real computational experiment was carried out. With perturbed initial data at remote initial approximations, the iterative scheme for determining the solution diverges. For this case, the regularization method was applied and a satisfactory solution was obtained.

2 Statement of the Inverse Problem

The equation for the solute transport in a saturated, homogeneous, isotropic porous medium, taking into account the effects of hydrodynamic dispersion, adsorption, and radioactive decay, can be written as [4]

$$\frac{\partial}{\partial t}(mc + (1 - m)S) = \frac{\partial}{\partial x_i}(mD_{i,j} \frac{\partial c}{\partial x_j} - u_i c) - \lambda(mc + (1 - m)S) - qc^* = 0 \quad (1)$$

where c is mass concentration of a substance in a solution kg/m^3 , s is mass concentration of adsorbed substances in solution kg/m^3 , m is the porosity of the porous medium, $D_{i,j}$ is hydrodynamic dispersion tensor, v_i is filtration rate in i th direction, λ is radioactive decay factor, s^{-1} , q is volumetric flow rate of fluid injection in the source per unit volume of the medium, (s^{-1}), c^* is the concentration of the substance in the injected fluid.

In the case of linear equilibrium adsorption, the isotherm is taken in the form

$$S = k_d c \quad (2)$$

where k_d is distribution coefficient. In the one-dimensional case from (1), (2)

$$R \frac{\partial c}{\partial t} - D \frac{\partial^2 c}{\partial x^2} + \frac{v}{m} \frac{\partial c}{\partial x} + \lambda R c - \frac{qc^*}{m} = 0, \quad (3)$$

where $R = 1 + (1 - m)/k_d$ is retardation factor; λ is coefficient of radioactive decay of substances.

To solve Eq. (3), we use the following initial and boundary conditions

$$c(0, x) = 0, \quad (4)$$

$$c(t, 0) = c_0, \quad c(t, l) = 0, \quad (5)$$

When solving the inverse problem, it is assumed that the retardation factor R is unknown. It needs to be determined based on some additional information.

We set the inverse problem as follows: Determine the coefficient R from the minimum condition for the following functional

$$J(R) = \sum_{k=1}^n \int_0^T [c(t, x_k) - z_k(t)]^2 dt, \quad (6)$$

where $z_k(t)$ is $c(t, x_k)$ function values of at points x_k , usually obtained by measurements, $k = 1, 2, \dots, n$.

3 Algorithm for Solving the Inverse Problem

The stationarity condition for functional (6) has the form [6]

$$\frac{dJ}{dR} = 2 \sum_{k=1}^n \int_0^T [c(t, x_k) - z_k(t)] w(t, x_k) dt = 0, \quad (7)$$

where $w = \frac{dc}{dR}$ is sensitivity function [6, 7]. Let us expand the function $c(t, x)$ in a neighborhood of some $\overset{s}{R}$ is in a series up to terms of the second order [6]

$$\overset{s+1}{c}(t, x) \approx \overset{s}{c}(t, x) + \left(\overset{s+1}{R} - \overset{s}{R} \right) \overset{s}{w}(t, x) \quad (8)$$

where $\overset{s+1}{c}$ means the value of $c(t, x)$ at $\overset{s+1}{R} = \overset{s}{R}$. We put expansion (8) into (7), we obtain the following relations

$$\sum_{k=1}^n \int_0^T \left[\overset{s}{c}(t, x_k) + \left(\overset{s+1}{R} - \overset{s+1}{R} \right) \overset{s}{w}(t, x_k) - z_k(t) \right] \overset{s}{w}(t, x_k) dt = 0, \quad (9)$$

from which it is easy to determine approximately $\overset{s+1}{R}$

$$\overset{s+1}{R} = \frac{\sum_{k=1}^n \int_0^T \left[\overset{s}{R} \overset{s}{w}(t, x_k) - \overset{s}{c}(t, x_k) + z_k(t) \right] \overset{s}{w}(t, x_k) dt}{\sum_{l=1}^n \int_0^T \left[\overset{s}{w}(t, x_k) \right]^2 dt}. \quad (10)$$

Differentiating problem (3)–(5) with respect to the coefficient R , we obtain the following problem

$$R \frac{\partial w}{\partial t} + \frac{\partial c}{\partial t} - D \frac{\partial^2 w}{\partial x^2} + \frac{v}{m} \frac{\partial w}{\partial x} + \lambda R w + \lambda c - \frac{qc^*}{m} = 0, \quad (11)$$

$$w(0, x) = 0, \quad (12)$$

$$w(t, 0) = 0, \quad w(t, l) = 0. \tag{13}$$

The algorithm for determining the coefficient R can be constructed as follows:

1. We choose the initial approximation R (assuming $s = 0$);
2. Solving problem (3)–(5) from $t = 0$ to we $t = T$ define the function $c(t, x)$. We find the value of the functional (6). We solve the problem (11)–(13) from $t = 0$ to $t = T$ and determine the $w(t, x)$ function;
3. Using relation (10), we calculate the approximations R^{s+1} ;
4. Steps 2, 3 are repeated until the condition is met

$$\left| \frac{J^{s+1} - J^s}{J^s} \right| \leq \varepsilon, \quad \left| \frac{R^{s+1} - R^s}{R^s} \right| \leq \varepsilon_1. \tag{14}$$

where $\varepsilon, \varepsilon_1$ are given small values.

In the framework of a quasi-real experiment [8], we first consider the direct problem (3) - (5) with known $R^{\text{exact}} = 2, 5$. This problem is solved numerically by the finite difference method [19]. In the area $D = \{0 \leq x \leq l, \quad 0 \leq t \leq T\}$ we introduce a uniform mesh grid

$$\Omega_{h\tau} = \{(x_i, t_j), \quad x_i = ih, \quad i = 0, 1, \dots, N, \quad h = l/N, \quad t_j = j\tau, \quad j = 0, 1, \dots, M, \quad \tau = T/M\}.$$

Based on the results of numerical calculations, the mesh grid function is determined $z_k^j = z_k(t_j), \quad j = 0, 1, \dots, M, \quad k = 1, 2, \dots, n$. In addition, when solving the inverse problem, the grid function $z_k(t)$ is perturbed with random errors [8] as follows:

$$z_{k, \delta}^j = z_k^j + 2\delta \left(\sigma^j - \frac{1}{2} \right), \tag{15}$$

where σ^j random function uniformly distributed over an interval $[0, 1]$, δ – error level. For the numerical solution of problem (3)–(5), the following initial values of the parameters were used:

$T = 250$ days, $L = 100$ m, $v_x = 0, 1$ m/day, $c_0 = 0, 1$ kg/m³; $m = 0, 2$ (m³/m³); $D_{xx} = 1, 0$ m²/day, $c^* = 0, 01$; $q = 5, 0 \cdot 10^{-6}$; $N = 201, M = 1001, h = 0, 5m, \tau = 0, 25$ day, $k_1 = 21; k_2 = 41, k_3 = 81; x_1 = 10$ m, $x_2 = 20$ m, $x_3 = 40$ m.

Graphics $z_{k, \delta}(t), k = 1, 2, 3$ are shown in Fig. 1.

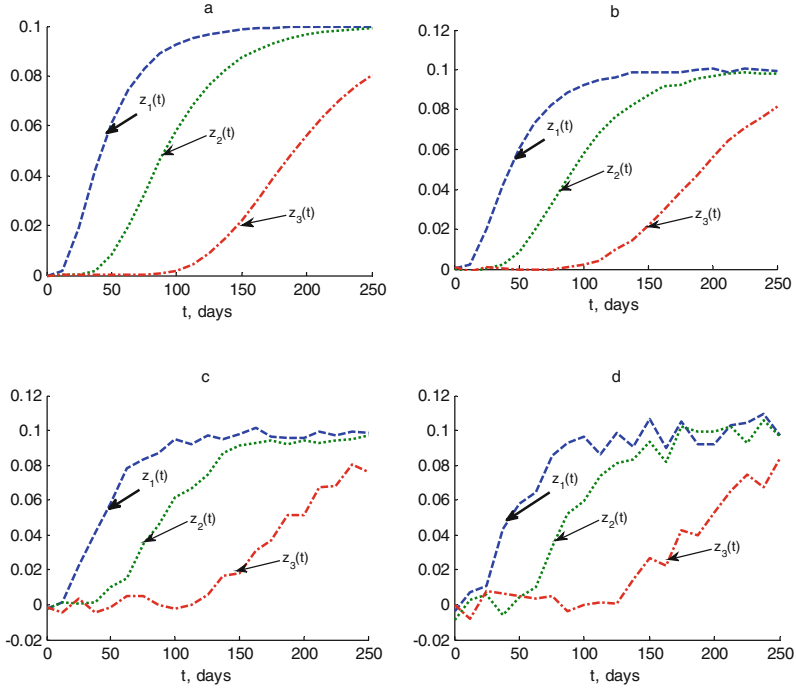


Fig. 1. Function Graph $z_{\delta}(t)$, at $\delta = 0, 0$ (a); $\delta = 0, 01$ (b); $\delta = 0, 05$ (c); (c) $\delta = 0, 10$ (d).

3.1 Difference Problems

Problems (3)–(5) and (11)–(13) are solved numerically with $R = \overset{s}{R}$ using the finite difference method [19]. We approximate problems (3)–(5) using a purely implicit scheme

$$\begin{aligned} \frac{\overset{s}{R} c_i^{j+1} - c_i^j}{\tau} - D \frac{c_{i+1}^{j+1} - 2c_i^{j+1} + c_{i-1}^{j+1}}{h^2} + \frac{v}{m} \frac{c_{i+1}^{j+1} - c_{i-1}^{j+1}}{2h} + \lambda \overset{s}{R} c_i^{j+1} - \frac{qc^*}{m} &= 0, \\ i = 1, 2, \dots, N-1, \quad j = 0, 1, \dots, M-1, & \\ c_i^0 = 0, \quad i = 0, 1, \dots, N, & \\ c_0^{j+1} = c_0, \quad c_N^{j+1} = 0, \quad j = 0, 1, \dots, M-1. & \end{aligned} \quad (16)$$

After approximation, functional (6) takes the form

$$J = \sum_{k=1}^n \sum_{j=0}^{M-1} \tau [c(t_{j+1}, x_k) - z_k(t_{j+1})]^2. \quad (17)$$

Problem (8)–(10) is approximated as follows

$$\begin{aligned} \frac{\overset{s}{R} w_i^{j+1} - w_i^j}{\tau} - \frac{c_i^{j+1} - c_i^j}{\tau} - D \frac{w_{i+1}^{j+1} - 2w_i^{j+1} + w_{i-1}^{j+1}}{h^2} + \frac{v}{m} \frac{w_{i+1}^{j+1} - w_{i-1}^{j+1}}{2h} + & \\ + \lambda \overset{s}{R} w_i^{j+1} + \lambda c_i^{j+1} - \frac{qc^*}{m} &= 0, \quad i = 1, 2, \dots, N-1, \quad j = 0, 1, \dots, M-1, \end{aligned} \quad (18)$$

$$w_i^0 = 0, \quad i = 0, 1, \dots, N,$$

$$w_0^{j+1} = 0, \quad w_N^{j+1} = 0, \quad j = 0, 1, \dots, M - 1.$$

Difference Eqs. (16), (18) can be solved by the Thomas' algorithm. After approximation, relation (10) takes the form

$${}^s R = \frac{\sum_{k=1}^n \sum_{j=0}^{M-1} \tau \left[{}^s R w(t_{j+1}, x_k) - c(t_{j+1}, x_k) + z_k(t_{j+1}) \right] w(t_{j+1}, x_k)}{\sum_{k=1}^n \sum_{j=0}^{M-1} \tau [w(t_{j+1}, x_k)]^2} \quad (19)$$

Difference Eqs. (16) are reduced to the form

$$A c_{i-1}^{j+1} - C c_i^{j+1} + B c_{i+1}^{j+1} = -F_i^j, \quad i = 1, 2, \dots, N - 1, \quad (20)$$

where

$$A = \frac{D\tau}{s R h^2} + \frac{\nu\tau}{2m R h}, \quad B = \frac{D\tau}{R h^2} - \frac{\nu\tau}{2m R h}, \quad C = 1 + \lambda + \frac{2D\tau}{s R h^2}, \quad F_i^j = c_i^j + \frac{q^s c^* \tau}{m R}.$$

The system of Eqs. (20) are solved by the Thomas' algorithm:

$$c_i^{j+1} = \alpha_{i+1} c_{i+1}^{j+1} + \beta_{i+1}, \quad i = N - 1, N - 2, \dots, 1, 0, \quad (21)$$

$$\alpha_{i+1} = \frac{B}{C - A\alpha_i}, \quad \beta_{i+1} = \frac{A\beta_i + F_i^j}{C - A\alpha_i}, \quad i = 1, 2, \dots, N - 1, \quad (22)$$

$$c_0^{j+1} = \alpha_1 c_1^{j+1} + \beta_1, \quad c_0^{j+1} = c_0, \quad \alpha_1 = 0, \quad \beta_1 = c_0. \quad (23)$$

$$c_N^{j+1} = 0. \quad (24)$$

Difference Eqs. (18) takes the form

$$A w_{i-1}^{j+1} - C w_i^{j+1} + B w_{i+1}^{j+1} = -W_i^j, \quad i = 1, 2, \dots, N - 1, \quad (25)$$

where

$$W_i^j = w_i^j - \frac{1}{s} (c_i^{j+1} - c_i^j) + \frac{q c^* \tau}{m R} - \frac{\lambda}{s} c_i^{j+1}.$$

The system of Eqs. (25) are also solved by the Thomas' algorithm

$$w_i^{j+1} = \alpha_{i+1}^{(1)} w_{i+1}^{j+1} + \beta_{i+1}^{(1)}, \quad i = N - 1, N - 2, \dots, 1, 0, \quad (26)$$

$$\alpha_{i+1}^{(1)} = \frac{B}{C - A\alpha_i^{(1)}}, \quad \beta_{i+1}^{(1)} = \frac{A\beta_i^{(1)} + W_i^j}{C - A\alpha_i^{(1)}}, \quad i = 1, 2, \dots, N - 1, \quad (27)$$

$$w_0^{j+1} = \alpha_i^{(1)} w_i^{j+1} + \beta_1^{(1)}, \quad w_0^{j+1} = 0, \quad \alpha_1^{(1)} = 0, \quad \beta_1^{(1)} = 0. \quad (28)$$

$$w_N^{j+1} = 0. \quad (29)$$

4 Numerical Results

The grid divides the $[0, 100]$ coordinate segment into 200 intervals, the $[0, 250]$ time segment into 1000 intervals. The “measurement data” of $z_{k, \delta}^j$, $k = 1, 2, 3$ is prepared based on this results at 20 “time” points. The results of calculations of the recovering of the retardation coefficient with unperturbed initial data for various initial approximations are shown in Fig. 2.

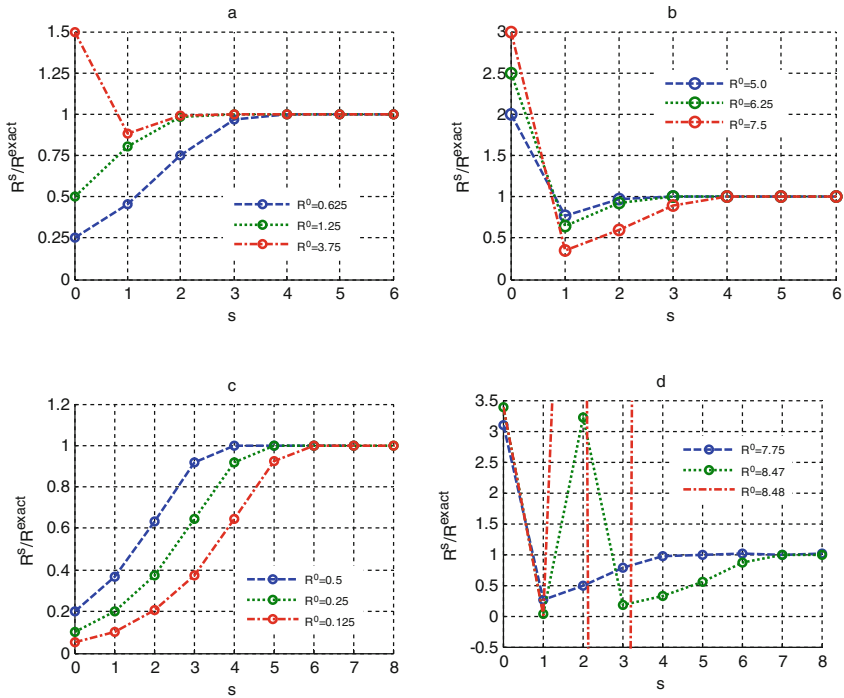


Fig. 2. Recovery of the coefficient R with unperturbed initial data (at $\delta = 0$), R^{exact} is the value of the parameter R used in the preparation of the initial data

In the case when the initial approximation R varies in the range from 0.06 to 8.47, 3–8 iterations are sufficient to restore the parameter R (Fig. 2. a-d).

In the case when the initial approximation is from two to twenty times less than the exact value of the desired coefficient, the coefficient R approaches the exact value monotonically and it requires 3–7 iterations to restore the coefficient R (Fig. 2.a., Fig. 2.c.). In the case when the initial approximation is from one and a half to three times greater than the exact value of the desired coefficient, the approximation will not be monotonic and 3–8 iterations are required (Fig. 2.a., Fig. 2.b.). With the removal of the initial approximation from the equilibrium point, the non-monotonicity of the approximation increases (Fig. 2.d, at $R = 8, 47$). At the value of the initial approximation $R = 8, 48$,

a undipped oscillations will appear (Fig. 2.d). To obtain a satisfactory recovery of R for remote initial approximations of R^0 , we use a modified first-order method [6]. On each iterative layer, instead of functional (4), we use the following functional [6]:

$$J_M \left(R^{(s+1)} \right) = J \left(R^{(s)} \right) + \alpha \left(R^{(s+1)} - R^{(s)} \right)^2,$$

where α is a regularizing parameter. Figure 3 shows the results of calculations of the retardation coefficient R for various initial approximations R^0 depending on the regularization parameter α . It can be seen that when the initial approximation R^0 is remote from the equilibrium point for different values of the regularization parameter α , the coefficient R is restored from (Fig. 3).

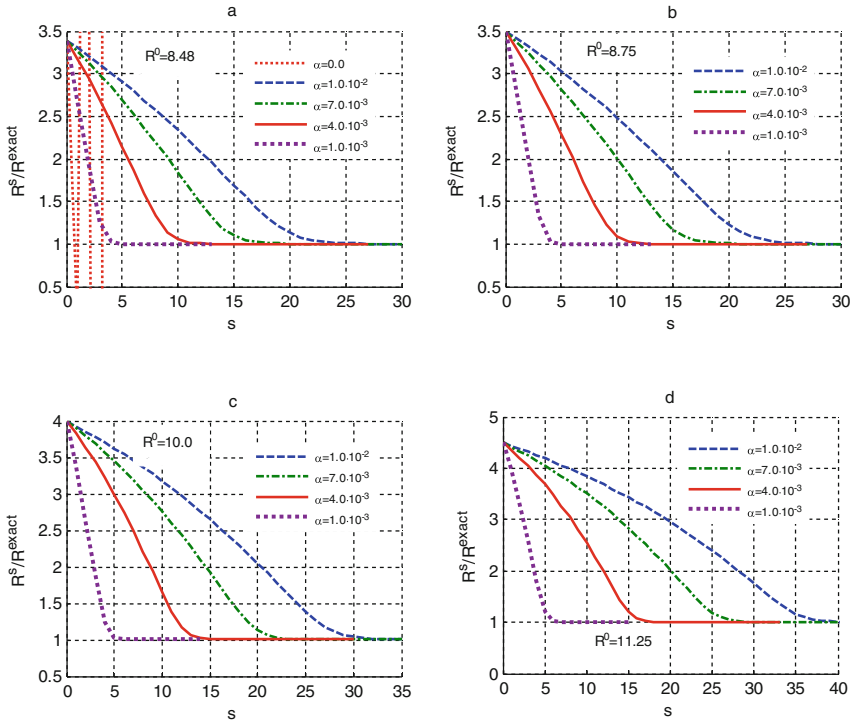


Fig. 3. Recovering of the coefficient R with unperturbed initial data (at $\delta = 0$), R^{exact} as in Fig. 2.

The calculation results show (Fig. 3 a, b, c and d) that as the initial approximation moves away from the equilibrium point, the required number of iterations increases. Depending on the initial approximations, the maximum number of iterations varies from five to forty (Fig. 3, respectively, a - d). For this series of calculations, the value of the regularization parameter $\alpha = 1 \cdot 10^{-3}$ is optimal. Therefore, in calculations with

perturbed initial data, we use this value of the regularization parameter α . Numerical calculations show that the more the initial approximation moves away from the equilibrium point, the greater the number of iterations is required.

The results of calculations with perturbed initial data during the restoration of the retardation coefficient R are given in Table 1. Numerical calculations were carried out with two initial approximations R_0 and R_1 , which are quite far from the equilibrium point from Fig. 1 shows that for an initial approximation of $R = 7,5 R_0$ with unperturbed data ($\delta = 0, 0$) the iterative process converges, and for with the iterative process diverges. Therefore, in the calculations of $R = 10, 0$, the regularization parameter $\alpha = 1 \cdot 10^{-3}$ was used. Relative errors of coefficient R recovery vary from 2323% to 2.1923%. As can be seen from the presented results, the relative error in determining R increases with increasing error in the initial data. In general, the use of the regularizing procedure makes it possible to obtain acceptable estimates of the desired parameter R for perturbed initial data and for choosing the initial approximation of the iterative procedure at a considerable distance from the equilibrium point.

Table 1. Recovery of the coefficient R with perturbed initial data

δ	$R_0 = 7,5$			$R_0 = 10, 0, \alpha = 1 \cdot 10^{-3}$		
	s	R_0	Relative error $\frac{ R_0 - R^{\text{exact}} }{R^{\text{exact}}} \cdot 100, \%$	s	R_0	Relative error $\frac{ R_0 - R^{\text{exact}} }{R^{\text{exact}}} \cdot 100, \%$
0,0	8	2,500000	0,0	not approaching		
0,01	9	2,498738	0,05048	14	2,502068	0,08272
0,05	9	2,488793	0,44828	16	2,514688	0,58752
0,10	9	2,445192	2,19232	14	2,474603	1,01588

5 Conclusion

In this paper, we consider the inverse problem of identifying the retardation coefficient of a parabolic differential equation describing the contaminant transport in a porous medium, taking into account equilibrium adsorption, diffusion, convective transport, radioactive decay. Concentration time curves at given two points of the medium are used as initial data. The retardation coefficient is determined by minimizing the quadratic residual functional. On the basis of a quasi-real experiment, it is shown that the desired parameter can be restored with sufficient accuracy. In this work, the identification of the retardation coefficient and the retardation coefficient from the solution of the inverse coefficient problem for the model equation of the inverse problems for solute transport of a homogeneous fluids in porous media is considered. In order to prepare additional

information for solving the inverse problem, the corresponding direct problem with known values of the retardation coefficient was considered. Thus, the “initial data” for solving the inverse problem is prepared. Calculations were also carried out with perturbed initial data, which were prepared by noisy data with random errors.

The solution of the inverse problem is found by minimizing the residual functional. The minimum of the functional is found from the condition of stationarity with respect to the desired parameter. The calculation results show that if the initial approximations in the iterative procedure are close to the equilibrium point, the retardation coefficient is restored rather quickly. But, with remote initial approximations, the desired coefficient is determined with a significant deviation from the real values, i.e. the iterative process diverges. In this case, the modified method with regularization gives a satisfactory result. From the numerical experiments performed, the optimal interval and the optimal value of the regularization parameter were determined. For certain values of the regularization parameter, the retardation coefficient and the regularization coefficient are restored with sufficient accuracy. As the initial iteration approximation moves away from the equilibrium point, the required number of iterations increases. In the case of perturbed initial data, the use of the method with regularization also gives satisfactory results.

References

1. Salama, A., Van Geel, P.J.: Flow and solute transport in saturated porous media: 1. The continuum hypothesis. *J. Porous Media* **11**, 403–413 (2008)
2. Bear, J., Cheng, A.H.-D.: *Modeling Groundwater Flow and Contaminant Transport*. Springer, Berlin, Germany (2010). <https://doi.org/10.1007/978-1-4020-6682-5>
3. Cussler, E.L.: *Diffusion Mass Transfer in Fluid Systems*, 3rd edn. Cambridge University Press, Cambridge (2009)
4. Lakshminarayana, V., Nayak, T.R.: *Modelling contaminant transport in saturated aquifers* Department of Civil Engineering Indian Institute of Technology, Kanpur, India
5. Alvarez, A.C., Hime, G., Marchesin, D., Bedrikovetsky, P.G.: The inverse problem of determining the filtration function and permeability reduction in flow of water with particles in porous media. *Transp. Porous Media* **70**, 43–62 (2007). <https://doi.org/10.1007/s11242-006-9082-3>
6. Babe, G.D., Bondarev, E.A., Voevodin, A.F., Kanibolotsky, M.A.: *Identification of Hydraulic Models*. Nauka, Novosibirsk (1980)
7. Samarski, A.A., Alifanov, O.M., Artyukhin, E.A., Rumyantsev, S.V.: *Extreme Methods for Solving Ill-Posed Problems*. Nauka, Moscow (1988)
8. Samarskii, A.A., Vabishchevich, P.N.: *Numerical Methods for Solving Inverse Problems of Mathematical Physics*. LKI Publishing House, Moscow (2009)
9. Alifanov, O.M.: *Inverse Problems of Heat Transfer*. Mashinostroenie, Moscow (1988)
10. Hao, D.N.: *Methods for Inverse Heat Conduction Problems*. – Lang. pub. Inc., Peter (1998)
11. Beck, J.V., Blackwell, B., Clair, C.R.: *Inverse Heat Conduction: Ill-Posed Problems*. Wiley (1985)
12. Khairullin, M.H., Abdullin, A.I., Morozov, P.E., Shamsiev, M.N.: The numerical solution of the inverse problem for the deformable porous fractured reservoir. *Matem. Mod.* **20**(11), 35–40 (2008)
13. Khairullin, M.H., et al.: Thermohydrodynamic studies of vertical wells with hydraulic fracturing of a reservoir. *High Temp.* **49**(5), 769–772 (2011)

14. Khuzhayorov, B., Kholiyarov, E.: Inverse problems of elastoplastic filtration of liquid in a porous medium. *J. Eng. Phys. Thermophys.* **8**(3), 517–525 (2007)
15. Khuzhayorov, B.Kh., Kholiyarov, E.Ch.: Determination of the flow coefficient and permeability during filtration of a homogeneous liquid in fractured-porous media. *Prob. Comput. Appl. Math.* (1(38)), 66–76 (2022)
16. Narmuradov, Ch.B., Kholiyarov, E.Ch., Gulomkodirov, K.A.: Numerical modeling of the inverse problem of relaxation filtration of a homogeneous fluid in a porous medium. *Prob. Comput. Appl. Math.* (2), 12–19 (2017)
17. Khuzhayorov, B., Ali, Md.F., Sulaymonov, F., Kholiyarov, E.: Inverse coefficient problem for mass transfer in two-zone cylindrical porous medium. *AIP Conf. Proc.* **1739**, 020028 (2016)
18. Begmatov, T.I., Kholiyarov, E.Ch., Fayziev, B.M.: Identification of the kinetic coefficient in the model of suspension filtration in a porous medium. *Prob. Comput. Appl. Mathem.* (1(38)), 9–17 (2022)
19. Samarsky, A.A.: *Theory of Difference Schemes*. Nauka, Moscow (1989)



Numerical Solution of Anomalous Solute Transport in a Two-Zone Fractal Porous Medium

Bakhtiyor Khuzhayorov^(✉), Azizbek Usmonov, and Fakhridin Kholliiev

Samarkand State University, Samarkand, Uzbekistan

b.khuzhayorov@mail.ru

Abstract. The process of anomalous solute transport in a porous medium is modeled by differential equations with a fractional derivative. The problem of the solute transport in a two-zone porous medium consisting of macropores and micropores. The profiles of changes in the concentrations of suspended particles in the macropore and micropore were determined. The influence of the order of the derivative with respect to the space and time coordinates is estimated, i.e. fractal dimension of the medium, on the characteristics of the solute transport in both zones.

Keywords: filtration · fractal structure · fractional derivative · solute transport · porous medium

1 Introduction

The problem of solute transport in porous media occurs in many technical processes. Mathematical models are widely used in the design and analysis of the solute transport in porous media [1, 2]. If the porous medium is nonhomogeneous (at the micro- and macrolevels), the process of solute transport can be anomalous, i.e., the solute transport does not obey Fick's law [3–5]. In most cases, a nonhomogeneous medium has a fractal dimension, and Fick's law is also written as a fractional derivative depending on the fractal dimension of the medium. This confirms that the process of solute transport proceeds anomalously. Solute transport equations based on Fick's law with fractional derivatives have the form of differential equations with fractional derivatives [7–10]. Such equations are not yet well understood. The equations have been studied only in some simple cases. Solving the equations of fractional derivatives by the finite difference method also has its own difficulties [11, 12].

The complex trajectories of liquid and solute particles in aggregates, fracture and porous blocks cause anomalous transport, and the conventional convective transport equation cannot adequately describe anomalous transport. Therefore, such media can be called fractal media. Recently, the interest of researchers in the anomalous solute transport has increased significantly. First of all, this is dictated by the relevance of the problem in terms of applications in various industries and technology. On the other hand, there are many unresolved theoretical issues, in particular, the question of the influence

of the anomalous nature of the transport on the hydrodynamic parameters has not been fully clarified [13, 14].

The rocks of many oil fields tend to be heterogeneous on both microscopic and macroscopic scales. A typical example of nonhomogeneous porous media are aggregated and fractured-porous media [15–17].

The equations for the solute transport in fractals were first proposed in [18]. In fractured-porous media, the transport equations were analyzed in [20, 21]. It is shown that the order of the fractional derivative in the equations depends on the fractal dimension of the medium.

2 Statement and Numerical Solution of the Problem

In the problem under consideration, the porous medium is divided into two zones, in one zone where the liquid is considered to be mobile, and in the second zone the liquid is considered immobile, but the movement of the solute is observed due to diffusion. Solute transport between these two zones is usually described by a first order kinetic equation. Such a process in the one-dimensional case (a semi-infinite medium) can be written by the following equations

$$\theta_m \frac{\partial c_m}{\partial t} + \gamma \theta_{im} \frac{\partial^\alpha c_{im}}{\partial t^\alpha} = \theta_m D_m \frac{\partial^\beta c_m}{\partial x^\beta} - v_m \theta_m \frac{\partial c_m}{\partial x}, \tag{1}$$

$$\gamma \theta_{im} \frac{\partial^\alpha c_{im}}{\partial t^\alpha} = \omega (c_m - c_{im}), \tag{2}$$

where θ_m , θ_{im} are the porosity coefficient, c_m , c_{im} are the volumetric concentration of the solute, v_m is the average velocity of the solution, γ is the mass transfer coefficient, $[\gamma] = T^{\alpha-1}$, $[\omega] = T^{-1}$.

The initial and boundary conditions have the form:

$$c_m(0, x) = 0, \quad c_{im}(0, x) = 0, \tag{3}$$

$$c_m(t, 0) = c_0, \quad c_m(t, \infty) = 0. \tag{4}$$

The orders of fractional derivatives α and β changes in the following interval: $0 < \alpha < 1$, $1 < \beta \leq 2$.

For the numerical solution of problem (1)–(4), we use the method of finite differences [22]. In the domain, $\Omega = \{0 \leq x \leq \infty, 0 \leq t \leq T\}$ we introduce a uniform grid $\omega_{h\tau} = \{(x_i, t_j), x_i = ih, i = \overline{0, N}, h = L/N, t_j = j\tau, j = \overline{0, M}, \tau = T/M\}$, where h is the grid step in coordinate x , where τ is the grid step in time, L is the characteristic length of the porous medium.

To approximate fractional time derivatives, we use the following relationship [23–25].

$$\frac{\partial^\alpha c_{im}}{\partial t^\alpha} = \frac{\tau^{1-\alpha}}{\Gamma(2-\alpha)} \left[\sum_{l=0}^{j-1} \frac{(c_{im})_i^{l+1} - (c_{im})_i^l}{\tau} \cdot ((j-l+1)^{1-\alpha} - (j-l)^{1-\alpha}) + \frac{(c_{im})_i^{j+1} - (c_{im})_i^j}{\tau} \right].$$

Difference approximations Eq. (1) has the form

$$\begin{aligned} & \theta_m \frac{(c_m)_i^{j+1} - (c_m)_i^j}{\tau} \\ & + \gamma \theta_{im} \left(\frac{1}{\Gamma(2-\alpha)\tau^\alpha} \left[\sum_{l=0}^{j-1} ((c_{im})_i^{l+1} - (c_{im})_i^l) ((j-l+1)^{1-\alpha} - (j-l)^{1-\alpha}) \right. \right. \\ & \left. \left. + ((c_{im})_i^{j+1} - (c_{im})_i^j) \right] \right) \\ & = \theta_m D_m \frac{1}{\Gamma(3-\beta) * h^\beta} * \left(\sum_{l=0}^{i-1} ((c_m)_{i-(l+1)}^j - 2(c_m)_{i-l}^j + (c_m)_{i-(l-1)}^j) \right) \\ & * ((l+1)^{2-\beta} - (l)^{2-\beta}) - v_m \theta_m \frac{(c_m)_{i+1}^j - (c_m)_{i-1}^j}{2h}, \end{aligned} \quad (5)$$

Difference approximations, the kinetics Eq. (2) takes the form

$$\begin{aligned} & \gamma \theta_{im} \left(\frac{1}{\Gamma(2-\alpha)\tau^\alpha} \left[\sum_{l=0}^{j-1} ((c_{im})_i^{l+1} - (c_{im})_i^l) ((j-l+1)^{1-\alpha} - (j-l)^{1-\alpha}) \right. \right. \\ & \left. \left. + ((c_{im})_i^{j+1} - (c_{im})_i^j) \right] \right) = \omega ((c_m)_i^j - (c_{im})_i^{j+1}) \end{aligned} \quad (6)$$

The initial and boundary conditions are approximated as follows

$$(c_m)_i^0 = 0 \quad (c_{im})_i^0 = 0 \quad (7)$$

$$(c_m)_0^j = 0 \quad (c_m)_N^j = 0 \quad (8)$$

Difference Eq. (5) after some operations takes the form

$$\begin{aligned} & \theta_m \frac{(c_m)_i^{j+1} - (c_m)_i^j}{\tau} \\ & + \gamma \theta_{im} \left(\frac{1}{\Gamma(2-\alpha)\tau^\alpha} \left[\sum_{l=0}^{j-1} ((c_{im})_i^{l+1} - (c_{im})_i^l) ((j-l+1)^{1-\alpha} - (j-l)^{1-\alpha}) \right. \right. \\ & \left. \left. + ((c_{im})_i^{j+1} - (c_{im})_i^j) \right] \right) \\ & = \theta_m D_m \frac{1}{\Gamma(3-\beta) * h^\beta} * \left(\sum_{l=0}^{i-1} ((c_m)_{i-(l+1)}^j - 2(c_m)_{i-l}^j + (c_m)_{i-(l-1)}^j) \right) \\ & * ((l+1)^{2-\beta} - (l)^{2-\beta}) - v_m \theta_m \frac{(c_m)_{i+1}^j - (c_m)_{i-1}^j}{2h}, \\ & (c_m)_i^{j+1} = \frac{\tau D_m}{\Gamma(3-\beta) * h^\beta} \left(\sum_{l=0}^{i-1} ((c_m)_{i-(l+1)}^j - 2(c_m)_{i-l}^j + (c_m)_{i-(l-1)}^j) \right) \end{aligned}$$

$$\begin{aligned}
 & \left((l+1)^{2-\beta} - l^{2-\beta} \right) - \tau v_m \frac{(c_m)_{i+1}^j - (c_m)_{i-1}^j}{2h} \\
 & - \frac{\gamma \tau \theta_{im}}{\theta_m \Gamma(2-\alpha) \tau^\alpha} \sum_{l=0}^{j-1} \left((c_{im})_i^{l+1} - (c_{im})_i^l \right) \left((j-l+1)^{1-\alpha} - (j-l)^{1-\alpha} \right) \\
 & + \left((c_{im})_i^{j+1} - (c_{im})_i^j \right) + (c_m)_i^j.
 \end{aligned}$$

After some simple arithmetic operations, the difference kinetics Eq. (6) has the form

$$\begin{aligned}
 & \sum_{l=0}^{j-1} \left((c_{im})_i^{l+1} - (c_{im})_i^l \right) \left((j-l+1)^{1-\alpha} - (j-l)^{1-\alpha} \right) + \left((c_{im})_i^{j+1} - (c_{im})_i^j \right) \\
 & = \frac{\Gamma(2-\alpha) \tau^\alpha}{\gamma \theta_{im}} \omega \left((c_m)_i^j - (c_{im})_i^{j+1} \right) \\
 & (c_{im})_i^{j+1} + \frac{\omega \Gamma(2-\alpha) \tau^\alpha}{\gamma \theta_{im}} (c_{im})_i^{j+1} = \frac{\omega \Gamma(2-\alpha) \tau^\alpha}{\gamma \theta_{im}} (c_m)_i^j + (c_{im})_i^j \\
 & - \sum_{l=0}^{j-1} \left((c_{im})_i^{l+1} - (c_{im})_i^l \right) \left((j-l+1)^{1-\alpha} - (j-l)^{1-\alpha} \right) \\
 & (c_{im})_i^{j+1} \left(1 + \frac{\omega \Gamma(2-\alpha) \tau^\alpha}{\gamma \theta_{im}} \right) = \frac{\omega \Gamma(2-\alpha) \tau^\alpha}{\gamma \theta_{im}} (c_m)_i^j + (c_{im})_i^j \\
 & - \sum_{l=0}^{j-1} \left((c_{im})_i^{l+1} - (c_{im})_i^l \right) \left((j-l+1)^{1-\alpha} - (j-l)^{1-\alpha} \right) \\
 & (c_{im})_i^{j+1} = \left((c_{im})_i^j - \sum_{l=0}^{j-1} \left((c_{im})_i^{l+1} - (c_{im})_i^l \right) \left((j-l+1)^{1-\alpha} - (j-l)^{1-\alpha} \right) \right) \\
 & + \frac{\omega \Gamma(2-\alpha) \tau^\alpha}{\gamma \theta_{im}} (c_m)_i^j / \left(1 + \frac{\omega \Gamma(2-\alpha) \tau^\alpha}{\gamma \theta_{im}} \right)
 \end{aligned}$$

3 Results and Discussion

For the numerical solution of the problem (1)–(4) the following initial data values were used: $v_m = 10^{-4} \text{ m/s}$, $D_m = 10^{-5} \text{ m}^\beta/\text{s}$, $\omega = 10^{-6} \text{ 1/s}$, $\tau = 1$, $h = 0.1$, $\theta_m = 0.4 \text{ m}^3/\text{m}^3$, $\theta_{im} = 0.1 \text{ m}^3/\text{m}^3$.

Some results of numerical calculations are shown in Fig. 1, 2, 3, 4, 5 and 6. As can be seen from Fig. 1, a decrease in the order of the derivative β from 2 will lead to a more diffuse distribution of the concentration field c_m . Comparing Fig. 1 a, b one can notice more advanced profiles c_m in the direction of x in case b) and c). This corresponds to the case of “fast diffusion”. This distribution of concentration in the macropore is also reflected in the distribution in the micropore. On Fig. 2 shows the results when decreasing α from 1. This leads to a slowdown in the spread of solute in the zone c_{im} (Fig. 2b). Comparing the Fig. 3a and Fig. 1a one can notice the intensification of the movement of the solute in c_m and c_{im} with a decrease in the values of β . Thus, “fast diffusion”

in c_m leads to the same fast diffusion” in c_{im} . Comparing the Fig. 4b and Fig. 2b one can notice a slowdown in the progress of the solute in the micropore with a decrease in the values of α . In Fig. 5 shows the change in the concentration profiles at different values of diffusion coefficient. With an increase in the value of the diffusion coefficient, a wider distribution of the concentration profiles in the zones c_m and, c_{im} respectively, is observed. In Fig. 6 shows the change in concentration profiles at different time points. Based on the results obtained, more widespread concentration profiles were determined with increasing time.

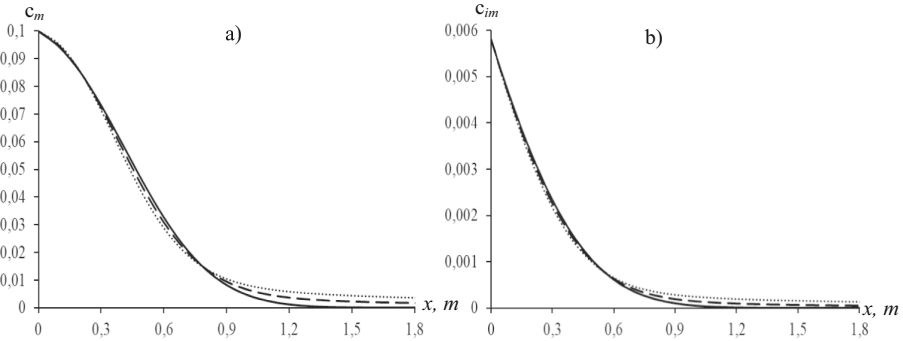


Fig. 1. Concentration c_m (a) and c_{im} , (b) profiles at $v_m = 10^{-4} m/s$, $D_m = 10^{-5} m^2/s$, $t = 3600 s$, $\omega = 10^{-6} 1/s$, $\gamma = 0.6$, $\alpha = 1$, $\beta = 1.6$ - - - , $\beta = 1.8$, ——— $\beta = 2$.

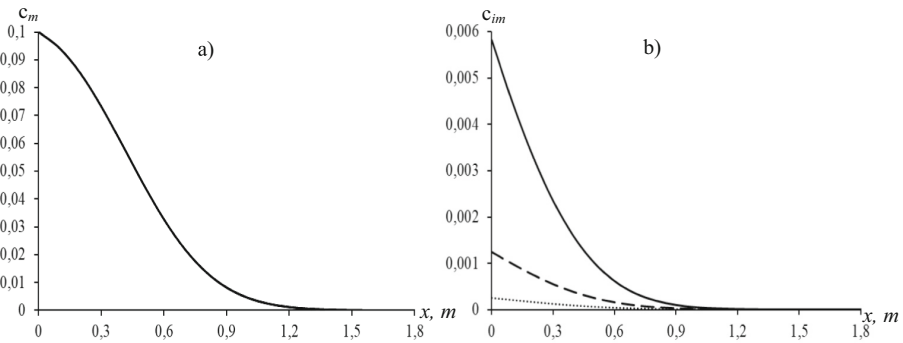


Fig. 2. Concentration c_m (a) and c_{im} , (b) profiles at $v_m = 10^{-4} m/s$, $D_m = 10^{-5} m^2/s$, $t = 3600 s$, $\omega = 10^{-6} 1/s$, $\gamma = 0.6$, $\beta = 2$ $\alpha = 0.6$ - - - -, $\alpha = 0.8$, ——— $\alpha = 1$.

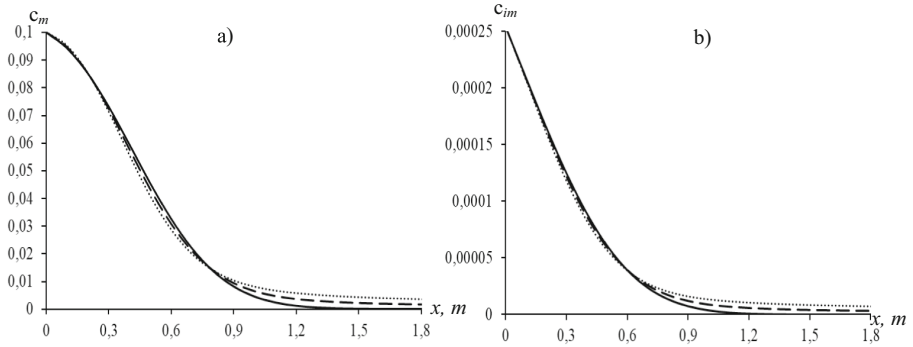


Fig. 3. Concentration c_m (a) and c_{im} , (b) profiles at $v_m = 10^{-4} m/s$, $D_m = 10^{-5} m^\beta/s$, $t = 3600 s$, $\omega = 10^{-6} 1/s$, $\gamma = 0.6$, $\alpha = 0.6$, , $\beta = 1.6$ - - - - , $\beta = 1.8$, ——— $\beta = 2$.

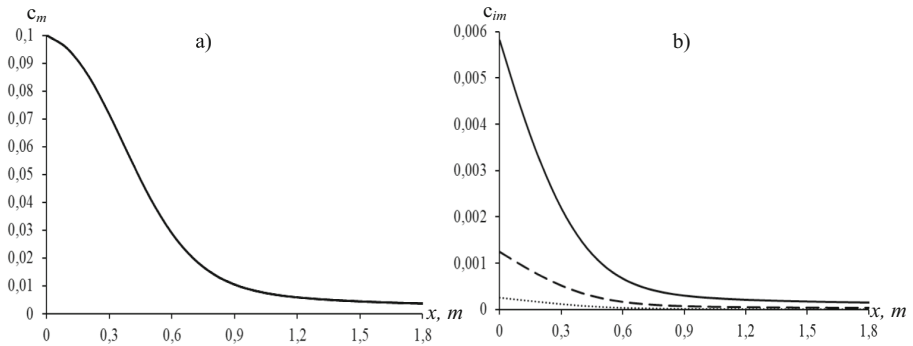


Fig. 4. Concentration c_m (a) and c_{im} , (b) profiles at $v_m = 10^{-4} m/s$, $D_m = 10^{-5} m^\beta/s$, $t = 3600 s$, $\omega = 10^{-6} 1/s$, $\gamma = 0.6$, $\beta = 1.6$, , $\alpha = 0.6$ - - - - , $\alpha = 0.8$, ——— $\alpha = 1$.

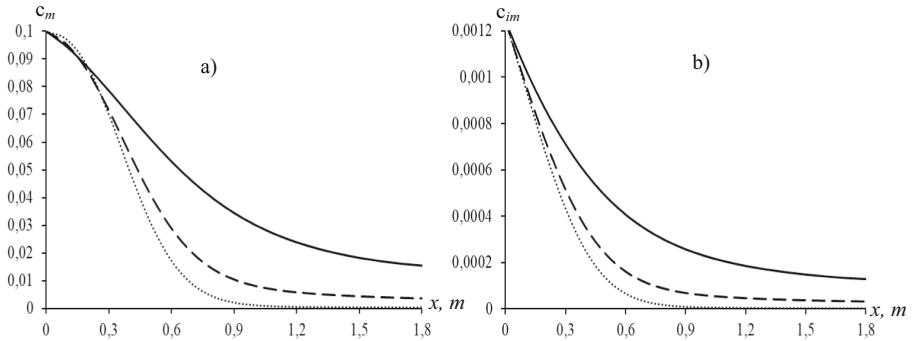


Fig. 5. Concentration c_m (a) and c_{im} , (b) profiles at $\alpha = 0.8$, $\beta = 1.6$, $v_m = 10^{-4} m/s$, $t = 3600 s$, $\omega = 10^{-6} 1/s$, $\gamma = 0.6$, , $D_m = 10^{-6}$ - - - - , $D_m = 10^{-5}$, ——— $D_m = 5 \cdot 10^{-5}$.

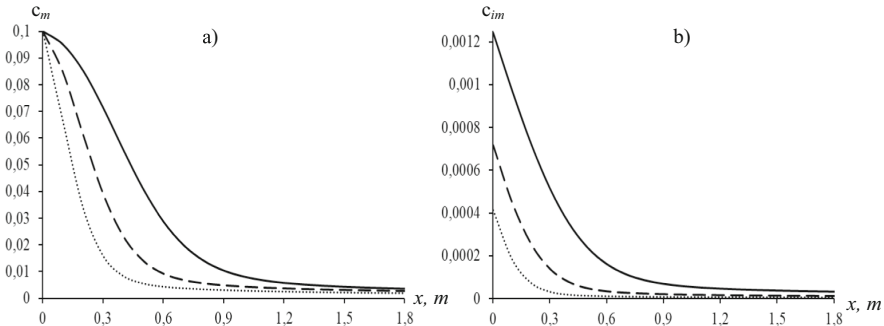


Fig. 6. Concentration c_m (a) and c_{im} . (b) profiles at $\alpha = 0.8, \beta = 1.6, v_m = 10^{-4} \text{ m/s}, D_m = 10^{-5} \text{ m}^2/\text{s}, \omega = 10^{-6} \text{ 1/s}, \gamma = 0.6, \dots\dots\dots t = 900 \text{ c}, - - - t = 1800 \text{ c}, \text{ — } t = 3600 \text{ c}$.

4 Conclusion

In this work, the problem of anomalous transport in porous media with a fractal structure is posed and numerically solved. For the numerical solution of the fractional differential equation, Caputo's definition was used. The numerical analysis shows that the anomalous process significantly affects the characteristics of the solute transport in both zones of the medium, i.e. in both micro and macropores. The anomaly of the transport is characterized by the order of the derivative in the diffusion terms of the transport equations in the macropore (β) and micropore (α). Reducing the order of the derivative in the diffusion terms of the transport equations in both zones leads to "fast diffusion". The decrease in α leads to "slow diffusion" in the micropore.

References

1. Khuzhayorov, B.Kh.: Filtration of nonhomogeneous liquids in porous media. Tashkent: Fan (2012). -280 p. Monograph (in Russian)
2. Khuzhayorov, B.Kh., Makhmudov, Zh.M.: Mathematical models of filtration of non-homogeneous liquids in porous media. Tashkent: Fan (2014). -280 p. Monograph (in Russian)
3. Korchagina, A.N., Merzhievsky, L.A.: Numerical modeling of diffusion processes in fractal media. Uchenye zapiski ZabGU **3**(50), 53–59 (2013). (in Russian)
4. Hassanizadeh, S.M.: On the transient non-Fickian dispersion theory. transport in Porous Media, **23**, 107–124 (1996)
5. Klages, R., Radons, G., Sokolov, I.M (Ed.): Anomalous Transport: Foundation and Applications. WILEY-VCH (2008)
6. Baeumer, B., Meerschaert, M.M.: Stochastic solutions for fractional Cauchy problems. *Frac. Calc. Appl. Anal.* **4**, 481–500 (2001)
7. Baeumer, B., Meerschaert, M.M., Benson, D.A., Wheatcraft, S.W.: Subordinated advection–dispersion equation for contaminant transport. *Water Resour. Res.* **37**, 1543–1550 (2001)
8. Barkai, E., Metzler, R., Klafter, J.: From continuous time random walks to the fractional Fokker-Planck equation. *Phys. Rev. E* **61**, 132–138 (2000)
9. Benson, D., Wheatcraft, S., Meerschaert, M.: Application of a fractional advection–dispersion equation. *Water Resour. Res.* **36**, 1403–1412 (2000)

10. Khuzhayorov, B.Kh., Usmonov, A.I.: Solute transport in porous cylindrical media with a fractal structure. Uzbek journal "Problems of Mechanics" (3), 39–44 (2019). (in Russian)
11. Meerschaert, M.M., Tadjeran, C.: Finite difference approximations for fractional advection-dispersion flow equations. *J. Comput. Appl. Math.* **172**, 65–77 (2004)
12. Liu, F., Zhuang, P., Anh, V., Turner, I., Burrage, K.: Stability and convergence of the difference methods for the space-time fractional advection-diffusion equation. *Appl. Math. Comput.* **191**, 12–20 (2007)
13. Chen, J.Sh., Chen, J.T., Chen, W.L., Liang, Ch.P., Lin, Ch.W.: Analytical solutions to two-dimensional advection–dispersion equation in cylindrical coordinates in finite domain subject to first- and third-type inlet boundary conditions. *J. Hydrol.* **405**, 522–531 (2011)
14. Chen, J.Sh., Chen, J.T., Chen, W.L., Liang, Ch.P., Lin, Ch.W.: Exact analytical solutions for two-dimensional advection–dispersion equation in cylindrical coordinates subject to third-type inlet boundary condition. *Advances in Water Res.* **34**, 365–374 (2011)
15. van Genuchten, M.Th.*, Wagenet, R.J.: Two-Site/Two-Region Models for Pesticide Transport and Degradation: Theoretical Development and Analytical Solutions. *Soil Sci. Soc. Am. J.* (1989)
16. Van Genuchten, M.Th., Wierenga, P.J.: Mass transfer studies in sorbing porous media: II. Experimental Evaluation with Tritium (H₂O). *Soil Sci. Soc. Am. J.* **41**, 272–278 (1977)
17. Sharma, P.K., Shukla, S.K.: Modeling for solute transport in mobile–immobile soil column experiment. *ISH J. Hydraulic Eng.* (2016)
18. Nigmatullin, R.R.: The realization of the generalized transfer equation in a medium with fractal geometry. *Phys. Stat. Sol. (b)* **133**, 425–430 (1986)
19. Fomin, S., Chugunov, V., Hashida, T.: Application of fractional differential equations for modeling the anomalous diffusion of contaminant from fracture into porous rock matrix with bordering alteration zone. *Transp. Porous Media* **81**, 187–205 (2010)
20. Fomin, S., Chugunov, V., Hashida, T.: Mathematical modeling of anomalous diffusion in porous media. *Fract. Diff. Cal.* **1**(1), 1–28 (2011)
21. Fomin, S., Chugunov, V., Hashida, T.: The effect of non-Fickian diffusion into surrounding rocks on contaminant transport in fractured porous aquifer. *Proc. Roy. Soc. A* **461**, 2923–2939 (2005)
22. Samarsky, A.A.: Theory of difference schemes. Nauka, Moscow (1989). -616 p (in Russian)
23. Xia, Y., Wu, J., Zhou, L.: Numerical solution of time-space fractional advection-dispersion equations. *ICES* **9**(2), 117–126 (2009)
24. Khuzhayorov, B., Usmonov, A., Nik Long, N.M.A., Fayziev, B.: Anomalous solute transport in a cylindrical two-zone medium with fractal structure. *Appl. Sci. (Switzerland)* **10**(15), 5349 (2020). <https://doi.org/10.3390/app10155349>
25. Li, G.S., Sun, C.L., Jia, X.Z., Du, D.H.: Numerical solution to the multi-term time fractional diffusion equation in a finite domain. *Number. Math. Theor. -Meth. Appl.* **9**, 337–357 (2016)



Initial-Boundary Value Problems for the Loaded Hallaire Equation with Gerasimov–Caputo Fractional Derivatives of Different Orders

Murat Beshtokov^(✉)

Institute of Applied Mathematics and Automation, Kabardino-Balkarian Scientific
Center of RAS, 89A Shortanova Str., 360000 Nalchik, Russia

beshtokov-murat@yandex.ru

Abstract. Boundary-value problems for the loaded Hallaire equation with variable coefficients and Gerasimov–Caputo fractional derivatives of different orders are studied. Using the method of energy inequalities for various relations between α and β , a priori estimates in differential and difference interpretations are obtained for the solution of the problem under consideration, from which the uniqueness and stability of the solution with respect to the initial data and the right-hand side, as well as the convergence of the solution of the difference problem to the solution of the differential problem.

Keywords: Hillaire’s equation · loaded equation · boundary value problem · Gerasimov–Caputo fractional derivative · a priori estimate · difference scheme · stability and convergence

1 Introduction

At present, it has become obvious that when solving many problems in mechanics, physics, and biology, one often encounters media and systems that are well interpreted as fractals; examples of the latter can be highly porous media, such as, for example, soil. The solution of various problems for such media leads to boundary value problems for differential equations with fractional derivatives. Partial differential equations of fractional order are a generalization of partial differential equations of integer order and are of great theoretical and practical interest [1–3].

Questions related to moisture transfer in soils lead to pseudo-parabolic equations or the Hallaire equation [4–6].

In this work boundary-value problems for the loaded Hallaire equation with variable coefficients and Gerasimov–Caputo fractional derivatives of different orders are studied. Using the method of energy inequalities for various relations between α and β , a priori estimates in differential and difference interpretations are obtained for the solution of the problem under consideration, from which

the uniqueness and stability of the solution with respect to the initial data and the right-hand side, as well as the convergence of the solution of the difference problem to the solution of the differential problem.

The works [3, 7–12] are devoted to numerical methods for solving boundary value problems for various equations of fractional order. In [7, 8], results were obtained that, as in the classical case ($\alpha = 1$), allow one to apply the method of energy inequalities to find a priori estimates of boundary value problems for a fractional order equation in differential and difference interpretations. In [9, 10], boundary value problems are studied for various loaded differential equations of integer and fractional orders.

This work is a continuation of the author’s series of works in this direction [10–12].

2 Materials and Methods

2.1 Problem Statement

In a closed rectangle $\bar{Q}_T = \{(x, t) : 0 \leq x \leq l, 0 \leq t \leq T\}$ consider the following problem

$$\begin{aligned} \partial_{0t}^\alpha u &= \frac{\partial}{\partial x} \left(k(x, t) \frac{\partial u}{\partial x} \right) + \partial_{0t}^\beta \frac{\partial}{\partial x} \left(\eta(x) \frac{\partial u}{\partial x} \right) + r(x, t) \frac{\partial u}{\partial x} - \\ &- \sum_{s=1}^m q_s(x, t) u(x_s, t) + f(x, t), \quad 0 < x < l, \quad 0 < t \leq T, \end{aligned} \tag{1}$$

$$u(0, t) = u(l, t) = 0, \quad 0 \leq t \leq T, \tag{2}$$

$$u(x, 0) = u_0(x), \quad 0 \leq x \leq l, \tag{3}$$

where

$$0 < c_0 \leq k(x, t), \eta(x) \leq c_1, \quad |q(x, t)|, |r(x, t)|, |r_x(x, t)|, |k_x(x, t)| \leq c_2, \tag{4}$$

$\partial_{0t}^\gamma u = \frac{1}{\Gamma(1-\gamma)} \int_0^t \frac{u_\tau(x, \tau)}{(t-\tau)^\gamma} d\tau$ is a fractional derivative in the sense of Gerasimov–Caputo order γ , $0 < \gamma < 1$, x_s ($s = 1, 2, \dots, m$) are arbitrary interval points $[0, l]$.

In what follows, we will assume that problem (1)–(3) has a unique solution possessing the necessary derivatives. We will also assume that the coefficients of the equation and boundary conditions satisfy the necessary smoothness conditions which provide the required order of approximation of the difference scheme.

Denote by $M_i, (i = 1, 2, \dots,)$ positive constant numbers depending only on the input data of the original problem.

2.2 A Priori Estimate in Differential Form

To obtain an a priori estimate for the solution of problem (1)–(3) in differential form, we multiply Eq. (1) scalarly by u

$$\begin{aligned} (\partial_{0t}^\alpha u, u) &= ((ku_x)_x, u) + \left(\partial_{0t}^\beta (\eta(x)u_x)_x, u \right) \\ &+ (ru_x(x, t), u) - \left(\sum_{s=1}^m q_s(x, t)u(x_s, t), u \right) + (f, u). \end{aligned} \quad (5)$$

where $(u, v) = \int_0^l uv dx$, $(u, u) = \|u\|_0^2$, where u, v are functions defined on $[0, l]$.

We transform the integrals in identity (5), using the Cauchy inequality with ε ([13], p. 100, [15]) and Lemma 1 from [7], we obtain:

$$(\partial_{0t}^\alpha u, u) \geq \frac{1}{2} (1, \partial_{0t}^\alpha u^2) = \frac{1}{2} \partial_{0t}^\alpha \|u\|_0^2, \quad (6)$$

$$((ku_x)_x, u) = \int_0^l u (ku_x)_x dx = uk u_x|_0^l - \int_0^l k u_x^2 dx, \quad (7)$$

$$\begin{aligned} \left(\partial_{0t}^\beta (\eta u_x)_x, u \right) &= \int_0^l u \partial_{0t}^\beta (\eta u_x)_x dx = u \partial_{0t}^\beta (\eta u_x)|_0^l - \int_0^l \eta(x) u_x \partial_{0t}^\beta u_x dx \\ &\leq u \partial_{0t}^\beta (\eta u_x)|_0^l - \frac{1}{2} \int_0^l \eta \partial_{0t}^\beta (u_x)^2 dx, \end{aligned} \quad (8)$$

$$(ru_x, u) = \int_0^l r u u_x dx \leq \frac{c_2^2}{4\varepsilon} \int_0^l u^2 dx + \varepsilon \int_0^l u_x^2 dx \leq M_1(\varepsilon) \|u\|_0^2 + \varepsilon \|u_x\|_0^2, \quad (9)$$

$$\begin{aligned} &- \left(\sum_{s=1}^m q_s(x, t)u(x_s, t), u \right) = - \int_0^l u \sum_{s=1}^m q_s(x, t)u(x_s, t) dx \\ &= - \sum_{s=1}^m u(x_s, t) \int_0^l q_s(x, t)u dx \leq \sum_{s=1}^m \left(u^2(x_s, t) + \frac{1}{4} \left(\int_0^l q_s(x, t)u dx \right)^2 \right) \\ &\leq M_2(\varepsilon) \|u\|_0^2 + \varepsilon \|u_x\|_0^2, \end{aligned} \quad (10)$$

$$(f, u) = \int_0^l f u dx \leq \frac{1}{2} \|u\|_0^2 + \frac{1}{2} \|f\|_0^2. \quad (11)$$

Taking into account (2) and transformations (6)–(11), from (5) we find

$$\frac{1}{2} \partial_{0t}^\alpha \|u\|_0^2 + \frac{1}{2} \partial_{0t}^\beta \int_0^l \eta (u_x)^2 dx + c_0 \|u_x\|_0^2 \leq \varepsilon \|u_x\|_0^2 + M_3(\varepsilon) \|u\|_0^2 + \frac{1}{2} \|f\|_0^2. \quad (12)$$

Choosing $\varepsilon = \frac{c_0}{2}$, from (12) we get

$$\partial_{0t}^\alpha \|u\|_0^2 + \partial_{0t}^\beta \int_0^l \eta (u_x)^2 dx + \|u_x\|_0^2 \leq M_4 \|u\|_0^2 + M_5 \|f\|_0^2. \quad (13)$$

1) Consider the case when $\alpha > \beta$, then applying the fractional integration operator $D_{0t}^{-\alpha}$ to both sides of (13), we obtain

$$\|u\|_0^2 + D_{0t}^{-(\alpha-\beta)}\|u_x\|_0^2 \leq M_6 D_{0t}^{-\alpha}\|u\|_0^2 + M_7 (D_{0t}^{-\alpha}\|f\|_0^2 + \|u_0(x)\|_0^2), \tag{14}$$

where $D_{0t}^{-\gamma}u = \frac{1}{\Gamma(\gamma)} \int_0^t \frac{u d\tau}{(t-\tau)^{1-\gamma}}$ is a fractional derivative in the sense of Riemann-Liouville of order $\gamma, 0 < \gamma < 1$.

Based on Lemma 2 [7], from (14) we find the following a priori estimate

$$\|u\|_1^2 \leq M_8 (D_{0t}^{-\alpha}\|f\|_0^2 + \|u_0(x)\|_0^2), \tag{15}$$

where $\|u\|_1^2 = \|u\|_0^2 + D_{0t}^{-(\alpha-\beta)}\|u_x\|_0^2$.

2) Consider the case when $\alpha = \beta$, then applying the fractional integration operator $D_{0t}^{-\alpha}$ to both sides of (13), we obtain

$$\|u\|_0^2 + \|u_x\|_0^2 \leq M_9 D_{0t}^{-\alpha}\|u\|_0^2 + M_{10} (D_{0t}^{-\alpha}\|f\|_0^2 + \|u_0(x)\|_0^2 + \|u'_0(x)\|_0^2). \tag{16}$$

Based on Lemma 2 [7], from (16) we find the following a priori estimate

$$\|u\|_2^2 \leq M_{11} (D_{0t}^{-\alpha}\|f\|_0^2 + \|u_0(x)\|_0^2 + \|u'_0(x)\|_0^2), \tag{17}$$

where $\|u\|_2^2 = \|u\|_0^2 + \|u_x\|_0^2$.

3) Consider the case when $\alpha < \beta$, then applying the fractional integration operator $D_{0t}^{-\beta}$ to both sides of (13), we obtain

$$\|u_x\|_0^2 + D_{0t}^{-(\beta-\alpha)}\|u\|_0^2 \leq M_{12} D_{0t}^{-\beta}\|u\|_0^2 + M_{13} (D_{0t}^{-\beta}\|f\|_0^2 + \|u'_0(x)\|_0^2). \tag{18}$$

By virtue of the condition (2), the inequality $\|u\|_0^2 \leq 2l^2\|u_x\|_0^2$ [14], is valid, then from (18) we obtain

$$\|u_x\|_0^2 + D_{0t}^{-(\beta-\alpha)}\|u\|_0^2 \leq M_{14} D_{0t}^{-\beta}\|u_x\|_0^2 + M_{13} (D_{0t}^{-\beta}\|f\|_0^2 + \|u'_0(x)\|_0^2). \tag{19}$$

Based on Lemma 2 [7], from (19) we find the following a priori estimate

$$\|u\|_3^2 \leq M_{15} (D_{0t}^{-\beta}\|f\|_0^2 + \|u'_0(x)\|_0^2), \tag{20}$$

where $\|u\|_3^2 = \|u_x\|_0^2 + D_{0t}^{-(\beta-\alpha)}\|u\|_0^2$.

2.3 Stability and Convergence of the Difference Scheme

For solution of problem (1)–(3) we use the finite difference method. In the closed rectangle \bar{Q}_T we introduce a uniform grid $\bar{\omega}_{h\tau} = \bar{\omega}_h \times \bar{\omega}_\tau$, where $\bar{\omega}_h = \{x_i = ih, i = 0, \bar{N}, h = l/N\}$, $\bar{\omega}_\tau = \{t_j = j\tau, j = 0, 1, \dots, j_0, \tau = T/j_0\}$. On the uniform grid $\bar{\omega}_{h\tau}$, we associate the differential problem (1)–(3) with the difference scheme

of the order of approximation $O(h^2 + \tau^2)$ for $\alpha = \beta$ and $O(h^2 + \tau^{2-\max\{\alpha,\beta\}})$ for $\alpha \neq \beta$:

$$\begin{aligned} \Delta_{0t_{j+\sigma}}^\alpha y &= \chi_i^j \left(a_i^j y_{\bar{x}}^{(\sigma)} \right)_{x,i} + \Delta_{0t_{j+\sigma}}^\beta (\gamma_i y_{\bar{x}})_{x,i} + b_i^{-j} a_i^j y_{\bar{x},i}^{(\sigma)} + b_i^{+j} a_{i+1}^j y_{x,i}^{(\sigma)} \\ &\quad - \sum_{s=1}^m d_{s,i}^j \left(y_{i_s}^{(\sigma)} x_{i_s}^- + y_{i_s+1}^{(\sigma)} x_{i_s}^+ \right) + \varphi_i^j, \quad (x, t) \in \omega_{h,\tau}, \end{aligned} \quad (21)$$

$$y_0^{(\sigma)} = y_N^{(\sigma)} = 0, \quad (22)$$

$$y(x, 0) = u_0(x), \quad (23)$$

where $\Delta_{0t_{j+\sigma}}^\gamma y = \frac{\tau^{1-\gamma}}{\Gamma(2-\gamma)} \sum_{s=0}^j c_{j-s}^{(\gamma,\sigma)} y_t^s$ is a discrete analogue of the fractional Gerasimova-Caputo derivative of order γ , $0 < \gamma < 1$, which assures order of accuracy $O(\tau^{3-\gamma})$ for $\sigma = 1 - \frac{\gamma}{2}$, and $O(\tau^{2-\gamma})$ for $\sigma = 0.5$ [8].

$$a_0^{(\gamma,\sigma)} = \sigma^{1-\gamma}, \quad a_l^{(\gamma,\sigma)} = (l + \sigma)^{1-\gamma} - (l - 1 + \sigma)^{1-\gamma}, \quad l \geq 1,$$

$$b_l^{(\gamma,\sigma)} = \frac{1}{2-\gamma} \left[(l + \sigma)^{2-\gamma} - (l - 1 + \sigma)^{2-\gamma} \right] - \frac{1}{2} \left[(l + \sigma)^{1-\alpha} + (l - 1 + \sigma)^{1-\gamma} \right], \quad l \geq 1,$$

if $j = 0$, then $c_0^{(\gamma,\sigma)} = a_0^{(\gamma,\sigma)}$;

$$\text{if } j > 0, \text{ then } c_s^{(\gamma,\sigma)} = \begin{cases} a_0^{(\gamma,\sigma)} + b_1^{(\gamma,\sigma)}, & s = 0; \\ a_s^{(\gamma,\sigma)} + b_{s+1}^{(\gamma,\sigma)} - b_s^{(\gamma,\sigma)}, & 1 \leq s \leq j - 1; \\ a_j^{(\gamma,\sigma)} - b_j^{(\gamma,\sigma)}, & s = j, \end{cases}$$

$$c_s^{(\gamma,\sigma)} > \frac{1-\gamma}{2} (s + \sigma)^{-\gamma} > 0, \quad \sigma = 1 - \frac{\gamma}{2}, \text{ for } \alpha = \beta, \text{ and } \sigma = 0.5, \text{ for } \alpha \neq \beta,$$

$$x_{i_s}^- = \frac{x_{i_s+1} - x_{i_s}}{h}, \quad x_{i_s}^+ = \frac{x_s - x_{i_s}}{h}, \quad x_{i_s} \leq x_s \leq x_{i_s+1},$$

$$a_i^j = k(x_{i-0.5}, t_{j+\sigma}), \quad \gamma_i = \eta(x_{i-0.5}), \quad b_i^{\pm j} = \frac{r^{\pm j}(x, t_{j+\sigma})}{k(x_i, t_{j+\sigma})}, \quad \varphi_i^j = f(x_i, t_{j+\sigma}),$$

$$r(x, t_{j+\sigma}) = r^+(x, t_{j+\sigma}) + r^-(x, t_{j+\sigma}), \quad |r(x, t_{j+\sigma})| = r^+(x, t_{j+\sigma}) - r^-(x, t_{j+\sigma}),$$

$$r^+(x, t_{j+\sigma}) = 0.5(r(x, t_{j+\sigma}) + |r(x, t_{j+\sigma})|) \geq 0,$$

$$r^-(x, t_{j+\sigma}) = 0.5(r(x, t_{j+\sigma}) - |r(x, t_{j+\sigma})|) \leq 0,$$

$$y^{(\sigma)} = \sigma y^{j+1} + (1 - \sigma) y^j, \quad d_{s,i}^j = q_s(x_i, t_{j+\sigma}), \quad a^{(+1)} = a_{i+1},$$

$\chi(x, t) = \frac{1}{1+R(x,t)}$, $R(x, t) = \frac{0.5h|r(x,t)|}{k(x,t)}$ is the difference Reynolds number.

An a priori estimate of the solution of problem (21)–(23) will be found by the method of energy inequalities; for this purpose, we introduce scalar products and a norm

$$(u, v) = \sum_{i=1}^{N-1} u_i v_i h, \quad (u, v] = \sum_{i=1}^N u_i v_i h, \quad (u, u) = (1, u^2) = \|u\|_0^2.$$

We multiply (21) scalarly by $y^{(\sigma)}$:

$$\begin{aligned} (\Delta_{0t_{j+\sigma}}^\alpha y, y^{(\sigma)}) &= (\chi (ay_{\bar{x}}^{(\sigma)})_x, y^{(\sigma)}) + (\Delta_{0t_{j+\sigma}}^\alpha (\gamma_i y_{\bar{x}})_x, y^{(\sigma)}) + (b^- ay_{\bar{x}}^{(\sigma)}, y^{(\sigma)}) \\ &+ (b^+ a^{(+1)} y_x^{(\sigma)}, y^{(\sigma)}) - \left(\sum_{s=1}^m d_{s,i}^j (y_{i_s}^{(\sigma)} x_{i_s}^- + y_{i_s+1}^{(\sigma)} x_{i_s}^+), y^{(\sigma)} \right) + (\varphi, y^{(\sigma)}). \end{aligned} \tag{24}$$

Estimating the sums in (24), using the ε -Cauchy inequality, (24) and Lemma 1 [8], we obtain:

$$(\Delta_{0t_{j+\sigma}}^\alpha y, y^{(\sigma)}) \geq \frac{1}{2} \Delta_{0t_{j+\sigma}}^\alpha \|y\|_0^2; \tag{25}$$

$$\begin{aligned} (\chi (ay_{\bar{x}}^{(\sigma)})_x, y^{(\sigma)}) &= \chi ay_{\bar{x}}^{(\sigma)} y^{(\sigma)} \Big|_0^N - (ay_{\bar{x}}^{(\sigma)}, (\chi y^{(\sigma)})_{\bar{x}}) \\ &= - (a\chi_{\bar{x}}, y_{\bar{x}}^{(\sigma)} y^{(\sigma)}) - (a\chi^{(-1)}, (y_{\bar{x}}^{(\sigma)})^2) \\ &\leq - (a\chi_{\bar{x}}, y_{\bar{x}}^{(\sigma)} y^{(\sigma)}) - \frac{1}{(1+hM_1)} (a\chi, (y_{\bar{x}}^{(\sigma)})^2); \end{aligned} \tag{26}$$

$$\begin{aligned} (\Delta_{0t_{j+\sigma}}^\beta (\gamma y_{\bar{x}})_x, y^{(\sigma)}) &= y^{(\sigma)} \Delta_{0t_{j+\sigma}}^\beta (\gamma y_{\bar{x}}) \Big|_0^N - (\gamma, y_{\bar{x}}^{(\sigma)} \Delta_{0t_{j+\sigma}}^\beta (y_{\bar{x}})) \\ &\leq - \left(\frac{\gamma}{2}, \Delta_{0t_{j+\sigma}}^\beta (y_{\bar{x}})^2 \right) \leq - \frac{c_0}{2} \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}}\|_0^2; \end{aligned} \tag{27}$$

$$\begin{aligned} &- \left(\sum_{s=1}^m d_s (y_{i_s}^{(\sigma)} x_{i_s}^- + y_{i_s+1}^{(\sigma)} x_{i_s}^+), y^{(\sigma)} \right) \\ &= - \sum_{s=1}^m \left((y_{i_s}^{(\sigma)} x_{i_s}^- + y_{i_s+1}^{(\sigma)} x_{i_s}^+) (d_s, y^{(\sigma)}) \right) \\ &\leq \sum_{s=1}^m \left(\frac{1}{2} (y_{i_s}^{(\sigma)} x_{i_s}^- + y_{i_s+1}^{(\sigma)} x_{i_s}^+)^2 + \frac{1}{2} (d_s, y^{(\sigma)})^2 \right) \\ &\leq \varepsilon \|y_{\bar{x}}^{(\sigma)}\|_0^2 + M_2(\varepsilon) \|y^{(\sigma)}\|_0^2; \end{aligned} \tag{28}$$

$$(\varphi, y^{(\sigma)}) \leq \frac{1}{2} \|y^{(\sigma)}\|_0^2 + \frac{1}{2} \|\varphi\|_0^2. \tag{29}$$

Taking into account the transformations (25)–(29), from (24) we get

$$\begin{aligned} \frac{1}{2} \Delta_{0t_{j+\sigma}}^\alpha \|y\|_0^2 + M_3 \|y_{\bar{x}}^{(\sigma)}\|_0^2 + \frac{c_0}{2} \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}}\|_0^2 &\leq \varepsilon \|y_{\bar{x}}^{(\sigma)}\|_0^2 - (a\chi_{\bar{x}}, y_{\bar{x}}^{(\sigma)} y^{(\sigma)}) \\ &+ (b^- ay_{\bar{x}}^{(\sigma)}, y^{(\sigma)}) + (b^+ a^{(+1)} y_x^{(\sigma)}, y^{(\sigma)}) + M_4(\varepsilon) \|y^{(\sigma)}\|_0^2 + \frac{1}{2} \|\varphi\|_0^2. \end{aligned} \tag{30}$$

We transform the the second, third and fourth expressions on the right-hand side of (30). Then using the Cauchy inequality with ε we obtain

$$\begin{aligned}
 & - \left(a\chi_{\bar{x}}, y_{\bar{x}}^{(\sigma)} y^{(\sigma)} \right] + \left(b^- a, y_{\bar{x}}^{(\sigma)} y^{(\sigma)} \right) + \left(b^+ a^{(+1)} y_x^{(\sigma)}, y^{(\sigma)} \right) \\
 & \leq \varepsilon \|y_{\bar{x}}^{(\sigma)}\|_0^2 + M_5(\varepsilon) \|y^{(\sigma)}\|_0^2.
 \end{aligned} \tag{31}$$

From (30), taking into account (31), with $\varepsilon = \frac{M_3}{2}$ we find

$$\Delta_{0t_{j+\sigma}}^\alpha \|y\|_0^2 + \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}}\|_0^2 + \|y_{\bar{x}}^{(\sigma)}\|_0^2 \leq M_6 \|y^{(\sigma)}\|_0^2 + M_7 \|\varphi\|_0^2. \tag{32}$$

We rewrite (32) in another form

$$\Delta_{0t_{j+\sigma}}^\alpha \|y\|_0^2 + \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}}\|_0^2 \leq M_8^\sigma \|y^{j+1}\|_0^2 + M_9^\sigma \|y^j\|_0^2 + M_7 \|\varphi\|_0^2. \tag{33}$$

1) Consider the case when $\alpha > \beta$, then, based on Lemma 7 [11], from (33) we obtain

$$\|y^{j+1}\|_0^2 \leq M_{10} \left(\|y^0\|_1^2 + \max_{0 \leq j' \leq j} \|\varphi^{j'}\|_0^2 \right), \tag{34}$$

where $\|y^0\|_1^2 = \|y^0\|_0^2 + \|y_{\bar{x}}^0\|_0^2$.

2) Consider the case when $\alpha \leq \beta$, then based on (22), the inequality $\|y\|_0^2 \leq 2l^2 \|y_{\bar{x}}\|_0^2$ [13] and Lemma 7 [11] from (32) we get:

$$\|y_{\bar{x}}^{j+1}\|_0^2 \leq M_{11} \left(\|y^0\|_1^2 + \max_{0 \leq j' \leq j} \|\varphi^{j'}\|_0^2 \right), \tag{35}$$

where $M_{10}, M_{11} = const > 0$, independent of h and τ .

2.4 Statement of the Boundary Value Problem with a Condition of the Third Kind and a Priori Estimate in Differential Form

We replace the second boundary condition in (2) by a condition of the third kind, then instead of (2) we have

$$\begin{cases} u(0, t) = 0, \\ -\Pi(l, t) = \beta(t)u(l, t) - \mu(t), \end{cases} \tag{36}$$

where

$$0 < c_0 \leq k, \eta \leq c_1, |\beta, r, q, r_x, k_x| \leq c_2, \Pi(x, t) = k(x, t)u_x + \partial_{0t}^\alpha (\eta u_x). \tag{37}$$

Repeating the reasoning (6)–(11), from (5) after some transformations we find the inequality

$$\begin{aligned}
 & \frac{1}{2} \partial_{0t}^\alpha \|u\|_0^2 + \frac{1}{2} \partial_{0t}^\beta \int_0^l \eta(u_x)^2 dx + c_0 \|u_x\|_0^2 \\
 & \leq u\Pi(x, t)|_0^l + M_1(\varepsilon) \|u\|_0^2 + \varepsilon \|u_x\|_0^2 + \frac{1}{2} \|f\|_0^2.
 \end{aligned} \tag{38}$$

We transform the first expression on the right-hand side of (38), then we get

$$\begin{aligned}
 u\Pi(x, t)\Big|_0^l &= \Pi(l, t)u(l, t) = u(l, t) (\mu(t) - \beta(t)u(l, t)) = -\beta(t)u^2(l, t) \\
 + \mu(t)u(l, t) &\leq M_2u^2(l, t) + \frac{1}{2}\mu^2(t) \leq M_3(\varepsilon)\|u\|_0^2 + \varepsilon\|u_x\|_0^2 + \frac{1}{2}\mu^2(t). \tag{39}
 \end{aligned}$$

Taking into account (39), from (38) with $\varepsilon = \frac{c_0}{2}$ we find

$$\partial_{0t}^\alpha\|u\|_0^2 + \partial_{0t}^\beta \int_0^l \eta(u_x)^2 dx + \|u_x\|_0^2 \leq M_4\|u\|_0^2 + M_5 (\|f\|_0^2 + \mu^2(t)). \tag{40}$$

- 1) Consider the case when $\alpha > \beta$, then applying the fractional integration operator $D_{0t}^{-\alpha}$ to both sides of the inequality (40), we obtain

$$\begin{aligned}
 &\|u\|_0^2 + D_{0t}^{-(\alpha-\beta)}\|u_x\|_0^2 \\
 &\leq M_6D_{0t}^{-\alpha}\|u\|_0^2 + M_7 (D_{0t}^{-\alpha} (\|f\|_0^2 + \mu^2(t)) + \|u_0(x)\|_0^2). \tag{41}
 \end{aligned}$$

Based on Lemma 2 [7], from (41) we find the following a priori estimate

$$\|u\|_1^2 \leq M_8 (D_{0t}^{-\alpha} (\|f\|_0^2 + \mu^2(t)) + \|u_0(x)\|_0^2). \tag{42}$$

- 2) Consider the case when $\alpha = \beta$, then applying the fractional integration operator $D_{0t}^{-\alpha}$ to both sides of the inequality (40), we obtain

$$\begin{aligned}
 &\|u\|_0^2 + \|u_x\|_0^2 \leq M_9D_{0t}^{-\alpha}\|u\|_0^2 \\
 &+ M_{10} (D_{0t}^{-\alpha} (\|f\|_0^2 + \mu^2(t)) + \|u_0(x)\|_0 + \|u'_0(x)\|_0^2). \tag{43}
 \end{aligned}$$

Based on Lemma 2 [7], from (43) we find the following a priori estimate

$$\|u\|_2^2 \leq M_{11} (D_{0t}^{-\alpha} (\|f\|_0^2 + \mu^2(t)) + \|u_0(x)\|_0 + \|u'_0(x)\|_0^2). \tag{44}$$

- 3) Consider the case when $\alpha < \beta$, then applying the fractional integration operator $D_{0t}^{-\beta}$ to both sides of the inequality (40), we obtain

$$\begin{aligned}
 &\|u_x\|_0^2 + D_{0t}^{-(\beta-\alpha)}\|u\|_0^2 \\
 &\leq M_{12}D_{0t}^{-\beta}\|u\|_0^2 + M_{13} \left(D_{0t}^{-\beta} (\|f\|_0^2 + \mu^2(t)) + \|u'_0(x)\|_0^2 \right). \tag{45}
 \end{aligned}$$

By virtue of the condition $u(0, t) = 0$, the inequality $\|u\|_0^2 \leq 2l^2\|u_x\|_0^2$ [14] is valid, then from (45) we get

$$\begin{aligned}
 &\|u_x\|_0^2 + D_{0t}^{-(\beta-\alpha)}\|u\|_0^2 \\
 &\leq M_{14}D_{0t}^{-\beta}\|u_x\|_0^2 + M_{13} \left(D_{0t}^{-\beta} (\|f\|_0^2 + \mu^2(t)) + \|u'_0(x)\|_0^2 \right). \tag{46}
 \end{aligned}$$

Based on Lemma 2 [7], from (46) we find the following a priori estimate

$$\|u\|_3^2 \leq M_{15} \left(D_{0t}^{-\beta} (\|f\|_0^2 + \mu^2(t)) + \|u'_0(x)\|_0^2 \right). \tag{47}$$

2.5 Stability and Convergence of the Difference Scheme

On the uniform grid $\bar{\omega}_{h\tau}$, we associate the differential problem (1), (36), (3) with the difference scheme:

$$\begin{aligned} \Delta_{0t_{j+\sigma}}^\alpha y &= \chi_i^j \left(a_i^j y_{\bar{x}}^{(\sigma)} \right)_{x,i} + \Delta_{0t_{j+\sigma}}^\beta (\gamma_i y_{\bar{x}})_{x,i} + b_i^{-j} a_i^j y_{\bar{x},i}^{(\sigma)} + b_i^{+j} a_{i+1}^j y_{x,i}^{(\sigma)} \\ &\quad - \sum_{s=1}^m d_{s,i}^j \left(y_{i_s}^{(\sigma)} x_{i_s}^- + y_{i_s+1}^{(\sigma)} x_{i_s}^+ \right) + \varphi_i^j, \quad (x, t) \in \omega_{h,\tau}, \end{aligned} \quad (48)$$

$$y(0, t) = 0, \quad t \in \bar{\omega}_\tau, \quad x = 0, \quad (49)$$

$$\begin{aligned} & - \left(\chi_N a_N y_{\bar{x},N}^{(\sigma)} + \Delta_{0t_{j+\sigma}}^\alpha (\gamma_N y_{\bar{x},N}) \right) = \beta y_N^{(\sigma)} \\ & + 0.5h \left(\Delta_{0t_{j+\sigma}}^\alpha y_N + d_N^j \left(y_{i_0}^{(\sigma)} x_{i_0}^- + y_{i_0+1}^{(\sigma)} x_{i_0}^+ \right) \right) - \tilde{\mu}, \quad x = l, \end{aligned} \quad (50)$$

$$y(x, 0) = u_0(x), \quad x \in \bar{\omega}_h, \quad (51)$$

where $\tilde{\mu}(t_{j+\sigma}) = \mu(t_{j+\sigma}) + 0.5h\varphi_N^j$.

Using the method of energy inequalities, we find an a priori estimate, for this we multiply Eq. (48) scalarly by y . Then, taking into account the transformations (24)–(29), after some transformations we obtain

$$\begin{aligned} & \left(\Delta_{0t_{j+\sigma}}^\alpha y, y^{(\sigma)} \right) + M_1 \|y_{\bar{x}}^{(\sigma)}\|_0^2 + \frac{C_0}{2} \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}}\|_0^2 \leq \varepsilon \|y_{\bar{x}}^{(\sigma)}\|_0^2 \\ & + \left(\chi_i^j a y_{\bar{x}}^{(\sigma)} + \Delta_{0t_{j+\sigma}}^\beta (\gamma_i y_{\bar{x}}) \right) y^{(\sigma)} \Big|_0^N + M_2(\varepsilon) \|y^{(\sigma)}\|_0^2 + \left(b_i^{-j} a_i^j y_{\bar{x},i}^{(\sigma)}, y^{(\sigma)} \right) \\ & + \left(b_i^{+j} a_{i+1}^j y_{x,i}^{(\sigma)}, y^{(\sigma)} \right) - \left(\sum_{s=1}^m d_{s,i}^j \left(y_{i_s}^{(\sigma)} x_{i_s}^- + y_{i_s+1}^{(\sigma)} x_{i_s}^+ \right), y^{(\sigma)} \right) + \left(\varphi, y^{(\sigma)} \right), \end{aligned} \quad (52)$$

where $(u, v] = \sum_{i=1}^N u_i v_i h$, $h = \begin{cases} 0.5h, & i = N; \\ h, & i \neq 0, N, \end{cases}$

$(u, u] = (1, u^2] = \|u\|_0^2$, $(u, v) = \sum_{i=1}^{N-1} u_i v_i h$.

We transform the first expression on the right-hand side of (52), then we get

$$\begin{aligned} & \left(\chi_i^j a y_{\bar{x}}^{(\sigma)} + \Delta_{0t_{j+\sigma}}^\beta (\gamma_i y_{\bar{x}}) \right) y^{(\sigma)} \Big|_0^N = \left(\chi_N^j a y_{\bar{x},N}^{(\sigma)} + \Delta_{0t_{j+\sigma}}^\beta (\gamma_N y_{\bar{x},N}) \right) y_N^{(\sigma)} \\ & = \left[\mu + 0.5h\varphi_N^j - \beta y_N^{(\sigma)} - 0.5h \left(\Delta_{0t_{j+\sigma}}^\alpha y_N - \sum_{s=1}^m d_{s,N} \left(y_{i_s}^{(\sigma)} x_{i_s}^- + y_{i_s+1}^{(\sigma)} x_{i_s}^+ \right) \right) \right] y_N^{(\sigma)} \\ & \leq -0.5h y_N^{(\sigma)} \Delta_{0t_{j+\sigma}}^\alpha y_N + M_2(\varepsilon) \|y^{(\sigma)}\|_0^2 + \varepsilon \|y_{\bar{x}}^{(\sigma)}\|_0^2 + M_3(\varepsilon) \mu^2 + 0.5h y_N^{(\sigma)} \varphi_N^j \\ & \quad + 0.5h y_N^{(\sigma)} \sum_{s=1}^m d_{s,N} \left(y_{i_s}^{(\sigma)} x_{i_s}^- + y_{i_s+1}^{(\sigma)} x_{i_s}^+ \right). \end{aligned} \quad (53)$$

Taking into account (53) with $\varepsilon = \frac{M_1}{2}$, from (52) we find

$$\begin{aligned} & \left(\Delta_{0t_{j+\sigma}}^\alpha y, y^{(\sigma)} \right) + \frac{M_1}{2} \|y_{\bar{x}}^{(\sigma)}\|_0^2 + \frac{c_0}{2} \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}}\|_0^2 \\ & \leq M_3 \|y^{(\sigma)}\|_0^2 + \left(b_i^{-j} a_i^j y_{\bar{x},i}^{(\sigma)}, y^{(\sigma)} \right) + \left(b_i^{+j} a_{i+1}^j y_{x,i}^{(\sigma)}, y^{(\sigma)} \right) \\ & - \left(\sum_{s=1}^m d_{s,i}^j \left(y_{i_s}^{(\sigma)} x_{i_s}^- + y_{i_s+1}^{(\sigma)} x_{i_s}^+ \right), y^{(\sigma)} \right) + \frac{1}{2} \mu^2 + \left(\varphi, y^{(\sigma)} \right). \end{aligned} \tag{54}$$

After simple transformations, taking into account the Cauchy inequality with ε , from (54) we obtain

$$\Delta_{0t_{j+\sigma}}^\alpha \|y\|_0^2 + \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}}\|_0^2 + \|y_{\bar{x}}^{(\sigma)}\|_0^2 \leq M_3 \|y^{(\sigma)}\|_0^2 + M_4 (\|\varphi\|_0^2 + \mu^2). \tag{55}$$

We rewrite (55) in another form

$$\Delta_{0t_{j+\sigma}}^\alpha \|y\|_0^2 + \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}}\|_0^2 \leq M_5^{\sigma} \|y^{j+1}\|_0^2 + M_6^{\sigma} \|y^j\|_0^2 + M_4 (\|\varphi\|_0^2 + \mu^2). \tag{56}$$

1) Consider the case when $\alpha > \beta$, then, based on Lemma 7 [11], from (56) we obtain

$$\|y\|_0^2 \leq M_7 \left(\|y^0\|_1^2 + \max_{0 \leq j' \leq j} (\|\varphi^{j'}\|_0^2 + \mu^2) \right), \tag{57}$$

where $\|y^0\|_1^2 = \|y^0\|_0^2 + \|y_{\bar{x}}^0\|_0^2$.

2) In the case when $\alpha \leq \beta$, from (56), taking into account (49) and the inequality $\|y^{(\sigma)}\|_0^2 \leq 2l^2 \|y_{\bar{x}}^{(\sigma)}\|_0^2$ [13], we obtain

$$\Delta_{0t_{j+\sigma}}^\alpha \|y\|_0^2 + \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}}\|_0^2 \leq M_8 \|y_{\bar{x}}^{(\sigma)}\|_0^2 + M_9 (\|\varphi\|_0^2 + \mu^2). \tag{58}$$

Based on Lemma 7 [11], from (58) we obtain

$$\|y_{\bar{x}}\|_0^2 \leq M_{10} \left(\|y^0\|_1^2 + \max_{0 \leq j' \leq j} (\|\varphi^{j'}\|_0^2 + \mu^2) \right), \tag{59}$$

where $M_7, M_{10} = const > 0$, independent of h and τ .

3 Results

Theorem 1. If $k(x, t) \in C^{1,0}(Q_T)$, $\eta(x) \in C^1[0, l]$, $r(x, t), q_s(x, t), f(x, t) \in C(Q_T)$, $u(x, t) \in C^{(2,0)}(Q_T) \cap C^{(1,0)}(\bar{Q}_T)$, $\partial_{0t}^\gamma u(x, t) \in C(Q_T)$, $\partial_{0t}^\gamma u_{xx}(x, t) \in C(\bar{Q}_T)$ and conditions (4) be satisfied, then estimates: (15) in the case when $\alpha > \beta$; (17) in the case when $\alpha = \beta$; (20) in the case when $\alpha < \beta$ are valid for the solution $u(x, t)$ of problem (1)–(3).

The a priori estimates (15), (17), (20) implies the uniqueness and stability of the solution with respect to the right-hand side and the initial data.

Theorem 2. Let conditions (4) be satisfied, then exist $\tau_0 = \tau_0(c_0, c_1, c_2, \alpha, \sigma)$, such that if $\tau \leq \tau_0$, then estimates: (34), in the case when $\alpha > \beta$; (35) in the case when $\alpha \leq \beta$ are valid for the solution of the difference problem (21)–(23).

The a priori estimates (34), (35) implies the uniqueness and stability of the solution to the difference scheme (21)–(23) with respect to the initial data and the right-hand side, as well as the convergence of the solution of the difference problem (21)–(23) to the solution of the differential problem (1)–(3) so that if there exist such τ_0 then for $\tau \leq \tau_0$ the a priori estimates are valid:

- 1) in case when $\alpha > \beta$: $\|y^{j+1} - u^{j+1}\|_0^2 \leq M (h^2 + \tau^{2-\max\{\alpha, \beta\}})$;
- 2) in case when $\alpha = \beta$: $\|y_{\bar{x}}^{j+1} - u_{\bar{x}}^{j+1}\|_0^2 \leq M (h^2 + \tau^2)$;
- 3) in case when $\alpha < \beta$: $\|y_{\bar{x}}^{j+1} - u_{\bar{x}}^{j+1}\|_0^2 \leq M (h^2 + \tau^{2-\max\{\alpha, \beta\}})$,

where $M - const > 0$, independent of h and τ .

Theorem 3. If $k(x, t) \in C^{1,0}(Q_T)$, $\eta(x) \in C^1[0, l]$, $r(x, t), q_s(x, t), f(x, t) \in C(Q_T)$, $u(x, t) \in C^{(2,0)}(Q_T) \cap C^{(1,0)}(\bar{Q}_T)$, $\partial_{0t}^\gamma u(x, t) \in C(Q_T)$, $\partial_{0t}^\gamma u_{xx}(x, t) \in C(\bar{Q}_T)$ and conditions (4), (37) be satisfied, then estimates: (42) in the case when $\alpha > \beta$; (44) in the case when $\alpha = \beta$; (47) in the case when $\alpha < \beta$ are valid for the solution $u(x, t)$ of problem (1), (36), (3).

The a priori estimates (42), (44), (47) implies the uniqueness and stability of the solution with respect to the right-hand side and the initial data

Theorem 4. Let conditions (4), (37), be satisfied, then exist $\tau_0 = \tau_0(c_0, c_1, c_2, \alpha, \sigma)$, such that if $\tau \leq \tau_0$, then estimates: (57), in the case when $\alpha > \beta$; (59) in the case when $\alpha \leq \beta$ are valid for the solution of the difference problem (48)–(51).

The a priori estimates (57), (59) implies the uniqueness and stability of the solution to the difference scheme (48)–(51) with respect to the initial data and the right-hand side, as well as the convergence of the solution of the difference problem (48)–(51) to the solution of the differential problem (1), (36), (3) so that if there exist such τ_0 then for $\tau \leq \tau_0$ the a priori estimates are valid:

- 1) in case when $\alpha > \beta$: $\|y^{j+1} - u^{j+1}\|_0^2 \leq M (h^2 + \tau^{2-\max\{\alpha, \beta\}})$;
- 2) in case when $\alpha = \beta$: $\|y_{\bar{x}}^{j+1} - u_{\bar{x}}^{j+1}\|_0^2 \leq M (h^2 + \tau^2)$;
- 3) in case when $\alpha < \beta$: $\|y_{\bar{x}}^{j+1} - u_{\bar{x}}^{j+1}\|_0^2 \leq M (h^2 + \tau^{2-\max\{\alpha, \beta\}})$,

where $M - const > 0$, independent of h and τ .

Comment. To solve numerically the boundary value problems considered in this paper for the Hallaire equation with two fractional Gerasimov-Caputo derivatives of different orders α, β , difference schemes (21)–(23) and (48)–(51) are reduced to calculated form; for this one can use the method of parametric sweep [16].

4 Conclusion

In this work boundary-value problems for the loaded Hallaire equation with variable coefficients and Gerasimov-Caputo fractional derivatives of different orders are studied. Using the method of energy inequalities for various relations between α and β , a priori estimates in differential and difference interpretations are obtained for the solution of the problem under consideration, from which the uniqueness and stability of the solution with respect to the initial data and the right-hand side, as well as the convergence of the solution of the difference problem to the solution of the differential problem.

References

1. Kochubei, A.N.: Diffusion of fractional order. *Differ. Eq.* **26**(4), 485–492 (1990)
2. Nigmatullin, R.R.: Fractional integral and its physical interpretation. *Theor. Math. Phys.* **90**(3), 242–251 (1992)
3. Goloviznin, V.M., Kiselev, V.P., Korotkij, I.A., Yurkov, Y.I.: Some features of computing algorithms for the equations fractional diffusion. Preprint. IBRAE-2002-01. Moscow: Nuclear Safety Institute RAS (2002)
4. Barenblat, G.I., Zheltov, Y.P., Kochina, I.N.: On the main ideas of the theory of filtration of homogeneous fluids in fractured rocks. *Appl. Math. Mech.* **25**(5), 852–864 (1960)
5. Hallaire, M.: Le potentiel efficace de l'eau dans le sol en regime de dessechement, L'Eau et la Production Vegetale. Paris: Institut National de la Recherche Agronomique. **9**, 27–62 (1964)
6. Chudnovsky, A.F.: *Thermal Physics of Soils*. Nauka, Moscow (1976)
7. Alikhanov, A.A.: A priori estimates for solutions of boundary value problems for fractional-order equations. *Differ. Eq.* **46**(5), 660–666 (2010)
8. Alikhanov, A.A.: A new difference scheme for the time fractional diffusion equation. *J. Comput. Phys.* **280**, 424–438 (2015)
9. Alikhanov, A.A. Berezgov, A.M. and Shkhanukov-Lafishev, M.KH.: Boundary value problems for certain classes of loaded differential equations and solving them by finite difference methods. *Comput. Math. Math. Phys.* **48**(9), 1581–1590 (2008)
10. Beshtokov, M.Kh.: The third boundary value problem for loaded differential Sobolev type equation and grid methods of their numerical implementation. *IOP Conf. Ser.: Mater. Sci. Eng.* **158**(1) (2016)
11. Beshtokov, M.K.: Numerical analysis of initial-boundary value problem for a Sobolev-type equation with a fractional-order time derivative. *Comput. Math. Math. Phys.* **59**(2), 175–192 (2019)
12. Beshtokov, M.K., Vodakhova, V.A.: Nonlocal boundary value problems for a fractional-order convection-diffusion equation, *Vestnik Udmurtskogo Universiteta. Matematika. Mekhanika. Komp'yuternye Nauki* **29**(4), 459–482 (2019)
13. Samarskiy, A.A.: *Theory of Difference Schemes*. Nauka, Moscow (1983)
14. Ladyzhenskaya, O.A.: *Boundary Value Problems of Mathematical Physics*. Nauka, Moscow (1973)
15. Samarskii, A.A., Gulin, A.V.: *Stability of Difference Schemes*. Nauka, Moscow (1973)
16. Voevodin, A.F., Shugrin, S.M.: *Numerical Methods for Calculating One-Dimensional Systems*, SO AS USSR. Nauka, Novosibirsk (1981)



Grid Method for Solving Local and Nonlocal Boundary Value Problems for a Loaded Moisture Transfer Equation with Two Fractional Differentiation Operators

Murat Beshtokov^(✉)

Institute of Applied Mathematics and Automation, Kabardino-Balkarian Scientific Center of RAS, 89A Shortanova Str., 360000 Nalchik, Russia
beshtokov-murat@yandex.ru

Abstract. This paper is devoted to the study of local and nonlocal boundary value problems for a loaded moisture transfer equation with two fractional Gerasimov-Caputo derivatives of different orders α , β . Using the method of energy inequalities for various relations between α and β , a priori estimates in differential and difference interpretations are obtained for solving the problems under consideration, which implies the uniqueness and stability of the solution with respect to the initial data and the right-hand side, as well as the convergence of the solution of the difference problem to the solution of the differential problem with the rate $O(h^2 + \tau^2)$ for $\alpha = \beta$ and $O(h^2 + \tau^{2-\max\{\alpha, \beta\}})$ for $\alpha \neq \beta$.

Keywords: loaded equation · moisture transfer equation · nonlocal boundary value problem · fractional Caputo derivative · a priori estimate · difference scheme · stability and convergence

1 Introduction

It is well known that the issues of fluid filtration in porous media [1, 2], heat transfer in a heterogeneous medium [3, 4], moisture transfer in soils [5], (see [6, c.137]) lead to modified diffusion equations, which are called pseudo-parabolic equations or Sobolev-type equations [7].

In [8], mathematical models of the water regime in soils with a fractal structure are proposed and investigated. These models are based on pseudo-parabolic equations with a fractional derivative in time.

Loaded differential equations in the literature (see [9–11]) are usually called equations containing in the coefficients or on the right side some functionals of the solution, in particular, the values of the solution or its derivatives on manifolds of smaller dimension. The study of such equations is of interest both from the point of view of constructing a general theory of differential equations, and

from the point of view of applications, moreover, applications both in mathematical modeling and in mathematics proper.

The nonclassical nature of the problems considered in this paper lies in the fact that instead of the first derivative in time, the equation contains a derivative of fractional order in the sense of Gerasimov-Caputo, while the equations are loaded. These non-classical directions are relevant due to the abundance of various applications in which such non-classical objects arise.

Papers [12–15] are devoted to numerical methods for solving boundary value problems for various diffusion equations. In [14, 15], results were obtained that allow, as in the classical case ($\alpha = 1$), to apply the method of energy inequalities to find a priori estimates of boundary value problems for a fractional order equation in differential and difference interpretations.

The works of the author [16, 17] are devoted to non-local boundary value problems for various moisture transfer equations, and to loaded equations - [18].

2 Materials and Methods

2.1 Problem Statement

In a closed rectangle $\overline{Q}_T = \{(x, t) : 0 \leq x \leq l, 0 \leq t \leq T\}$ consider the first boundary value problem for the loaded moisture transfer equation with two Gerasimov–Caputo fractional differentiation operators of different orders α and β

$$\partial_{0t}^\alpha u = \frac{\partial}{\partial x} \left(k(x) \frac{\partial u}{\partial x} \right) + \eta \partial_{0t}^\beta \frac{\partial^2 u}{\partial x^2} + \sum_{s=1}^m r_s(x, t) \frac{\partial u}{\partial x}(x_s, t) - q(x, t)u + f(x, t),$$

$$0 < x < l, \quad 0 < t \leq T, \tag{1}$$

$$u(0, t) = u(l, t) = 0, \quad 0 \leq t \leq T, \tag{2}$$

$$u(x, 0) = u_0(x), \quad 0 \leq x \leq l, \tag{3}$$

where

$$0 < c_0 \leq k(x) \leq c_1, \quad |k_x(x)|, |r_s(x, t)|, |q(x, t)| \leq c_2, \eta = const > 0, \tag{4}$$

$\partial_{0t}^\gamma u = \frac{1}{\Gamma(1-\gamma)} \int_0^t \frac{u_\tau(x, \tau)}{(t-\tau)^\gamma} d\tau$ is a fractional derivative in the sense of Gerasimov–Caputo order γ , $0 < \gamma < 1$, x_s ($s = 1, 2, \dots, m$) are arbitrary interval points $[0, l]$.

We will assume that the problem (1)–(3) has a unique solution possessing the derivatives required in the course of the presentation, the coefficients of the equation and boundary conditions satisfy the smoothness conditions necessary in the course of the presentation, which ensure the required order of approximation of the difference scheme.

We will also use positive constants $M_i, (i = 1, 2, \dots,)$ depending only on the input data of the original problem.

2.2 A Priori Estimate in Differential Form

To obtain an a priori estimate for the solution of problem (1)–(3) in differential form, we multiply Eq. (1) scalarly by $U = \partial_{0t}^\alpha u + \partial_{0t}^\beta u - u_{xx}$:

$$\begin{aligned} (\partial_{0t}^\alpha u, U) &= ((ku_x)_x, U) + \left(\eta \partial_{0t}^\beta u_{xx}, U \right) + \\ &+ \left(\sum_{s=1}^m r_s u_x(x_s, t), U \right) - (qu, U) + (f, U). \end{aligned} \quad (5)$$

where $(u, v) = \int_0^l uv dx$, $(u, u) = \|u\|_0^2$, where u, v are functions defined on $[0, l]$.

We transform the integrals in the identity (5), using (2), the Cauchy inequality with ε ([13], p. 100, [15]) and Lemma 1 from [14], we obtain:

$$\begin{aligned} &\frac{1}{2} \partial_{0t}^\alpha \|u_x\|_0^2 + \frac{\eta}{2} \partial_{0t}^\beta \|u_{xx}\|_0^2 + \frac{c_0}{2} \partial_{0t}^\alpha \|u_x\|_0^2 + \frac{c_0}{2} \partial_{0t}^\beta \|u_x\|_0^2 + \|\partial_{0t}^\alpha u\|_0^2 + \\ &+ \eta \|\partial_{0t}^\beta u_x\|_0^2 + \left(\partial_{0t}^\alpha u, \partial_{0t}^\beta u \right) + \eta \left(\partial_{0t}^\alpha u_x, \partial_{0t}^\beta u_x \right) + c_0 \|u_{xx}\|_0^2 \leq \varepsilon_1 M_1 \|\partial_{0t}^\alpha u\|_0^2 + \\ &+ \varepsilon_2 M_2 \|\partial_{0t}^\beta u\|_0^2 + \varepsilon_3 M_3 \|u_{xx}\|_0^2 + M_4^{\varepsilon_1, \varepsilon_2, \varepsilon_3} \|u_x\|_0^2 + M_5^{\varepsilon_1, \varepsilon_2, \varepsilon_3} \|f\|_0^2. \end{aligned} \quad (6)$$

Now we will show that $(\partial_{0t}^\alpha u, \partial_{0t}^\beta u) \geq 0$, then we have

$$\begin{aligned} (\partial_{0t}^\alpha u, \partial_{0t}^\beta u) &= \int_0^l \left(\frac{1}{\Gamma(1-\alpha)} \int_0^t \frac{u_\tau(x, \tau)}{(t-\tau)^\alpha} d\tau \frac{1}{\Gamma(1-\beta)} \int_0^t \frac{u_\tau(x, \tau)}{(t-\tau)^\beta} d\tau \right) dx = \\ &= \frac{1}{\Gamma(1-\alpha)\Gamma(1-\beta)} \int_0^l \left(\int_0^t \frac{u_\tau(x, \tau)}{(t-\tau)^\alpha} d\tau \int_0^t \frac{u_\tau(x, \tau)}{(t-\tau)^\beta} d\tau \right) dx \geq \\ &\geq \frac{1}{\Gamma(1-\alpha)\Gamma(1-\beta)} \int_0^l \left(\int_0^t \frac{u_\tau(x, \tau)}{(t-\tau)^{\frac{\alpha+\beta}{2}}} d\tau \right)^2 dx \geq 0. \end{aligned}$$

Taking into account due to (2) that $\|u\|_0^2 \leq 2l^2 \|u_x\|_0^2$ and choosing $\varepsilon_1 = \frac{1}{2M_1}$, $\varepsilon_2 = \frac{\eta}{2M_2}$, $\varepsilon_3 = \frac{c_0}{2M_3}$, from (6) we find

$$\begin{aligned} \partial_{0t}^\alpha \|u_x\|_0^2 + \partial_{0t}^\beta (\|u_x\|_0^2 + \|u_{xx}\|_0^2) + \|u_{xx}\|_0^2 + \|\partial_{0t}^\alpha u\|_0^2 + \|\partial_{0t}^\beta u_x\|_0^2 \leq \\ \leq M_6 \|u_x\|_0^2 + M_7 \|f\|_0^2, \end{aligned} \quad (7)$$

1) Let $\alpha > \beta$, then applying the fractional integration operator $D_{0t}^{-\alpha}$ to both sides of the inequality (7), we obtain

$$\begin{aligned} \|u_x\|_0^2 + D_{0t}^{-(\alpha-\beta)} (\|u_x\|_0^2 + \|u_{xx}\|_0^2) + D_{0t}^{-\alpha} (\|u_{xx}\|_0^2 + \|\partial_{0t}^\alpha u\|_0^2 + \|\partial_{0t}^\beta u_x\|_0^2) \leq \\ \leq M_6 D_{0t}^{-\alpha} \|u_x\|_0^2 + M_8 (D_{0t}^{-\alpha} \|f\|_0^2 + \|u'_0(x)\|_0^2), \end{aligned} \quad (8)$$

where $D_{0t}^{-\gamma} u = \frac{1}{\Gamma(\gamma)} \int_0^t \frac{u d\tau}{(t-\tau)^{1-\gamma}}$ is a fractional derivative in the sense of Riemann-Liouville of order $\gamma, 0 < \gamma < 1$.

Based on Lemma 2 [14], from (8) we find the following a priori estimate

$$\|u\|_1^2 \leq M_8 (D_{0t}^{-\alpha} \|f\|_0^2 + \|u_0(x)\|_0^2), \tag{9}$$

where $\|u\|_1^2 = \|u_x\|_0^2 + D_{0t}^{-(\alpha-\beta)} \|u_{xx}\|_0^2 + D_{0t}^{-\beta} (\|\partial_{0t}^\alpha u\|_0^2 + \|\partial_{0t}^\beta u_x\|_0^2)$, $M_8 = const > 0$, dependent only on input data (1)–(3).

2) Let $\alpha \leq \beta$, then from (7) we get

$$\begin{aligned} \|u_x\|_0^2 + \|u_{xx}\|_0^2 + D_{0t}^{-(\beta-\alpha)} \|u_x\|_0^2 + D_{0t}^{-\beta} (\|u_{xx}\|_0^2 + \|\partial_{0t}^\alpha u\|_0^2 + \|\partial_{0t}^\beta u_x\|_0^2) \leq \\ \leq M_5 D_{0t}^{-\beta} \|u_x\|_0^2 + M_7 (D_{0t}^{-\beta} \|f\|_0^2 + \|u'_0(x)\|_0^2 + \|u''_0(x)\|_0^2). \end{aligned} \tag{10}$$

Based on Lemma 2 [14], from (10) we find the following a priori estimate

$$\|u\|_2^2 \leq M_9 (D_{0t}^{-\beta} \|f\|_0^2 + \|u'_0(x)\|_0^2 + \|u''_0(x)\|_0^2), \tag{11}$$

where $\|u\|_2^2 = \|u_x\|_0^2 + \|u_{xx}\|_0^2 + D_{0t}^{-\beta} (\|\partial_{0t}^\alpha u\|_0^2 + \|\partial_{0t}^\beta u_x\|_0^2)$, $M_9 = const > 0$, dependent only on input data (1)–(3).

2.3 Stability and Convergence of the Difference Scheme

For solution of problem (1)–(3) we use the finite difference method. In the closed rectangle \overline{Q}_T we introduce a uniform grid $\overline{w}_{h\tau} = \overline{w}_h \times \overline{w}_\tau$, where $\overline{w}_h = \{x_i = ih, i = \overline{0}, \overline{N}, h = l/N\}$, $\overline{w}_\tau = \{t_j = j\tau, j = 0, 1, \dots, j_0, \tau = T/j_0\}$. On the uniform grid $\overline{w}_{h\tau}$, we associate the differential problem (1)–(3) with the difference scheme of the order of approximation $O(h^2 + \tau^2)$ for $\alpha = \beta$ and $O(h^2 + \tau^{2-\max\{\alpha,\beta\}})$ for $\alpha \neq \beta$:

$$\Delta_{0t_{j+\sigma}}^\alpha y_i = \left(ay_{\overline{x}}^{(\sigma)} \right)_{x,i} + \eta \Delta_{0t_{j+\sigma}}^\beta y_{\overline{x},i} + \sum_{s=1}^m r_{s,i}^j \left(y_{\overline{x},i_s}^{(\sigma)} x_{i_s}^- + y_{\overline{x},i_s+1}^{(\sigma)} x_{i_s}^+ \right) - d_i^j y_i^{(\sigma)} + \varphi_i^j, \tag{12}$$

$$y_0^{(\sigma)} = y_N^{(\sigma)} = 0, \tag{13}$$

$$y(x, 0) = u_0(x), \tag{14}$$

where $\Delta_{0t_{j+\sigma}}^\gamma y = \frac{\tau^{1-\gamma}}{\Gamma(2-\gamma)} \sum_{s=0}^j c_{j-s}^{(\gamma,\sigma)} y_s^s$ is a discrete analogue of the fractional Gerasimova-Caputo derivative of order $\gamma, 0 < \gamma < 1$, which assures order of accuracy $O(\tau^{3-\gamma})$ for $\sigma = 1 - \frac{\gamma}{2}$, and $O(\tau^{2-\gamma})$ for $\sigma = 0.5$ [15].

Where

$$a_0^{(\gamma,\sigma)} = \sigma^{1-\gamma}, \quad a_l^{(\gamma,\sigma)} = (l + \sigma)^{1-\gamma} - (l - 1 + \sigma)^{1-\gamma}, \quad l \geq 1,$$

$$b_l^{(\gamma,\sigma)} = \frac{1}{2-\gamma} [(l+\sigma)^{2-\gamma} - (l-1+\sigma)^{2-\gamma}] - \frac{1}{2} [(l+\sigma)^{1-\gamma} + (l-1+\sigma)^{1-\gamma}], \quad l \geq 1,$$

if $j = 0$, then $c_0^{(\gamma,\sigma)} = a_0^{(\gamma,\sigma)}$;

$$\text{if } j > 0, \text{ then } c_s^{(\gamma,\sigma)} = \begin{cases} a_0^{(\gamma,\sigma)} + b_1^{(\gamma,\sigma)}, & s = 0; \\ a_s^{(\gamma,\sigma)} + b_{s+1}^{(\gamma,\sigma)} - b_s^{(\gamma,\sigma)}, & 1 \leq s \leq j-1; \\ a_j^{(\gamma,\sigma)} - b_j^{(\gamma,\sigma)}, & s = j, \end{cases}$$

$\sigma = 1 - \frac{\gamma}{2}$, for $\alpha = \beta$ and $\sigma = 0.5$, for $\alpha \neq \beta$,

$$c_s^{(\gamma,\sigma)} > \frac{1-\gamma}{2}(s+\sigma)^{-\gamma} > 0; \quad y^{(\sigma)} = \sigma y^{j+1} + (1-\sigma)y^j; \quad a_i = k(x_{i-0.5}),$$

$$r_{s,i}^j = r_s(x_i, t^{j+\sigma}), \quad d_i^j = d(x_i, t^{j+\sigma}), \quad \varphi_i^j = f(x_i, t^{j+\sigma}),$$

$$x_{i_s} \leq x_s \leq x_{i_s+1}, \quad x_{i_s}^- = \frac{x_{i_s+1} - x_s}{h}, \quad x_{i_s}^+ = \frac{x_s - x_{i_s}}{h} \quad c_{s-1}^{(\gamma,\sigma)} > c_s^{(\gamma,\sigma)}.$$

Now let us find the a priori estimate by the method of energy inequalities, for this purpose, we multiply (12) scalarly by $\bar{y} = \Delta_{0t_{j+\sigma}}^\alpha y + \Delta_{0t_{j+\sigma}}^\beta y - y_{\bar{x}\bar{x}}^{(\sigma)}$:

$$\begin{aligned} & \left(\Delta_{0t_{j+\sigma}}^\alpha y, \bar{y} \right) = \left(\left(ay_{\bar{x}}^{(\sigma)} \right)_x, \bar{y} \right) + \left(\eta \Delta_{0t_{j+\sigma}}^\beta y_{\bar{x}\bar{x}}, \bar{y} \right) + \\ & + \left(\sum_{s=1}^m r_{s,i}^j \left(y_{\bar{x},i_s}^{(\sigma)} x_{i_s}^- + y_{\bar{x},i_s+1}^{(\sigma)} x_{i_s}^+ \right), \bar{y} \right) - \left(dy^{(\sigma)}, \bar{y} \right) + (\varphi, \bar{y}). \end{aligned} \quad (15)$$

From (15), taking into account (13) and Lemma 1 [15], we find

$$\begin{aligned} & \left(\frac{1}{2} + \frac{c_0}{2} \right) \Delta_{0t_{j+\sigma}}^\alpha \|y_{\bar{x}}\|_0^2 + \frac{\eta}{2} \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}\bar{x}}\|_0^2 + \frac{c_0}{2} \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}}\|_0^2 + \|\Delta_{0t_{j+\sigma}}^\alpha y\|_0^2 \\ & + \eta \|\Delta_{0t_{j+\sigma}}^\beta y_{\bar{x}\bar{x}}\|_0^2 + \left(\Delta_{0t_{j+\sigma}}^\alpha y, \Delta_{0t_{j+\sigma}}^\beta y \right) + \eta \left(\Delta_{0t_{j+\sigma}}^\alpha y_{\bar{x}}, \Delta_{0t_{j+\sigma}}^\alpha y_{\bar{x}} \right) \\ & + c_0 \|y_{\bar{x}\bar{x}}^{(\sigma)}\|_0^2 \leq \varepsilon_1 M_1 \|\Delta_{0t_{j+\sigma}}^\alpha y\|_0^2 + \varepsilon_2 M_2 \|\Delta_{0t_{j+\sigma}}^\beta y\|_0^2 + \varepsilon_3 M_3 \|y_{\bar{x}\bar{x}}^{(\sigma)}\|_0^2 \\ & + M_4^{\varepsilon_1, \varepsilon_2, \varepsilon_3} \|y_{\bar{x}}^{(\sigma)}\|_0^2 + M_5^{\varepsilon_1, \varepsilon_2, \varepsilon_3} \|\varphi\|_0^2. \end{aligned} \quad (16)$$

By virtue of the condition (13), that the inequality $\|y^\sigma\|_0^2 \leq 2l^2 \|y_{\bar{x}}^\sigma\|_0^2$ is valid, and choosing $\varepsilon_1 = \frac{1}{2M_1}$, $\varepsilon_2 = \frac{\eta}{2M_2}$, $\varepsilon_3 = \frac{c_0}{2M_3}$, from (16) we get

$$\begin{aligned} & \Delta_{0t_{j+\sigma}}^\alpha \|y_{\bar{x}}\|_0^2 + \Delta_{0t_{j+\sigma}}^\beta (\|y_{\bar{x}}\|_0^2 + \|y_{\bar{x}\bar{x}}\|_0^2) + \|\Delta_{0t_{j+\sigma}}^\alpha y\|_0^2 + \|\Delta_{0t_{j+\sigma}}^\beta y_{\bar{x}\bar{x}}\|_0^2 \\ & + \|y_{\bar{x}\bar{x}}^{(\sigma)}\|_0^2 \leq M_6 \|y_{\bar{x}}^{(\sigma)}\|_0^2 + M_7 \|\varphi\|_0^2, \end{aligned} \quad (17)$$

We rewrite (17) in another form

$$\begin{aligned} & \Delta_{0t_{j+\sigma}}^\alpha \|y_{\bar{x}}\|_0^2 + \Delta_{0t_{j+\sigma}}^\beta (\|y_{\bar{x}}\|_0^2 + \|y_{\bar{x}\bar{x}}\|_0^2) \\ & \leq M_8^\sigma \|y_{\bar{x}}^{j+1}\|_0^2 + M_9^\sigma \|y_{\bar{x}}^j\|_0^2 + M_{10} \|\varphi\|_0^2. \end{aligned} \quad (18)$$

Based on Lemma 7 [17], from (18) we obtain

1) In case when $\alpha > \beta$:

$$\|y_{\bar{x}}^{j+1}\|_0^2 \leq M_{11} \left(\|y^0\|_1^2 + \max_{0 \leq j' \leq j} \|\varphi^{j'}\|_0^2 \right), \tag{19}$$

2) In case when $\alpha \leq \beta$:

$$\|y^{j+1}\|_1^2 \leq M_{12} \left(\|y^0\|_1^2 + \max_{0 \leq j' \leq j} \|\varphi^{j'}\|_0^2 \right), \tag{20}$$

where $M_{11}, M_{12} = const > 0$, independent of h and τ , $\|y\|_1^2 = \|y_{\bar{x}}\|_0^2 + \|y_{\bar{x}x}\|_0^2$.

2.4 Statement of the Non-local Boundary Value Problem and a Priori Estimate in Differential Form

Consider instead of the second condition in (2) a non-local condition, then instead of the condition (2) we will consider

$$\begin{cases} u(0, t) = 0, \\ -\Pi(l, t) = \beta_1(t)u(l, t) + \beta_2(t)\partial_{0t}^\alpha u(l, t) + \beta_3(t)\partial_{0t}^\beta u(l, t) - \mu(t), \end{cases} \tag{21}$$

where

$$\eta = const > 0, \quad 0 < c_0 \leq k, \beta_2, \beta_3 \leq c_1, \quad |\beta_1|, |r|, |q|, |k_x| \leq c_2, \tag{22}$$

$$\Pi(x, t) = ku_x + \eta\partial_{0t}^\beta u_x,$$

We multiply the Eq. (1) scalarly by $U = u + \partial_{0t}^\alpha u + \partial_{0t}^\beta u - u_{xx}$

$$\begin{aligned} (\partial_{0t}^\alpha u, U) &= ((ku_x)_x, U) + (\eta\partial_{0t}^\beta u_{xx}, U) \\ &+ \left(\sum_{s=1}^m r_s u_x(x_s, t), U \right) - (qu, U) + (f, U). \end{aligned} \tag{23}$$

By virtue of the condition (21), that the inequalities $\|u\| \leq 2l^2\|u_x\|$, $\partial_{0t}^\alpha u \partial_{0t}^\beta u > 0$ are valid, from (23) after some simple transformations, we obtain the inequality

$$\begin{aligned} &\frac{1}{2}\partial_{0t}^\alpha \|u\|_0^2 + \left(\frac{c_0 + 1}{2}\right)\partial_{0t}^\alpha \|u_x\|_0^2 + \left(\frac{c_0 + \eta}{2}\right)\partial_{0t}^\beta \|u_x\|_0^2 + \left(\frac{\eta}{2}\right)\partial_{0t}^\beta \|u_{xx}\|_0^2 \\ &+ \|\partial_{0t}^\alpha u\|_0^2 + \eta\|\partial_{0t}^\beta u_x\|_0^2 + c_0 (\|u_x\|_0^2 + \|u_{xx}\|_0^2) \leq u\Pi(x, t)|_{x=l} \\ &+ (\Pi(x, t) + u_x) \partial_{0t}^\alpha u|_{x=l} + \Pi(x, t)\partial_{0t}^\beta u|_{x=l} + M_2^\varepsilon \|u_x\|_0^2 + M_3^\varepsilon \|f\|_0^2. \end{aligned} \tag{24}$$

We transform the first, second and third expressions on the right-hand side of (24), then we get

$$\begin{aligned}
& \left[u\Pi(x, t) + (\Pi(x, t) + u_x) \partial_{0t}^\alpha u + \Pi(x, t) \partial_{0t}^\beta u \right]_{x=l} \\
&= \left(\mu(t) - \beta_1(t)u(l, t) - \beta_2(t)\partial_{0t}^\alpha u(l, t) - \beta_3(t)\partial_{0t}^\beta u(l, t) \right) u(l, t) \\
&+ \left(u_x(l, t) + \mu(t) - \beta_1(t)u(l, t) - \beta_2(t)\partial_{0t}^\alpha u(l, t) - \beta_3(t)\partial_{0t}^\beta u(l, t) \right) \partial_{0t}^\alpha u(l, t) \\
&+ \left(\mu(t) - \beta_1(t)u(l, t) - \beta_2(t)\partial_{0t}^\alpha u(l, t) - \beta_3(t)\partial_{0t}^\beta u(l, t) \right) \partial_{0t}^\beta u(l, t) \\
&\leq -\frac{\beta_2}{2} (\partial_{0t}^\alpha u(l, t))^2 - \frac{\beta_2}{2} \partial_{0t}^\alpha u^2(l, t) - \frac{\beta_3}{2} \left(\partial_{0t}^\beta u(l, t) \right)^2 - \frac{\beta_3}{2} \partial_{0t}^\beta u^2(l, t) \\
&+ \beta_2 \partial_{0t}^\alpha u(l, t) \partial_{0t}^\beta u(l, t) + \varepsilon M_2 \|u_{xx}\|_0^2 + M_6 \|u_x\|_0^2 + M_7 \mu^2(t). \tag{25}
\end{aligned}$$

Considering transformations (25), from (24) with $\varepsilon = \frac{c_0}{2}$ find

$$\begin{aligned}
& \partial_{0t}^\alpha (\|u\|_0^2 + \|u_x\|_0^2) + \partial_{0t}^\beta (\|u_x\|_0^2 + \|u_{xx}\|_0^2) + \|\partial_{0t}^\alpha u\|_0^2 + \|\partial_{0t}^\beta u_x\|_0^2 \\
&+ c_0 (\|u_x\|_0^2 + \|u_{xx}\|_0^2) \leq M_2 \|u_x\|_0^2 + M_3 (\|f\|_0^2 + \mu^2(t)). \tag{26}
\end{aligned}$$

1) Let $\alpha > \beta$, then applying the fractional integration operator $D_{0t}^{-\alpha}$ to both sides of the inequality (26), based on Lemma 2 [14], we obtain the a priori estimate

$$\|u\|_3^2 \leq M_{10} (D_{0t}^{-\alpha} (\|f\|_0^2 + \mu^2) + \|u_0(x)\|_0^2 + \|u'_0(x)\|_0^2), \tag{27}$$

where $\|u\|_3^2 = \|u\|_0^2 + \|u_x\|_0^2 + D_{0t}^{-(\alpha-\beta)} \|u_{xx}\|_0^2 + D_{0t}^{-\alpha} (\|\partial_{0t}^\alpha u\|_0^2 + \|\partial_{0t}^\beta u_x\|_0^2)$, $M_{10} = \text{const} > 0$, dependent only on input data (1), (21), (3).

2) Let $\alpha = \beta$, then applying the fractional integration operator $D_{0t}^{-\alpha}$ to both sides of the inequality (26), based on Lemma 2 [14], we obtain the a priori estimate

$$\|u\|_4^2 \leq M_{11} (D_{0t}^{-\alpha} (\|f\|_0^2 + \mu^2) + \|u_0(x)\|_0^2 + \|u'_0(x)\|_0^2 + \|u''_0(x)\|_0^2), \tag{28}$$

where $\|u\|_4^2 = \|u\|_0^2 + \|u_x\|_0^2 + \|u_{xx}\|_0^2 + D_{0t}^{-\alpha} (\|\partial_{0t}^\alpha u\|_0^2 + \|\partial_{0t}^\alpha u_x\|_0^2)$, $M_{11} = \text{const} > 0$, dependent only on input data (1), (21), (3),

3) Let $\alpha < \beta$, then applying the fractional integration operator $D_{0t}^{-\beta}$ to both sides of the inequality (26), based on Lemma 2 [14], we obtain the a priori estimate

$$\|u\|_5^2 \leq M_{12} (D_{0t}^{-\alpha} (\|f\|_0^2 + \mu^2) + \|u'_0(x)\|_0^2 + \|u''_0(x)\|_0^2), \tag{29}$$

where $\|u\|_5^2 = \|u_x\|_0^2 + \|u_{xx}\|_0^2 + D_{0t}^{-(\beta-\alpha)} \|u\|_0^2 + D_{0t}^{-\beta} (\|\partial_{0t}^\alpha u\|_0^2 + \|\partial_{0t}^\alpha u_x\|_0^2)$, $M_{12} = \text{const} > 0$, dependent only on input data (1), (21), (3),

2.5 Stability and Convergence of the Difference Scheme

On the uniform grid $\bar{\omega}_{h\tau}$, we associate the differential problem (1), (21), (3) with the difference scheme of the order of approximation $O(h^2 + \tau^2)$ for $\alpha = \beta$ and $O(h^2 + \tau^{2-\max\{\alpha,\beta\}})$ for $\alpha \neq \beta$:

$$\Delta_{0t_{j+\sigma}}^\alpha y_i = \left(a^j y_{\bar{x}}^{(\sigma)} \right)_{x,i} + \eta \Delta_{0t_{j+\sigma}}^\alpha y_{\bar{x}x} + \sum_{s=1}^m r_{s,i}^j \left(y_{\bar{x},i_s}^{(\sigma)} x_{i_s}^- + y_{\bar{x},i_s+1}^{(\sigma)} x_{i_s}^+ \right) - d_i^j y_i^{(\sigma)} + \varphi_i^j, \tag{30}$$

$$y(0, t) = 0, \quad t \in \bar{\omega}_\tau, \quad x = 0, \tag{31}$$

$$\begin{aligned} & - \left(a_N y_{\bar{x},N}^{(\sigma)} - \Delta_{0t_{j+\sigma}}^\alpha y_{\bar{x},N} \right) + 0.5h \sum_{s=1}^m r_{s,N}^j \left(y_{\bar{x},i_s}^{(\sigma)} x_{i_s}^- + y_{\bar{x},i_s+1}^{(\sigma)} x_{i_s}^+ \right) \\ & = \tilde{\beta}_1 y_N^{(\sigma)} + \tilde{\beta}_2 \Delta_{0t_{j+\sigma}}^\alpha y_N + \beta_3 \Delta_{0t_{j+\sigma}}^\beta y_N - \tilde{\mu}, \quad t \in \bar{\omega}_\tau, \quad x = l, \end{aligned} \tag{32}$$

$$y(x, 0) = u_0(x), \quad x \in \bar{\omega}_h, \tag{33}$$

where $\tilde{\beta}_1(t_{j+\sigma}) = \beta_1(t_{j+\sigma}) + 0.5hd_N^j$, $\tilde{\beta}_2(t_{j+\sigma}) = \beta_2(t_{j+\sigma}) + 0.5h$, $\tilde{\mu}(t_{j+\sigma}) = \mu(t_{j+\sigma}) + 0.5h\varphi_N^j$.

Using the method of energy inequalities, we find an a priori estimate, for this we multiply Eq. (30) scalarly by $\bar{y} = y^{(\sigma)} + \Delta_{0t_{j+\sigma}}^\alpha y + \Delta_{0t_{j+\sigma}}^\beta y - y_{\bar{x}x}^{(\sigma)}$:

$$\begin{aligned} & \left(\Delta_{0t_{j+\sigma}}^\alpha y, \bar{y} \right) = \left((ay_{\bar{x}}^{(\sigma)})_x, \bar{y} \right) + \eta \left(\Delta_{0t_{j+\sigma}}^\alpha y_{\bar{x}x}, \bar{y} \right) \\ & + \left(\sum_{s=1}^m r_{s,i}^j \left(y_{\bar{x},i_s}^{(\sigma)} x_{i_s}^- + y_{\bar{x},i_s+1}^{(\sigma)} x_{i_s}^+ \right), \bar{y} \right) - \left(dy^{(\sigma)}, \bar{y} \right) + (\varphi, \bar{y}), \end{aligned} \tag{34}$$

where $(u, v] = \sum_{i=1}^N u_i v_i h$, $h = \begin{cases} 0.5h, & i = N; \\ h, & i \neq 0, N, \end{cases}$

$(u, u] = (1, u^2] = \|u\|_0^2$, $(u, v) = \sum_{i=1}^{N-1} u_i v_i h$.

After some transformations from (34) we obtain

$$\begin{aligned} & \left(\Delta_{0t_{j+\sigma}}^\alpha y, \bar{y} \right) + c_0 \|y_{\bar{x}}^{(\sigma)}\|_0^2 + \frac{c_0}{2} \Delta_{0t_{j+\sigma}}^\alpha \|y_{\bar{x}}\|_0^2 + \frac{\eta + c_0}{2} \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}}\|_0^2 \\ & + \frac{c_0}{2} \|y_{\bar{x}x}^{(\sigma)}\|_0^2 + \eta \|\Delta_{0t_{j+\sigma}}^\beta y_{\bar{x}}\|_0^2 + \frac{\eta}{2} \Delta_{0t_{j+\sigma}}^\beta \|y_{\bar{x}x}\|_0^2 + \left(\eta \Delta_{0t_{j+\sigma}}^\alpha y_{\bar{x}}, \Delta_{0t_{j+\sigma}}^\beta y_{\bar{x}} \right) \\ & \leq \left(ay_{\bar{x}}^{(\sigma)} + \eta \Delta_{0t_{j+\sigma}}^\beta y_{\bar{x}} \right) \left(y^{(\sigma)} + \Delta_{0t_{j+\sigma}}^\alpha y + \Delta_{0t_{j+\sigma}}^\beta y \right) \Big|_0^N + M_1 \|y_{\bar{x}}^{(\sigma)}\|_0^2 \\ & + \left(\sum_{s=1}^m r_{s,i}^j \left(y_{\bar{x},i_s}^{(\sigma)} \frac{x_{i_s+1} - x_s}{h} + y_{\bar{x},i_s+1}^{(\sigma)} \frac{x_s - x_{i_s}}{h} \right), \bar{y} \right) - \left(dy^{(\sigma)}, \bar{y} \right) + (\varphi, \bar{y}). \end{aligned} \tag{35}$$

We transform the first expression on the right-hand side of (35), then we get

$$\begin{aligned}
& \left(ay_{\bar{x}}^{(\sigma)} + \eta \Delta_{0t_{j+\sigma}}^{\beta} y_{\bar{x}} \right) \left(y^{(\sigma)} + \Delta_{0t_{j+\sigma}}^{\alpha} y + \Delta_{0t_{j+\sigma}}^{\beta} y \right) \Big|_0^N \\
&= \left(a_N y_{\bar{x},N}^{(\sigma)} + \eta \Delta_{0t_{j+\sigma}}^{\beta} y_{\bar{x},N} \right) \left(y_N^{(\sigma)} + \Delta_{0t_{j+\sigma}}^{\alpha} y_N + \Delta_{0t_{j+\sigma}}^{\beta} y_N \right) \\
&= \left[\tilde{\mu} - \tilde{\beta}_1 y_N^{(\sigma)} - \tilde{\beta}_2 \Delta_{0t_{j+\sigma}}^{\alpha} y_N - \beta_3 \Delta_{0t_{j+\sigma}}^{\beta} y_N + 0.5h \sum_{s=1}^m \left(r_{s,N} \left(y_{\bar{x},i_s}^{(\sigma)} \frac{x_{i_s+1} - x_s}{h} \right. \right. \right. \\
&\quad \left. \left. \left. + y_{\bar{x},i_s+1}^{(\sigma)} \frac{x_s - x_{i_s}}{h} \right) \right) \right] \left(y_N^{(\sigma)} + \Delta_{0t_{j+\sigma}}^{\alpha} y_N + \Delta_{0t_{j+\sigma}}^{\beta} y_N \right) \\
&= \left(\mu - \beta_1 y_N^{(\sigma)} - \beta_2 \Delta_{0t_{j+\sigma}}^{\alpha} y_N - \beta_3 \Delta_{0t_{j+\sigma}}^{\beta} y_N \right) \left(y_N^{(\sigma)} + \Delta_{0t_{j+\sigma}}^{\alpha} y_N + \Delta_{0t_{j+\sigma}}^{\beta} y_N \right) \\
&+ 0.5h \left(\varphi_N - d_N y_N^{(\sigma)} - \Delta_{0t_{j+\sigma}}^{\alpha} y_N + \sum_{s=1}^m \left(r_{s,N} \left(y_{\bar{x},i_s}^{(\sigma)} \frac{x_{i_s+1} - x_s}{h} + y_{\bar{x},i_s+1}^{(\sigma)} \frac{x_s - x_{i_s}}{h} \right) \right) \right) \\
&\quad * \left(y_N^{(\sigma)} + \Delta_{0t_{j+\sigma}}^{\alpha} y_N + \Delta_{0t_{j+\sigma}}^{\beta} y_N \right). \tag{36}
\end{aligned}$$

We transform the first expression on the right-hand side of (36), then we find

$$\begin{aligned}
& \left(\mu - \beta_1 y_N^{(\sigma)} - \beta_2 \Delta_{0t_{j+\sigma}}^{\alpha} y_N - \beta_3 \Delta_{0t_{j+\sigma}}^{\beta} y_N \right) \left(y_N^{(\sigma)} + \Delta_{0t_{j+\sigma}}^{\alpha} y_N + \Delta_{0t_{j+\sigma}}^{\beta} y_N \right) \\
&\leq -\frac{\beta_2}{2} \left(\Delta_{0t_{j+\sigma}}^{\alpha} y_N \right)^2 - \frac{\beta_2}{2} \Delta_{0t_{j+\sigma}}^{\alpha} y_N^2 - (\beta_2 + \beta_3) \Delta_{0t_{j+\sigma}}^{\alpha} y_N \Delta_{0t_{j+\sigma}}^{\beta} y_N \\
&- \frac{\beta_3}{2} \left(\Delta_{0t_{j+\sigma}}^{\beta} y_N \right)^2 - \frac{\beta_3}{2} \Delta_{0t_{j+\sigma}}^{\beta} y_N^2 + M_2 \left(\|y^{(\sigma)}\|_0^2 + \|y_{\bar{x}}^{(\sigma)}\|_0^2 \right) + M_3 \mu^2. \tag{37}
\end{aligned}$$

Taking into account (36) and (37), from (35) we get

$$\begin{aligned}
& \left[\Delta_{0t_{j+\sigma}}^{\alpha} y, y^{(\sigma)} + \Delta_{0t_{j+\sigma}}^{\alpha} y + \Delta_{0t_{j+\sigma}}^{\beta} y \right] + c_0 \|y_{\bar{x}}^{(\sigma)}\|_0^2 + \frac{c_0}{2} \Delta_{0t_{j+\sigma}}^{\alpha} \|y_{\bar{x}}\|_0^2 + \frac{c_0}{2} \|y_{\bar{x}x}^{(\sigma)}\|_0^2 \\
&+ \frac{\eta + c_0}{2} \Delta_{0t_{j+\sigma}}^{\beta} \|y_{\bar{x}}\|_0^2 + \eta \|\Delta_{0t_{j+\sigma}}^{\beta} y_{\bar{x}}\|_0^2 + \frac{\eta}{2} \Delta_{0t_{j+\sigma}}^{\beta} \|y_{\bar{x}x}\|_0^2 + \left(\eta \Delta_{0t_{j+\sigma}}^{\alpha} y_{\bar{x}}, \Delta_{0t_{j+\sigma}}^{\beta} y_{\bar{x}} \right) \\
&\quad + \frac{\beta_2}{2} \left(\Delta_{0t_{j+\sigma}}^{\alpha} y_N \right)^2 + \frac{\beta_2}{2} \Delta_{0t_{j+\sigma}}^{\alpha} y_N^2 + (\beta_2 + \beta_3) \Delta_{0t_{j+\sigma}}^{\alpha} y_N \Delta_{0t_{j+\sigma}}^{\beta} y_N \\
&+ \frac{\beta_3}{2} \left(\Delta_{0t_{j+\sigma}}^{\beta} y_N \right)^2 + \frac{\beta_3}{2} \Delta_{0t_{j+\sigma}}^{\beta} y_N^2 \leq \left(\Delta_{0t_{j+\sigma}}^{\alpha} y, y_{\bar{x}}^{(\sigma)} \right) + M_4 \left(\|y^{(\sigma)}\|_0^2 + \|y_{\bar{x}}^{(\sigma)}\|_0^2 \right) \\
&+ \left[\sum_{s=1}^m r_s \left(y_{\bar{x},i_s}^{(\sigma)} \frac{x_{i_s+1} - x_s}{h} + y_{\bar{x},i_s+1}^{(\sigma)} \frac{x_s - x_{i_s}}{h} \right), y^{(\sigma)} + \Delta_{0t_{j+\sigma}}^{\alpha} y + \Delta_{0t_{j+\sigma}}^{\beta} y \right] \\
&\quad - \left(\sum_{s=1}^m r_s \left(y_{\bar{x},i_s}^{(\sigma)} \frac{x_{i_s+1} - x_s}{h} + y_{\bar{x},i_s+1}^{(\sigma)} \frac{x_s - x_{i_s}}{h} \right), y_{\bar{x}x}^{(\sigma)} \right) + M_7 \mu^2 \\
&\quad - \left[dy^{(\sigma)}, y^{(\sigma)} + \Delta_{0t_{j+\sigma}}^{\alpha} y + \Delta_{0t_{j+\sigma}}^{\beta} y \right] + \left(dy^{(\sigma)}, y_{\bar{x}x}^{(\sigma)} \right) \\
&\quad + \left[\varphi, y^{(\sigma)} + \Delta_{0t_{j+\sigma}}^{\alpha} y + \Delta_{0t_{j+\sigma}}^{\beta} y \right] + \left(\varphi, y_{\bar{x}x}^{(\sigma)} \right). \tag{38}
\end{aligned}$$

By virtue of the condition (31), that the inequality $\|y^{(\sigma)}\|_0^2 \leq 2l^2 \|y_{\bar{x}}^{(\sigma)}\|_0^2$ is valid, from (38) after some simple transformations, we obtain the inequality

$$\begin{aligned} & \Delta_{0t_{j+\sigma}}^\alpha (\|y\|_0^2 + \|y_{\bar{x}}\|_0^2 + y_N^2) + \Delta_{0t_{j+\sigma}}^\beta (\|y_{\bar{x}}\|_0^2 + \|y_{\bar{x}x}\|_0^2 + y_N^2) + \|\Delta_{0t_{j+\sigma}}^\alpha y\|_0^2 \\ & + \|\Delta_{0t_{j+\sigma}}^\beta y_{\bar{x}}\|_0^2 + \|y_{\bar{x}}^{(\sigma)}\|_0^2 + \|y_{\bar{x}x}^{(\sigma)}\|_0^2 + \left(\Delta_{0t_{j+\sigma}}^\alpha y_N\right)^2 + \left(\Delta_{0t_{j+\sigma}}^\beta y_N\right)^2 \\ & \leq M_9 \|y_{\bar{x}}^{(\sigma)}\|_0^2 + M_{10} (\|\varphi\|_0^2 + \mu^2). \end{aligned} \quad (39)$$

We rewrite (39) in another form

$$\begin{aligned} & \Delta_{0t_{j+\sigma}}^\alpha (\|y\|_0^2 + \|y_{\bar{x}}\|_0^2 + y_N^2) + \Delta_{0t_{j+\sigma}}^\beta (\|y_{\bar{x}}\|_0^2 + \|y_{\bar{x}x}\|_0^2 + y_N^2) \\ & \leq M_{11} \|y_{\bar{x}}^{j+1}\|_0^2 + M_{12} \|y_{\bar{x}}^j\|_0^2 + M_{13} (\|\varphi\|_0^2 + \mu^2). \end{aligned} \quad (40)$$

Based on Lemma 7 [17], from (40) we obtain a priori estimates:

1) In case when $\alpha > \beta$:

$$\|y\|_2^2 \leq M_{14} \left(\|y^0\|_2^2 + \max_{0 \leq j' \leq j} (\|\varphi^{j'}\|_0^2 + \mu^2) \right), \quad (41)$$

where $\|y\|_2^2 = \|y\|_0^2 + \|y_{\bar{x}}\|_0^2$.

2) In case when $\alpha = \beta$:

$$\|y\|_3^2 \leq M_{15} \left(\|y^0\|_3^2 + \max_{0 \leq j' \leq j} (\|\varphi^{j'}\|_0^2 + \mu^2) \right), \quad (42)$$

where $\|y\|_3^2 = \|y\|_0^2 + \|y_{\bar{x}}\|_0^2 + \|y_{\bar{x}x}\|_0^2$.

3) In case when $\alpha < \beta$:

$$\|y\|_1^2 \leq M_{16} \left(\|y\|_2^2 + \max_{0 \leq j' \leq j} (\|\varphi^{j'}\|_0^2 + \mu^2) \right), \quad (43)$$

where $M_{14}, M_{15}, M_{16} = \text{const} > 0$, independent of h and τ .

3 Results

Theorem 1. If $k(x, t) \in C^{1,0}(Q_T)$, $r_s(x, t), q(x, t), f(x, t) \in C(Q_T)$, $u(x, t) \in C^{(2,0)}(Q_T) \cap C^{(1,0)}(\bar{Q}_T)$, $\partial_{0t}^\alpha u(x, t) \in C(Q_T)$, $\partial_{0t}^\alpha u_{xx}(x, t) \in C(\bar{Q}_T)$ and conditions (4) be satisfied, then estimates: (9) in the case, when $\alpha > \beta$; (11) in the case, when $\alpha \leq \beta$ are valid for the solution $u(x, t)$ of problem (1)–(3).

The a priori estimates (9), (11) implies the uniqueness and stability of the solution with respect to the right-hand side and the initial data

Theorem 2. Let conditions (4) be satisfied, then exist $\tau_0 = \tau_0(c_0, c_1, c_2, \alpha, \sigma)$, such that if $\tau \leq \tau_0$, then estimates: (19), in the case when $\alpha > \beta$; (20) in the case when $\alpha \leq \beta$ are valid for the solution of the difference problem (12)–(14).

The a priori estimates (19), (20) implies the uniqueness and stability of the solution to the difference scheme (12)–(14) with respect to the initial data and the right-hand side, as well as the convergence of the solution of the difference problem (12)–(14) to the solution of the differential problem (1)–(3) so that if there exist such τ_0 then for $\tau \leq \tau_0$ the a priori estimates are valid:

- 1) in case, when $\alpha > \beta$: $\|y_{\bar{x}}^{j+1} - u_{\bar{x}}^{j+1}\|_0^2 \leq M (h^2 + \tau^{2-\max\{\alpha,\beta\}})$;
- 2) in case, when $\alpha = \beta$: $\|y^{j+1} - u^{j+1}\|_1^2 \leq M (h^2 + \tau^2)$;
- 3) in case, when $\alpha < \beta$: $\|y^{j+1} - u^{j+1}\|_1^2 \leq M (h^2 + \tau^{2-\max\{\alpha,\beta\}})$.

where $M - const > 0$, independent of h and τ .

Theorem 3. If $k(x, t) \in C^{1,0}(Q_T)$, $r_s(x, t), q(x, t), f(x, t) \in C(Q_T)$, $u(x, t) \in C^{(2,0)}(Q_T) \cap C^{(1,0)}(\bar{Q}_T)$, $\partial_{0t}^\alpha u(x, t) \in C(Q_T)$, $\partial_{0t}^\alpha u_{xx}(x, t) \in C(\bar{Q}_T)$ and conditions (4), (22), be satisfied, then estimates: (27) in the case when $\alpha > \beta$; (28) in the case when $\alpha = \beta$; (29) in the case when $\alpha < \beta$ are valid for the solution $u(x, t)$ of problem (1), (21), (3).

The a priori estimates (27), (28), (29) implies the uniqueness and stability of the solution with respect to the right-hand side and the initial data

Theorem 4. Let conditions (4), (22), be satisfied, then exist $\tau_0 = \tau_0(c_0, c_1, c_2, \alpha, \sigma)$, such that if $\tau \leq \tau_0$, then estimates: (41), in the case, when $\alpha > \beta$; (42) in the case, when $\alpha = \beta$; (43) in the case, when $\alpha < \beta$ are valid for the solution of the difference problem (30)–(33).

The a priori estimates (41)–(43) implies the uniqueness and stability of the solution to the difference scheme (30)–(33) with respect to the initial data and the right-hand side, as well as the convergence of the solution of the difference problem (30)–(33) to the solution of the differential problem (1), (21), (3) so that if there exist such τ_0 then for $\tau \leq \tau_0$ the a priori estimates are valid:

- 1) in case, when $\alpha > \beta$: $\|y^{j+1} - u^{j+1}\|_2^2 \leq M (h^2 + \tau^{2-\max\{\alpha,\beta\}})$;
- 2) in case, when $\alpha = \beta$: $\|y^{j+1} - u^{j+1}\|_3^2 \leq M (h^2 + \tau^2)$;
- 3) in case, when $\alpha < \beta$: $\|y^{j+1} - u^{j+1}\|_1^2 \leq M (h^2 + \tau^{2-\max\{\alpha,\beta\}})$,

where $M - const > 0$, independent of h and τ .

Comment. To solve numerically the boundary value problems considered in this paper for the loaded moisture transfer equation with two fractional Gerasimov-Caputo derivatives of different orders α, β , difference schemes (12)–(14) and (30)–(33) are reduced to calculated form; for this purpose one can use the method of parametric sweep [20].

4 Conclusion

This paper is devoted to the study of local and nonlocal boundary value problems for the loaded moisture transfer equation with two fractional Gerasimov-Caputo derivatives of different orders α , β . Using the method of energy inequalities for various relations between α and β , a priori estimates in differential and difference interpretations are obtained for solving the problems under consideration, which implies the uniqueness and stability of the solution with respect to the initial data and the right-hand side, as well as the convergence of the solution of the difference problem to the solution of the differential problem with the rate $O(h^2 + \tau^2)$ for $\alpha = \beta$ and $O(h^2 + \tau^{2-\max\{\alpha,\beta\}})$ for $\alpha \neq \beta$.




References

1. Barenblatt, G.I., Zheltov, Y.P., Kochina, I.N.: On basic concepts of the theory of filtration of homogeneous liquids in cracked rocks. *Prikl. Mat. Mekh.* **25**(5), 852–864 (1960)
2. Dzektsler, Ye.S.: The equations of motion of groundwater with a free surface in multilayer environments. *DAN SSSR* **220**(3), 540–543 (1975)
3. Rubinshteyn, L.I.: To the question of the process of heat distribution in heterogeneous media. *AN SSSR, Ser. Geogr.* **12**(1), 27–45 (1948)
4. Ting, T.W.: A cooling process according to two-temperature theory of heat conduction. *J. Math. Anal. Appl* **45**(9), 23–31 (1974)
5. Hallaire, M.: On a theory of moisture-transfer. *Inst. Rech. Agronom.* **3**, 60–72 (1964)
6. Chudnovskiy, A.F.: (1976): Thermophysics of soils. Nauka, Moscow (1976)
7. Sveshnikov, A.A., Al'shin, A.B., Korpusov, M.O., Pletner, Y.D.: Linear and Nonlinear Sobolev-Type Equations. Fizmatlit, Moscow (2007)
8. Bedanokova, S.YU.: The equation of motion of soil moisture and a mathematical model of the moisture content of the soil layer based on the Hillaire's equation. *Vestnik Adygeyskogo gosudarstvennogo universiteta. Seriya 4: Yestestvenno matematicheskiye i tekhnicheskkiye nauki.* 4, pp. 68–71 (2007)
9. Nakhushhev, A.M.: Equations of Mathematical Biology. Vysshaya Shkola, Moscow (1995)
10. Dzhaneliyev, M.T.: On a quadratic functional in the Cauchy problem for a loaded first-order differential operator equation. *Differ. Eq.* **31**(12), 2029–2037 (1995)
11. Cannon, J.R., Yin, N.M.: (1989) On a class of nonlinear nonclassical parabolic problems. *J. Different. Equat.* **79**, 266–288 (1989)
12. Goloviznin, V.M., Kiselev, V.P., Korotkij, I.A., Yurkov, Y.I.: Some features of computing algorithms for the equations fractional diffusion. Preprint. IBRAE-2002-01. Moscow: Nuclear Safety Institute RAS (2002)
13. Alikhanov, A.A., Berezgov, A.M., Shkhanukov-Lafishev, M.K.H.: Boundary value problems for certain classes of loaded differential equations and solving them by finite difference methods. *Comput. Math. Math. Phys.* **48**(9), 1581–1590 (2008)
14. Alikhanov, A.A.: A priori estimates for solutions of boundary value problems for fractional-order equations. *Differ. Eq.* **46**(5), 660–666 (2010)
15. Alikhanov, A.A.: A new difference scheme for the time fractional diffusion equation. *J. Comput. Phys.* **280**, 424–438 (2015)

16. Beshtokov, M.K.: Boundary value problems for degenerating and nondegenerating Sobolev-type equations with a nonlocal source in differential and difference forms. *Differ. Eq.* **54**(2), 250–267 (2018)
17. Beshtokov, M.K.: Numerical analysis of initial-boundary value problem for a sobolev-type equation with a fractional-order time derivative. *Comput. Math. Math. Phys.* **59**(2), 175–192 (2019)
18. Beshtokov, M.Kh.: The third boundary value problem for loaded differential Sobolev type equation and grid methods of their numerical implementation. *IOP Conf. Ser.: Mater. Sci. Eng.* **158**(1) (2016)
19. Samarskiy, A.A.: *Theory of Difference Schemes*. Nauka, Moscow (1983)
20. Voevodin, A.F., Shugrin, S.M.: *Numerical methods for calculating one-dimensional systems*, SO AS USSR. Nauka, Novosibirsk (1981)



Determining Frequencies of Free Longitudinal Vibrations of Rods by Analytical and Numerical Methods

Kh. P. Kulterbaev¹ (✉) , M. M. Lafisheva¹ , and L. A. Baragunova² 

¹ North-Caucasus Center for Mathematical Research, North-Caucasus Federal University, Stavropol, Russia
kulthp@mail.ru

² Kabardino-Balkaria State University H.M. Berbekov, 173, Chernyshevsky Street, Nalchik, Russia

Abstract. The longitudinal vibrations of a steel rod of constant cross-section along the length are considered. The left end of the rod is pinched, a concentrated mass is attached to the right end. The mathematical model consists of a hyperbolic partial differential equation and boundary conditions. The D'Alembert's principle is used. Free oscillations are undamped and harmonic without initial conditions. The purpose of solving the problem is to determine the frequencies of free oscillations. Analytical and numerical-graphical methods of solving the problem are used. In the first case, the oscillation frequency is determined from the transcendental equation. In the second case, the basic equation and boundary conditions are replaced by a system of algebraic equations. The required oscillation frequencies are defined as the eigenvalues of a square matrix. At the final stage of determining the frequencies of free oscillations, a numerical-graphical method is used, implemented in the environment of the Matlab computing complex. A concrete example of the solution is given. Conclusions are drawn.

Keywords: Longitudinal oscillations of a steel rod · Partial differential equations · The principle of D'Alembert · Finite difference method · The oscillation frequency · Homogeneous system of algebraic equations · Eigenvalues · Matlab computing complex

1 Introduction

Steel rods are often found in construction structures, machine-building and machine-tool parts, in structures of the oil and gas and chemical industries. At the same time, they are supplemented with a variety of carried masses and supports, loaded with dynamic and kinematic effects. The frequencies of their oscillations are of great interest, since their coincidence with the frequencies of loads leads to resonance and the occurrence of large stresses and deformations that threaten the destruction of structures.

An extensive bibliography is devoted to longitudinal vibrations of rods and related problems. The issues of solving such problems by traditional methods are presented

in scientific and educational publications [1–6]. In recent years, the use of numerical methods, computer systems and computer technologies have allowed us to move on to the design of more complex and optimal structures. In this case, analytical and numerical methods are used to solve the determination of the frequencies of free oscillations. The result was numerous publications [7–10].

In this article, the frequencies of free vibrations are determined by the analytical and numerical methods currently used. The purpose of the study is to determine the most preferred of them for practical use.

2 Problem Statement

Figure 1 shows a steel rod of constant cross-section. Its characteristics are: length l , elastic modulus E , linear mass $m = \rho A$, material density ρ , cross-section area A .

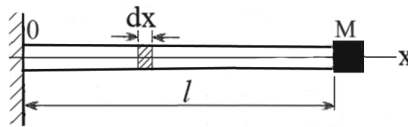


Fig. 1. Calculation scheme

The left end of the rod is fixed, there is a concentrated mass M at the right end. The cross sections of the rod make free oscillations in the longitudinal direction. The basic equation of vibrations is derived by considering the vibrations of the rod element (Fig. 2), to which forces are applied in sections $N, N + N' dx$ and the d'Alembert inertia force $dl = m\ddot{u}dx$.

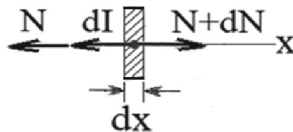


Fig. 2. Calculation scheme

As a result of simple calculations using the d'Alembert principle, it is possible to obtain the basic equation of longitudinal oscillations [2].

$$\ddot{u} - a^2 u'' = 0, a^2 = \frac{E}{\rho}, x \in (0, l), t > -\infty. \tag{1}$$

Here the points above u correspond to differentiation by t , the strokes in the upper indices correspond to differentiation by the spatial coordinate x . Additional boundary conditions are added to (1) at the left and right ends of the rod

$$u(0, t) = 0, u'(l, t) + b\ddot{u}(l, t) = 0, b = \frac{M}{EA}, t > -\infty. \tag{2}$$

We will assume that the hypothesis of plane sections is valid, and we will neglect the transverse movements of mass particles. The displacements of the sections are characterized by the function $u(x, t)$.

Equations (1), (2) form a mathematical model for determining the natural frequencies of longitudinal vibrations of the rod.

We will solve the problem using traditional analytical and numerical-graphical methods.

3 Analytical Solution Method

Let's write out the general solution of Eq. (1)

$$u(x, t) = X(x)e^{i\omega t}. \quad (3)$$

Its substitution in (1) gives

$$a^2 X''(x) + \omega^2 X(x) = 0. \quad (4)$$

and the shape of the oscillations

$$X(x) = C \sin kx + D \cos kx, k = \omega/a, a = \sqrt{E/\rho}. \quad (5)$$

Boundary conditions (2) and representation (3) give

$$X(0) = 0, X'(l) - bX(l)\omega^2 = 0. \quad (6)$$

Differentiating (5), we get

$$X'(x) = k(C \cos kx - D \sin kx). \quad (7)$$

Substituting (5), (7) into (6) and performing simple transformations, we have

$$C \neq 0, \rightarrow C \cos kl = 0, k \cos kl - b\omega^2 \sin kl = 0. \quad (8)$$

By solving this transcendental equation, it is possible to obtain the values of the natural frequencies of ω_i . Let's represent the left side of the equation as some function

$$f(\omega) = 0. \quad (9)$$

Those values of ω , that turn the values of the function to zero will be the natural frequencies of free oscillations.

Let's take a concrete example.

Example 1. A steel rod of constant cross-section is given according to Fig. 1 with parameters $l = 2 \text{ m}$, $E = 2 \cdot 10^{11} \text{ N/m}^2$, $\rho = 7800 \text{ kg/m}^3$, $M = 100 \text{ kg}$.

It is required to find the roots of the transcendental Eq. (7), which are the frequencies of free oscillations.

The analytical method of solving the problem under difficult boundary conditions usually leads to the need to solve transcendental equations, as in this example. Therefore, we will use the capabilities of the Matlab computing complex.

According to the algorithms presented above, a computer program was compiled, implemented in the environment of the Matlab computing complex. The Matlab complex has a tool that, like a magnifying glass, is able to multiply the fragments of the graph. Using them, it is possible to obtain natural frequencies with a high degree of accuracy. The obtained result is shown in Fig. 3, where the first natural oscillation frequency $\omega_1 = 8176s^{-1}$ is marked with a bold dot on the abscissa axis. Other higher frequencies are determined similarly. This method of determining eigenvalues was first used in the article [11] and subsequently called the numerical-graphical method. The following high frequencies are of only theoretical interest. Therefore, we will not define them.

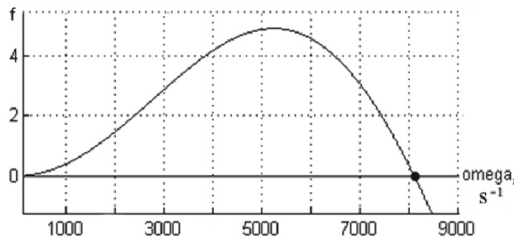


Fig. 3. Frequencies of free oscillations

4 Numerical Solution Method

Let’s consider the solution of the previous example by the finite difference method.

Let’s move from differential operators to finite difference operators and to the finite difference method (FDM). The area of continuous change of the argument of the function $x \in [0, l]$ is replaced by a uniform grid with a step h (Fig. 4)

$$l_h = \{x_i = (i - 1)h, h = l/(n - 1), i = 1, 2, \dots, n, \}$$

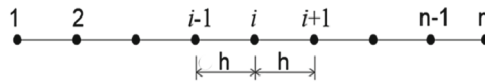


Fig. 4. Uniform grid

Then the function $X(x)$ corresponds to the grid function $X_i(x_i)$ at the nodes, i.e. $X_i(x_i) \approx X(x_i)$. Next, we will use finite-difference operators that ensure the accuracy of $O(h^2)$ [3].

The functions of the argument x and derivatives in the Eqs. (4), (6), (7) are presented in the form of:

$$\begin{aligned} X &= \{X_1, X_2, \dots, X_n\}, X''(x) = \frac{1}{2h}(-X_{i-1} + X_{i+1}), \\ X''(x) &= \frac{1}{h^2}(X_{i-1} - 2X_i + X_{i+1}). \end{aligned} \tag{10}$$

As in the previous case, the capabilities of the Matlab software package were used. In the coordinate system $\omega - det$ a curved line corresponding to Eq. (16) is obtained, almost completely coinciding with the curve of Fig. 3. The first natural frequency, up to an integer, coincides with the analytical method obtained above, $\omega_1 = 8176s^{-1}$.

Such an exact coincidence of the results obtained by analytical and numerical methods guarantees the reliability and validity of problem statements and solutions.

5 Conclusion




1. The solution of the well-known problem of longitudinal free oscillations of rods by determining natural frequencies is significantly simplified when using the numerical-graphical method.
2. The used computing program in the environment of the Matlab computing complex has the property of universality and easily adapts to changes in the input data of the task.
3. Verification of the analytical and numerical methods proposed in the work is ensured by the exact coincidence of the frequencies of free oscillations.

References

1. Tikhonov, A.N., Samarsky, A.A.: Equations of mathematical physics. In: Publishing House of Moscow State University, p. 799 (1999)
2. Kulterbaev, Kh.P.: Fundamentals of the theory of oscillations. Fundamentals of theory, tasks for homework, examples of solutions. In: Kabardino-Balkarian State University, p. 130 (2003)
3. Gurbatov, S.N., Gryaznova, I.Y., Demin, I.u., Klemina, A.V., Kurin, V.V., Pronchatov-Rubtsov, N.V., Lisin, A.A.: Oscillations of mechanical systems with distributed parameters. In: Nizhny Novgorod State University, p. 28 (2021)
4. Nurimbetov, A.U., Dzhunisbekov, M.Sh.: Vibrations of a rectangular rod. In: Pr. of the VIII th. scientific.pract.conf. "Engineering systems - 2015", pp. 97–102. RUDN, Moscow (2015)
5. Biderman, V.L.: Applied theory of mechanical oscillations. In: Higher school, p. 416. Moscow (1972)
6. Ed. Bolotina, V.V.: Vibrations in technology. Handbook in 6 volumes. Volume 1. Oscillations of linear systems. In: Mashinostroenie, p. 352. Moscow (1978)
7. Karamansky, T.D.: Numerical methods of structural mechanics. In: Stroyizdat, p. 436. Moscow (1981)
8. Verzhbitsky, V.M.: Fundamentals of numerical methods. In: Higher school, p. 840. Moscow (2002)
9. Samarsky, A.A., Gulin, A.V.: Numerical methods. In: Nauka, p. 432. Moscow (1989)
10. Samarskii, A.A., Mikhailov, A.P.: Math modeling. Ideas. Methods. Examples. In: FIZ-MATLIT, p. 320. Moscow (2005)
11. Kulterbaev, Kh.P., Baragunova, L.A., Shogenova, M.M., Senov, Kh.M.: About a High-Precision Graphoanalytical Method of Determination of Critical Forces of an Oblate Rod. In: Proceedings 2018 IEEE International Conference "Quality Management, Transport and Information Security, Information Technologies"(IT&QM&IS), pp. 794–796. St. Petersburg, Russia (2018)



Forced Longitudinal Oscillations of a Rod with a Mass at the End

Kh. P. Kulterbaev¹ (✉) , M. M. Lafisheva¹ , and L. A. Baragunova² 

¹ North-Caucasus Center for Mathematical Research, North-Caucasus Federal University, Stavropol, Russia
kulthp@mail.ru

² Kabardino-Balkaria State University H.M. Berbekov, 173, Chernyshevsky Street, Nalchik, Russia

Abstract. Forced longitudinal vibrations of a steel rod of constant cross-section along its length are considered. The left end of the rod is pinched, a concentrated mass is attached to the right end. The source of the oscillations is the harmonic force acting along the axis. The mathematical model consists of a hyperbolic partial differential equation and boundary conditions. The D’Alembert’s principle is used. Forced oscillations are undamped and harmonic without initial conditions. The amplitudes of forced oscillations are determined, their dependences on the frequency of the driving force are analyzed. Analytical and numerical-graphical methods of solving the problem are used. In the first case, by traditional methods, the problem is reduced to a transcendental equation, from which the amplitude of the oscillations is determined. In the second case, the basic equation and boundary conditions are replaced by a system of algebraic equations. The desired oscillation amplitudes are determined from the matrix-vector equation. At the final stage of determining the amplitudes, a numerical-graphical method is used, implemented in the environment of the Matlab computing complex. Examples are given. Conclusions are drawn.

Keywords: Longitudinal forced vibrations of a steel rod · Partial differential equations · The principle of D Alembert · Finite difference method · The amplitude of the oscillations · The transcendental equation · an inhomogeneous system of algebraic equations · Matlab computing complex

1 Introduction

Rods are widely used in various fields of technology: in machines, building structures and appliances. Their configurations are very different: beams, trusses, frames, racks, frames of cars and airplanes, towers, etc. In real conditions, dynamic loads often act on the rods, which leads to fluctuations. Therefore, the theory of oscillations currently has an extensive bibliography in the form of monographs, textbooks and journal articles.

Rods are often supplemented with a variety of load-bearing masses and supports, loaded with dynamic and static forces. The oscillation amplitudes of such structures are

of great interest, since normal stresses in cross sections and deformations of the rod, which threaten the destruction of structures, significantly depend on their magnitude.

An extensive bibliography is devoted to longitudinal free and forced vibrations of rods and related problems. The issues of solving such problems by traditional methods are presented in scientific and educational publications [1–7]. In recent years, the use of numerical methods, computer systems and computer technologies [8–10] has allowed us to move on to designing more complex and optimal structures.

The purpose of this article is to determine by analytical and numerical methods the displacements of the cross sections of the rod under the longitudinal force and to show the high efficiency of the practical application of computer systems.

2 Mathematical model of the problem

Figure 1 shows a steel structure consisting of a rod of constant cross-section and a mass concentrated at the right end. It oscillates in the horizontal direction, the source of which is the harmonic force $F(t)$. The characteristics of the oscillatory system are: length l , elastic modulus E , linear mass $m = \rho A$, ρ is the density of the material, A is the cross-sectional area, M is the mass, f is the amplitude of the driving force, ω is the circular frequency of the driving force. The basic equation of vibrations is determined by considering the vibrations of the rod element (Fig. 2), to which longitudinal forces in sections $N, N + dN$ and the Dalember inertia force $dI = m\ddot{u}dx$.



Fig. 1. Calculation scheme

When deriving the oscillation equations, we will assume that the hypothesis of plane sections is valid, and we will neglect the transverse displacements of mass particles. The longitudinal displacements of the sections are denoted by $u(x,t)$. Let's take the rod element (Fig. 2) and determine the relative deformation.

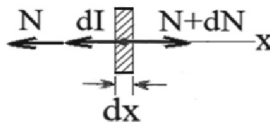


Fig. 2. Elementary section of the rod

It is known that the relative deformation ε is determined by the formula

$$\varepsilon = \frac{\Delta dx}{dx}$$

where dx is the elongation of the element dx . From the drawing

$$\Delta dx = u + u'dx - u = u'dx.$$

Therefore,

$$\varepsilon = \frac{u'dx}{dx} = u'.$$

In the above formulas, the strokes in the upper indices mean differentiation by the spatial coordinate x . According to Fig. 2, the equilibrium equation of the element has the form

$$dI = dN.$$

As a result of simple calculations using the D'Alembert principle, it is possible to obtain the basic equation of longitudinal oscillations.

$$dI = mdx\ddot{u}, dN = N'dx, N = \sigma A = E\varepsilon A = EAu'.$$

They lead to equality

$$-m\ddot{u} + EAu'' = 0 \quad (1)$$

Convert (1) to a more convenient form

$$\ddot{u} - a^2 u'' = 0, a^2 = \frac{E}{\rho}, x \in (0, l), t > -\infty. \quad (2)$$

The result obtained is the basic equation of longitudinal oscillations [2]. Here the points above u correspond to the differentiation in time t . Additional boundary conditions at the left and right ends of the rod are added to (2). The left end is sealed. Therefore, there is no moving it.

$$u(0, t) = 0 \quad (3)$$

The boundary conditions at the right end are determined taking into account the presence of mass M and force F (Fig. 3). The equilibrium condition follows from the figure

$$-M\ddot{u} - N + F = 0, F = fe^{i\omega t} \quad (4)$$

Equations (2)–(4) form a mathematical model for determining the parameters of longitudinal vibrations of the rod. We will solve the problem by two methods: traditional analytical and numerical-graphical.



Fig. 3. Right end

3 Analytical Solution Method

The general solution of Eq. (2) has the form

$$u(x, t) = X(x)e^{i\omega t}. \tag{5}$$

The general solution of Eq. (2) will be

$$a^2 X''(x) + \omega^2 X(x) = 0. \tag{6}$$

The general solution of Eq. (2) will be.

$$X(x) = C \sin kx + D \cos kx, k = \omega/a, a = \sqrt{E/\rho}$$

Taking into account the boundary condition (3) gives.

$$D = 0, X(x) = C \sin kx. \tag{7}$$

Differentiating (7), we get.

$$X'(x) = Ck \cos kx. \tag{8}$$

Boundary conditions (3), (4) and representation (5) give.

$$X(0) = 0, \tag{9}$$

$$C(EAk \cos kl - M \sin k l \omega^2) = f \tag{10}$$

Here the second equation is transcendental. With its help, it is possible to obtain the values of the natural frequencies of ω_i , the amplitude of oscillations at given values of the frequency of the force F and other parameters of forced oscillations. When determining the natural frequencies in Eq. (10), the right part should be zero and the transcendental equation will take the form.

$$EAk \cos kl - M \sin k l \omega^2 = 0 \tag{11}$$

It is difficult to find the exact values of the roots of Eq. (11), which are the eigenvalues of free oscillations. Therefore, we will use the capabilities of the Matlab computing complex.

Let us take a concrete example.

Example 1. A steel rod of constant cross-section is given according to Fig. 1 with the parameters: $l = 2 \text{ m}$, $E = 2,1 \cdot 10^{11} \text{ N/m}^2$, $A = 2 \text{ cm}^2$, $\rho = 7800 \text{ kg/m}^3$, $M = 100 \text{ kg}$.

It is required to find the roots of the transcendental Eq. (11), which are the frequencies of free oscillations.

The analytical method of solving the problem under difficult boundary conditions usually leads to the need to solve transcendental equations, as in this example.

According to the algorithms presented above, a computer program was compiled, implemented in the Matlab environment. The Matlab complex has a tool that, like a magnifying glass, is able to multiply the fragments of the graph. Using them, it is possible to obtain natural frequencies with a high degree of accuracy. The obtained result is shown in Fig. 4, where the first natural oscillation frequency $\omega_1 = 8176, 15 \text{ s}^{-1}$ is marked with a bold dot on the abscissa axis. Other higher.

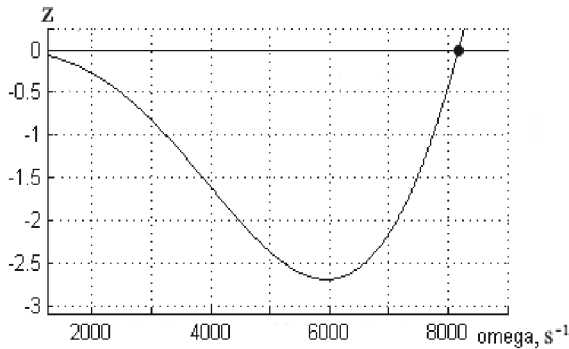


Fig. 4. The first frequency of free oscillations.

frequencies are determined similarly. This method of determining eigenvalues was first used in the article [11] and subsequently called the numerical-graphical method. The following high frequencies in most cases are of only theoretical interest. But they arise in high-frequency technology and measuring devices.

Let us turn to such a case and consider the range of calculations up to $\omega = 100000 \text{ s}^{-1}$. The computer program gave the result shown in Fig. 5. There were 12 natural frequencies in this range. All of them are marked with dots, their numerical values are easy to read when the picture is enlarged. In this way, the first three natural frequencies are established $\Omega = \{8176, 16314, 24460\} \text{ s}^{-1}$.

Let's move on to calculating the amplitudes of forced vibrations for the steel rod considered in Example 1. In this case, we will use the function (7) obtained above.

Example 2. A steel rod of constant cross-section is given (Fig. 1) with the parameters: $l = 2 \text{ m}$, $E = 2,1 \cdot 10^{11} \text{ N/m}^2$, $A = 2 \text{ cm}^2$, $\rho = 7800 \text{ kg/m}^3$, $M = 100 \text{ kg}$, $\omega = \{6000, 15000, 24000\} \text{ s}^{-1}$, $f = 10000 \text{ kN}$.

Required: calculate the values of $X(x)$ at $x \in [0, l]$ and show three graphs of oscillation amplitudes.

The results of the calculation are shown by curves 1, 2, 3 (Fig. 6). In the $x - X$ coordinate system, three curved lines corresponding to the three frequencies of this problem are obtained.

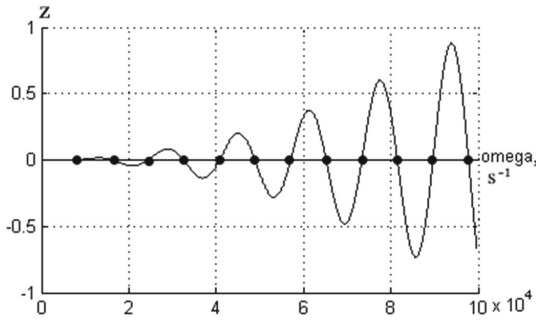


Fig. 5. Frequencies of free oscillations.

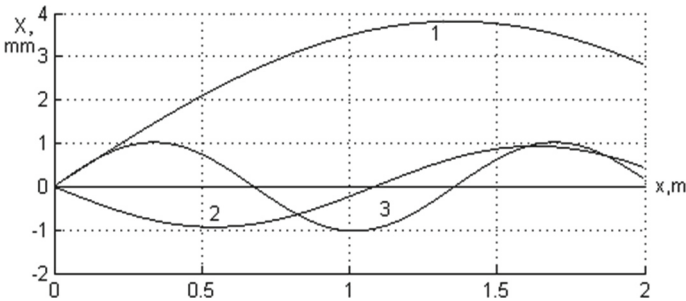


Fig. 6. Frequencies of free oscillations obtained by the analytical method.

Analysis of the curves shows that an increase in the frequency of the driving force is accompanied by an increase in the curvature and undulation of deformations along the rod. At the same time, for small longitudinal deformations, a force with an amplitude f of a significant magnitude was required.

4 Numerical Solution Method

When using representation (5), the basic Eq. (2) will take the form

$$X'' + k^2X = 0, k = \omega/a \tag{12}$$

Similarly, we transform the boundary conditions (3), (4) taking into account the vibrational nature of the force F and obtain.

$$X(0) = 0, EAX'(1) + M\omega^2X(1) = f. \tag{13}$$

At this stage of solving the problem, the mathematical model is represented by Eq. (12) and additional conditions (13).

Let's consider the solution of the previous example by the numerical finite difference method. Let's move from differential operators to finite difference ones. The area of

The zero elements of the matrix are not written out.

Let us solve the previous example by the finite difference method.

Example 3. A steel rod of constant cross-section is given according to Fig. 1 with the parameters: $l = 2 \text{ m}$, $E = 2,1 \cdot 10^{11} \text{ N/m}^2$, $A = 2 \text{ cm}^2$, $\rho = 7800 \text{ kg/m}^3$, $M = 100 \text{ kg}$, $\omega = \{6000, 15000, 24000\} \text{ s}^{-1}$, $f = 10000 \text{ kN}$, $n = 10001$.

Required: calculate the values of $X(x)$ at $x \in [0, l]$ and show three graphs of oscillation amplitudes.

As in the previous case, the capabilities of the Matlab software package were used. In the $x - X$ coordinate system, three curved lines are obtained (Fig. 8) corresponding to Eq. (17) and almost completely coinciding with the curves of Fig. 6.

Such an exact coincidence of the results obtained by analytical and numerical methods guarantees the accuracy and reliability of problem statements and solutions.

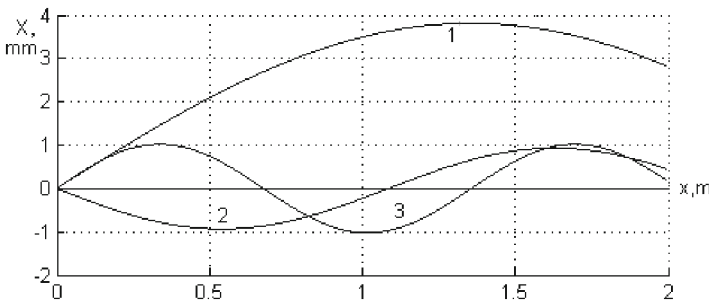


Fig. 8. Graphs of forced oscillations obtained by numerical method

It should be noted that the numerical-graphical method used in Example 2 has the property of universality. Replacing the problem data with other numbers makes it possible to investigate the behavior of a mechanical system in a wide range of parameters.

5 Conclusion

Verification of the results obtained by analytical and numerical methods used in the work is ensured by the exact coincidence of the graphs of cross-section movements along the axis of the rod. The solution of the problem of longitudinal forced vibrations of rods by determining the displacements of sections is significantly simplified when using analytical-graphical and numerical-graphical methods. The use of analytical-graphical and numerical-graphical methods in combination with the capabilities of computer systems provides highly efficient universal solutions to problems.

References

1. Tikhonov, A.N., Samarsky, A.A.: Equations of mathematical physics. In: Publishing House of Moscow State University, p. 799 (1999)
2. Svetlitsky, V.A.: Rod mechanics. In two parts. Part II. Dynamics. In: Higher School, Moscow, p. 304 (1987)
3. Kulterbaev, Kh.P.: Fundamentals of the theory of oscillations. Fundamentals of theory, tasks for homework, examples of solutions. In: Kabardino-Balkarian State University, p.130 (2003)
4. Gurbatov, S.N., Gryaznova, I.Yu., Demin, I.Yu., Klemina, A.V., Kurin, V.V., Pronchatov-Rubtsov, N.V., Lisin, A.A.: Oscillations of mechanical systems with distributed parameters. In: Nizhny Novgorod State University, p. 28 (2021)
5. Nurimbetov, A.U., Dzhunisbekov, M.Sh.: Vibrations of a rectangular rod. In: Pr. of the VIII th. scientific.pract.conf. "Engineering systems - 2015", RUDN, Moscow, pp. 97–102 (2015)
6. Biderman, V.L.: Applied theory of mechanical oscillations. In: Higher school, Moscow, p. 416 (1972)
7. Bolotina, V.V.: Vibrations in technology. In: Handbook in 6 volumes. Volume 1. Oscillations of linear systems. Mashinostroenie, Moscow, p. 352 (1978)
8. Karamansky T.D.: Numerical methods of structural mechanics. In: Stroyizdat, Moscow, p. 436 (1981)
9. Verzhbitsky V.M.: Fundamentals of numerical methods. In: Higher school, Moscow, p. 840 (2002)
10. Samarsky, A.A., Gulin, A.V.: Numerical methods. In: Nauka, Moscow, p. 432 (1989)
11. Kulterbaev, Kh.P., Baragunova, L.A., Shogenova, M.M., Senov, Kh. M.: About a high-precision graphoanalytical method of determination of critical forces of an oblate rod. In: Proceedings 2018 IEEE International Conference "Quality Management, Transport and Information Security, Information Technologies"(IT&QM&IS), In: St. Petersburg. Russia, pp. 794–796 (2018)



On the Unassociated Matrices Number of the n Order and a Given Determinant

Urusbi Pachev^{1,2(✉)} and Rezuan Dokhov²

¹ Kabardino–Balkarian State University, Nalchik, Russia
urusbi@rambler.ru

² North Caucasus Center of Mathematical Research, North Caucasian Federal University, Stavropol, Russia
rezuan.dokhov@yandex.ru

Abstract. The main object of the present work is to derive new relations between the number of n -order non-associated matrices and a determinant N , which can subsequently be put into use. In this study we mainly employ the Hermite triangular form of n -order full matrices and the determinant N . The following new results are obtained in the work:

1. formula for $\sigma_0(n, p_1 \cdots p_k)$ n -order non-associated primitive matrices with non-square determinants values $N = p_1 \cdots p_k$, where p_i are primes;
2. formula for $\sigma_0(n, p^\alpha)$ primitive non-associated n -order matrices $N = p^\alpha$, where p is a prime;
3. the recurrent relations is established for $\sigma_0(n, N)$ by order of matrices considered;
4. an upper estimate for the number of the considered n order matrices and the determinant is obtained close to the precise value of $\sigma(n, N)$ in the case where the canonical expansion of N is not given;
5. the relationship between $\sigma(n, p^\alpha)$ as well as $\sigma_0(n, p^\alpha)$ and the Gaussian coefficients by combinatorics is established.

Keywords: integer matrix · primitive matrix · triangular canonical form · Mobius function · Euler function · unassociated matrices · recurrence relation · canonical decomposition · upper bound

1 Introduction

In connection with the problem of representing integers by ternary quadratic forms, Yu.V. Linnik [1] developed an original method of analytic number theory, which uses non-commutative arithmetic of quaternions and matrices and with its further development, was called the discrete ergodic method.

The method of Yu.V. Linnik is associated with several useful concepts, one of which is the concept of non-associated square matrices of a given order. This concept already proves the main lemma of this method when estimating

from below the number of left-unassociated second-order primitive matrices K_i in matrix decomposition

$$l + L_i = K_i T_i \quad (i = 1, \dots, r),$$

where $SpL_i = 0$, $l = lE$ — is a scalar matrix; $\det K_i = p^s$, p — is a prime number; the left-hand of the matrices K_i is non-associated; $r > c(\det L_i)^{\frac{1}{2}-\varepsilon}$, $\varepsilon > 0$ — is an arbitrarily small number.

In [1] a formula is given for the number of non-associated second order matrices $\sigma_0(n)$ where n is odd.

Later, in [2] the proof of the formula

$$\sigma_0(n) = |n| \cdot \prod_{p|n} \left(1 + \frac{1}{p}\right)$$

is given for any value of the determinant $n \neq 0$, moreover, we employ the Hermite normal form for the matrix ring $M_2(\mathbb{Z})$ as well as the Möbius inversion.

The specified result obtained for $\sigma_0(n)$ you can find in [1–6]. In further studies [7] a formula for $\sigma_0(2, A)$ value of the second order non-associated primitive matrices divisible by a given matrix $A \in M_2(\mathbb{Z})$.

In [8], the concept of associativity is transferred to the ring of indefinite anisotropic quaternions, but for the number of pairwise non-associated quaternions of a given norm, only estimates from below and from above are obtained.

In [9], all the mentioned results were transferred from the matrices ring $M_2(\mathbb{Z})$ to the third order full matrices ring $M_3(\mathbb{Z})$. We note that more complicated case with n order and the results are presented in [10].

In the present work, we obtained the formula for the number of pairwise n -order non-associated primitive matrices $\sigma_0(n, p_1 \cdots p_k)$ with the non-square determinant $p_1 \cdots p_k$, where p_i are distinct primes using the canonical triangular form of non-associated matrices with respect to the $M_n(\mathbb{Z})$ ring.

The formula for the number of n -order non-associated primitive matrices $\sigma_0(n, p^\alpha)$ and the determinant p^α equal to the prime power is also found using the Möbius inversion.

The upper bound for $\sigma(n, N)$ is obtained for the case when the canonical expansion of N is unknown.

Hence a general formula for $\sigma_0(n, N)$ can be easily deduced if we know the canonical expansion of N . The relationship between $\sigma(n, p^\alpha)$ and $\sigma_0(n, p^\alpha)$ is established using the Gaussian coefficient of the above combinatorics.

2 Lemmas About Integer Matrices of the n Order

Here we give auxiliary statements about integer matrices of the n order (in special cases for $n = 2; 3$ they were used, for example, in [7, 9]).

Matrix

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

It has called an integer if its elements are $a_{ij} \in Z$. The ring of integer matrices of order n will be denoted by $M_n(Z)$.

Matrix $\mathcal{E} \in M_n(Z)$ is called invertible, if $\mathcal{E}^{-1} \in M_n(Z)$ and means, $\det \mathcal{E} = \pm 1$.

Matrix $A, A' \in M_n(Z)$ are called associated from the right, if there is an invertible matrix $U \in M_n(Z)$, for which $A' \in AU$, in other cases A and A' will unassociated on the right (similarly, the associativity of matrices on the left is determined).

The associativity of matrices on the right (left) is an equivalence relation that splits the ring $M_n(Z)$ into classes of matrices associated on the right (left), and the number of such classes is finite.

In each class of right-associated matrices from $M_n(Z)$, a single canonical triangular matrix can be selected.

Lemma 1 (on canonical matrix form). *For every non-degenerate matrix $A \in M_n(Z)$, there is a unique matrix associated to it on the right of the form*

$$T = \begin{pmatrix} \delta_1 & \varepsilon_{12} & \varepsilon_{13} & \cdots & \varepsilon_{1n} \\ 0 & \delta_2 & \varepsilon_{23} & \cdots & \varepsilon_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \cdots & \delta_n \end{pmatrix}$$

where

$$\begin{aligned} \delta_1, \delta_2, \dots, \delta_n &> 0; & \delta_1 \cdot \delta_2 \cdots \delta_n &= \Delta, \\ 0 \leq \varepsilon_{1j} &< \delta_1 & (j = 2, \dots, n); \\ 0 \leq \varepsilon_{2j} &< \delta_2 & (j = 3, \dots, n); \\ & \dots & \dots & \dots & \dots & \dots \\ 0 \leq \varepsilon_{n-1,n} &< \delta_{n-1}; \end{aligned}$$

Δ – matrix determinant T .

The proof is given in [[11], Chapter II], where it is carried out using elementary column transformations.

Lemma 2. *All integer matrices of the determinant Δ from the ring $M_n(Z)$ are divided into a finite number of classes of associated matrices, i.e. for the number $\sigma(n, \Delta)$ of pairwise unassociated matrices of the n -th order and the determinant Δ , the formula is valid*

$$\sigma(n, \Delta) = \sum_{\Delta = \delta_1 \cdot \delta_2 \cdots \delta_n} \delta_1^{n-1} \cdot \delta_2^{n-2} \cdots \delta_{n-1},$$

where the sum is taken over all representations of the number Δ in the form of $\Delta = \delta_1 \cdot \delta_2 \cdots \delta_n$.

The proof is based on combinatorial reasoning related to placements with repetitions and using Lemma 1.

Regarding the result related to Lemma 2, see [11–13].

Lemma 3. *For the number of unassociated matrices of the determinant N , the recurrence relation holds*

$$\sigma(n, N) = \sum_{d|N} \sigma\left(n-1, \frac{N}{d}\right) \cdot \frac{N}{d},$$

where n and $n-1$ – the orders of the matrices under consideration; summation is carried out over all natural divisors of the number N .

Proof. Consider the triangular canonical Hermite form for a matrix of order n . We have

$$\begin{pmatrix} d_1 & c_{12} & c_{13} & \cdots & c_{1,n-1} & c_{1n} \\ 0 & d_2 & c_{23} & \cdots & c_{2,n-1} & c_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & d_{n-1} & c_{n-1,n} \\ 0 & 0 & 0 & \cdots & 0 & d_n \end{pmatrix}$$

where

$$\begin{aligned} N &= d_1 \cdot d_2 \cdot \cdots \cdot d_n, \\ 0 &\leq c_{1i} < d_1, \quad 2 \leq i \leq n; \\ 0 &\leq c_{2i} < d_2, \quad 3 \leq i \leq n; \\ &\dots\dots\dots\dots\dots\dots \\ 0 &\leq c_{n-1,n} < d_{n-1}. \end{aligned}$$

By Lemma 2 we have

$$\begin{aligned} \sigma(n, N) &= \sum_{d_1 \cdot d_2 \cdots d_n = N} d_1^{n-1} \cdot d_2^{n-2} \cdot \cdots \cdot d_{n-1} \\ &= \sum_{d_1 \cdot d_2 \cdots d_n = N} (d_1 \cdot d_2 \cdots d_{n-1}) d_1^{n-2} \cdot d_2^{n-3} \cdots d_{n-2} \\ &= \sum_{d_n | N} \frac{N}{d_n} \cdot d_1^{n-2} d_2^{n-3} \cdots d_{n-2} \\ &\quad d_1 \cdot d_2 \cdots d_{n-1} = \frac{N}{d_n} \\ &= \sum_{d_n | N} \frac{N}{d_n} \cdot \sum_{d_1 \cdot d_2 \cdots d_{n-1} = \frac{N}{d_n}} d_1^{n-2} d_2^{n-3} \cdots d_{n-2} \\ &= \sum_{d_n | N} \frac{N}{d_n} \cdot \sigma\left(n-1, \frac{N}{d_n}\right) \\ &= \sum_{d|N} \sigma\left(n-1, \frac{N}{d}\right) \frac{N}{d}. \end{aligned}$$

Lemma 3 is proved.

Remark. A proof of a similar statement in the case of matrices over an abstract Euclidean ring with a multiplicative norm is given in [11]. But here we preferred a proof relating to the more visual case of the ring of integer matrices.

Lemma 4. *If*

$$N = \prod_{p|N} p^{\alpha_p}$$

– canonical decomposition of a natural number N , then

$$\sigma(n, N) = \prod_{p^{\alpha_p}|N} \frac{(p^{\alpha_p+1} - 1)(p^{\alpha_p+2} - 1) \dots (p^{\alpha_p+n-1} - 1)}{(p - 1)(p^2 - 1) \dots (p^{n-1} - 1)}.$$

The result of Lemma 4 is given and used in [14, 15], but there is no proof of the formula for $\sigma(n, N)$ in them, and as far as we know, the proof is still not found in the mathematical literature.

In particular cases, the proof of formulas for $\sigma(n, N)$ for $n = 2; 3$ are given in [2, 9].

3 Main Results on the Number of Nonassociated Primitive Matrices of the n Order

Based on Lemmas 1–4, we obtain the main results of this work on the number of unassociated primitive matrices of the n order given by the determinant N .

First, let’s consider one general auxiliary statement.

Lemma 5. *For the number $\sigma_0(n, N)$ of nonassociated right (left) primitive matrices of the n th order of the norm (determinant) N , the formula is valid*

$$\sigma_0(n, N) = \sum_{d^n|N} \mu(d) \sigma\left(n, \frac{N}{d^n}\right),$$

where $\mu(d)$ – the Mobius function; $\sigma\left(n, \frac{N}{d^n}\right)$ – the number of unassociated matrices in order n and determinant $\frac{N}{d^n}$; in this case, the summation is carried out for all d , for which $d^n | N$.

The proof of this lemma is given in [10].

Lemma 5 allows us to obtain formulas for the number of unassociated primitive matrices of the n order of some special types.

Theorem 1. *For the number $\sigma_0(n, p_1 \cdot p_2 \cdot \dots \cdot p_k)$ of unassociated primitive matrices of the n order with a square-free determinant value, the formula is valid*

$$\sigma_0(n, p_1 \cdot p_2 \cdot \dots \cdot p_k) = \prod_{i=1}^k \frac{p_i^n - 1}{p_i - 1}.$$

Proof. By virtue of Lemma 5, for $N = p_1 \cdot p_2 \cdots p_k$ with various simple factors we will have

$$\sigma_0(n, p_1 \cdot p_2 \cdots p_k) = \sum_{d^n | p_1 \cdot p_2 \cdots p_k} \mu(d) \sigma\left(n, \frac{p_1 \cdot p_2 \cdots p_k}{d^n}\right).$$

Since $d^n \cdot p_1 \cdot p_2 \cdots p_k$, then $d = 1$ by $n > 1$.

Therefore, we have

$$\begin{aligned} \sigma_0(n, p_1 \cdot p_2 \cdots p_k) &= \sigma(n, p_1 \cdot p_2 \cdots p_k) = \sigma(n, p_1) \cdot \sigma(n, p_2) \cdots \sigma(n, p_k) = \\ &= \frac{p_1^n - 1}{p_1 - 1} \cdot \frac{p_2^n - 1}{p_2 - 1} \cdots \frac{p_k^n - 1}{p_k - 1} = \prod_{i=1}^k \frac{p_i^n - 1}{p_i - 1}. \end{aligned}$$

Theorem 1 is proved.

The obtained theorem 1 generalizes one result from [10] relating to the case $k = 1$.

Let us now consider an important case of a special kind of matrix, when its determinant is the power of a prime number, i.e. we are talking about calculating $\sigma_0(n, p^\alpha)$, where p – prime number.

Theorem 2. *For the number $\sigma_0(n, p^\alpha)$ of unassociated primitive matrices of the n order and the determinant p^α , where p is a prime number, the formulas are valid*

$$\begin{aligned} \sigma_0(n, p^\alpha) &= \frac{(p^{\alpha+1} - 1)(p^{\alpha+2} - 1) \cdots (p^{\alpha+n-1} - 1)}{(p - 1)(p^2 - 1) \cdots (p^{n-1} - 1)} - \\ &- \frac{(p^{\alpha-n+1} - 1)(p^{\alpha-n+2} - 1) \cdots (p^{\alpha-1} - 1)}{(p - 1)(p^2 - 1) \cdots (p^{n-1} - 1)} \end{aligned}$$

if $n \leq \alpha$ and

$$\sigma_0(n, p^\alpha) = \sigma(n, p^\alpha),$$

if $n > \alpha$.

Proof. By virtue of Lemma 5 we have

$$\sigma_0(n, N) = \sum_{d^n | N} \mu(d) \sigma\left(n, \frac{N}{d^n}\right).$$

Putting $N = p^\alpha$, where p is a prime number, we have

$$\sigma_0(n, p^\alpha) = \sum_{d^n | p^\alpha} \mu(d) \sigma\left(n, \frac{p^\alpha}{d^n}\right). \tag{1}$$

By the property of the Mobius function we have that from the divisibility of $d^n | p^\alpha$ for $n \leq \alpha$ follows $d = 1; p$. Then and (1) we have

$$\sigma_0(n, p^\alpha) = \sigma(n, p^\alpha) - \sigma(n, p^{\alpha-n}). \tag{2}$$

Applying now Lemma 3 to the right side of equality (2) for $n \leq \alpha$, we will have the following equalities

$$\sigma(n, p^\alpha) = \frac{(p^{\alpha+1} - 1)(p^{\alpha+2} - 1) \dots (p^{\alpha+n-1} - 1)}{(p - 1)(p^2 - 1) \dots (p^{n-1} - 1)}, \tag{3}$$

$$\sigma(n, p^{\alpha-n}) = \frac{(p^{\alpha-n+1} - 1)(p^{\alpha-n+2} - 1) \dots (p^{\alpha-1} - 1)}{(p - 1)(p^2 - 1) \dots (p^{n-1} - 1)}. \tag{4}$$

Subtracting these equalities now, and taking into account equality (2), we get

$$\begin{aligned} \sigma_0(n, p^\alpha) &= \frac{(p^{\alpha+1} - 1)(p^{\alpha+2} - 1) \dots (p^{\alpha+n-1} - 1)}{(p - 1)(p^2 - 1) \dots (p^{n-1} - 1)} \\ &\quad - \frac{(p^{\alpha-n+1} - 1)(p^{\alpha-n+2} - 1) \dots (p^{\alpha-1} - 1)}{(p - 1)(p^2 - 1) \dots (p^{n-1} - 1)}. \end{aligned}$$

The lower equality for $\sigma_0(n, p^\alpha)$ for $n > \alpha$ is obtained as follows. Applying Lemma 5 for $n > \alpha$, $N = p^\alpha$ and $d = p^\beta$ we have

$$\sigma_0(n, p^\alpha) = \sum_{p^{n\beta} | p^\alpha} \mu(p^\beta) \sigma\left(n, \frac{p^\alpha}{p^\beta}\right) = \sigma(n, p^\alpha)$$

(here it was used that $p^{n\beta} | p^\alpha$ is possible only when $\beta = 0$).

Theorem 2 is proved.

In connection with Lemma 3, we consider the question of an analogous recurrence relation for the number $\sigma_0(n, N)$ of unassociated primitive matrices of the n th order and the determinant N .

Theorem 3. *For the number $\sigma_0(n, N)$ of unassociated primitive matrices of the n th order with a square-free value of the determinant N , the relation is valid*

$$\sigma_0(n, N) = \sum_{d|N} \sigma_0\left(n - 1, \frac{N}{d}\right) \frac{N}{d}.$$

Proof. Let $N = p_1 \cdot p_2 \cdot \dots \cdot p_k$ be a square-free number and hence p_1, p_2, \dots, p_k are distinct primes. Using Lemma 5, we will have

$$\begin{aligned} \sigma_0(n, p_1 \cdot p_2 \cdot \dots \cdot p_k) &= \sum_{d^n | p_1 \cdot p_2 \cdot \dots \cdot p_k} \mu(d) \sigma\left(n, \frac{p_1 \cdot p_2 \cdot \dots \cdot p_k}{d^n}\right) \\ &= \sigma(n, p_1 \cdot p_2 \cdot \dots \cdot p_k), \end{aligned}$$

that is, on the same square-free numbers N , the values of the quantities under consideration coincide $\sigma_0(n, N) = \sigma(n, N)$.

$$\begin{aligned} \sigma_0(n, N) &= \sigma_0(n, p_1 \cdot p_2 \cdot \dots \cdot p_k) \\ &= \sum_{d|p_1 \cdot p_2 \cdot \dots \cdot p_k} \sigma\left(n-1, \frac{p_1 \cdot p_2 \cdot \dots \cdot p_k}{d}\right) \frac{p_1 \cdot p_2 \cdot \dots \cdot p_k}{d} \\ &= \sum_{d|p_1 \cdot p_2 \cdot \dots \cdot p_k} \sigma_0\left(n-1, \frac{p_1 \cdot p_2 \cdot \dots \cdot p_k}{d}\right) \frac{p_1 \cdot p_2 \cdot \dots \cdot p_k}{d} \\ &= \sum_{d|N} \sigma_0\left(n-1, \frac{N}{d}\right) \frac{N}{d}, \end{aligned}$$

where N – a square-free number.

Theorem 3 is proved.

Remark 1. The square-free condition of the number N in Theorem 3 is essential, as the following calculations show in the special case

$$\begin{aligned} \sigma_0(3, p^2) &= \frac{(p^4 - 1)(p^3 - 1)}{(p^2 - 1)(p - 1)} = p^4 + p^3 + 2p^2 + p + 1; \\ \sum_{d|p^2} \sigma_0\left(2, \frac{p^2}{d}\right) \frac{p^2}{d} &= \sigma_0(2, p^2) p^2 + \sigma_0(2, p) p + \sigma_0(2, 1) \\ &= p^4 + p^3 + p^2 + p + 1; \end{aligned}$$

that is, the values obtained do not match.

4 An Estimate for the Number of Unassociated Matrices of the n Order

Exact formulas for the number of pairwise unassociated matrices of the n order and a given determinant N do not allow calculating values for $\sigma(n, N)$ when the canonical decomposition of the number N is unknown, and even if such a decomposition is known, the exact formulas for them have a rather cumbersome form. In this regard, the question of obtaining an estimate or an asymptomatic formula for $\sigma(n, N)$ depending on N is of interest.

To derive the upper bound for $\sigma(n, N)$, we need the following auxiliary sentence.

Lemma 6. *For any natural number $n \geq 3$, the estimate is valid*

$$\frac{n}{\varphi(n)} = O(\ln \ln n),$$

where φ – Euler function.

The proof is given in [16], where this property is given as a problem.

Theorem 4. *For the number of unassociated matrices of the n order and the determinant of N the upper bound is valid*

$$\sigma(n, N) = O(N^{n-1} \ln \ln N).$$

Proof. May

$$N = \prod_{p^{\alpha p} \| N} p^{\alpha p}$$

– canonical decomposition of the number N . Then by Lemma 4 we have -

$$\begin{aligned} & \sigma(n, N) \\ = & N^{n-1} \prod_{p^{\alpha p} \| N} \frac{\left(1 - \frac{1}{p^{\alpha p + 1}}\right) \cdot \left(1 - \frac{1}{p^{\alpha p + 2}}\right) \cdots \left(1 - \frac{1}{p^{\alpha p + n - 1}}\right)}{\left(1 - \frac{1}{p}\right) \left(1 - \frac{1}{p^2}\right) \cdots \left(1 - \frac{1}{p^{n-1}}\right)} \\ < & N^{n-1} \prod_{p \| N} \frac{1}{\left(1 - \frac{1}{p}\right) \left(1 - \frac{1}{p^2}\right) \cdots \left(1 - \frac{1}{p^{n-1}}\right)} \\ = & N^{n-1} \prod_{p \| N} \frac{1}{\left(1 - \frac{1}{p}\right)^{n-1} \left(1 + \frac{1}{p}\right) \left(1 + \frac{1}{p} + \frac{1}{p^2}\right) \cdots \left(1 + \frac{1}{p} + \frac{1}{p^2} + \cdots + \frac{1}{p^{n-2}}\right)} \\ < & N^{n-1} \prod_{p \| N} \frac{1}{\left(1 - \frac{1}{p}\right)^{n-1} \cdot \left(1 + \frac{1}{p}\right)^{n-2}} \\ = & \frac{N^n}{\varphi(N)} \cdot \prod_{p \| N} \frac{1}{\left(1 - \frac{1}{p^2}\right)^{n-2}} \\ < & \frac{N^n}{\varphi(N)} \cdot \left\{ \prod_p \left(1 - \frac{1}{p^2}\right)^{-1} \right\}^{n-2} \\ = & \frac{N^n}{\varphi(N)} \cdot \left(\frac{\pi^2}{6}\right)^{n-2} = N^{n-1} \cdot \frac{N}{\varphi(N)} \cdot \left(\frac{\pi^2}{6}\right)^{n-2}. \end{aligned}$$

Applying Lemma 6 now we get

$$\sigma(n, N) = O(N^{n-1} \ln \ln N).$$

Theorem 4 is proved.

5 Conclusion

It will be interesting to give a proof of the formula for $\sigma(n, N)$ indicated in Lemma 4 (in works [14, 15], a result about this value is used, but there are no instructions regarding its proof).

It is also of interest to refine the upper bound for the number of unassociated anisotropic indefinite quaternions of a given norm (see [8]).

The proved theorem 2 allows us to establish a relationship between the values $\sigma(n, p^\alpha)$ and $\sigma_0(n, p^\alpha)$, where p is a prime number, with Gaussian coefficients from [17], namely

$$\sigma(n, p^\alpha) - \sigma_0(n, p^\alpha) = \binom{\alpha - 1}{n - 1}_p,$$

where

$$\binom{\alpha - 1}{n - 1}_p = \frac{(p^{\alpha-1} - 1)(p^{\alpha-2} - 1) \dots (p^{\alpha-n+1} - 1)}{(p^{n-1} - 1)(p^{n-2} - 1) \dots (p - 1)}$$

– gaussian coefficient. Further study of this connection is of interest.

References

1. Linnik, Y.V.: Ergodic Properties of Algebraic Fields. Publishing House of the Leningrad University, Leningrad (1967)
2. Malyshev, A.V., Pachev, U.M.: On the arithmetic of second-order matrices. Notes LOMI Sci. Seminars. **93**, 41–86 (1980)
3. Pachev, U.M.: On the distribution of integer points on some two-cavity hyperboloids. Notes LOMI Sci. Seminars. **93**, 87–141 (1980)
4. Pachev, U.M.: On the number of reduced integer indefinite binary quadratic forms with the condition of divisibility of the first coefficients. Chebyshevsky Collection **4**(3), 92–105 (2003)
5. Pachev, U.M.: Representation of integers by isotropic ternary quadratic forms. Izv. RAS. Ser. math. **70**(3), 167–184 (2006)
6. Pachev, U.M.: On the asymptotics of the number of reduced binary quadratic forms with the condition of divisibility of the first coefficients. Siberian Math. J. **48**(2), 376–388 (2007)
7. Pachev, U.M.: On the number of primitive unassociated matrices of the second order of the determinant n , divisible by a given matrix. Vladikavk. Matem. J. **17**(2), 62–67 (2015)
8. Pachev, U.M., Shakova, T.A.: On the units of the quaternion order of an indefinite anisotropic ternary quadratic form. Chebyshevsky Sat. **20**(4), 270–280 (2019)
9. Dokhov, R.A., Pachev, U.M.: On the number of primitive unassociated matrices of the third order of a given determinant. Chebyshevsky Sat. **22**(5), 129–137 (2021)
10. Pachev, U.M.: On the arithmetic of matrices of the ring of integer matrices of the n th order. Vladikavk. Matem. J. **10**(1), 75–78 (2008)
11. Newmann, M.: Integral Matrices, p. 244. Academie Press, New York (1972)
12. Venkov, V.A.: On the integral invariant of the group of unimodular linear substitutions. Leningr. Un. **144**(93), 3–25 (1952)
13. Tolstikov A.V.: Arithmetic of n -th order matrices over a Euclidean ring. DEP_VINITI No. 6444-V88. Cherepovets 120 (1988)
14. Skubenko, B.F.: On the asymptotics of integer matrices of the n th order and on the integral invariant of the group of unimodular matrices. DAN USSR **153**(2), 290–291 (1963)

15. Skubenko, B.F.: On the distribution of integer matrices and the calculation of the volume of the fundamental domain of a unimodular group of matrices. Tr. MIAN USSR **80**, 129–144 (1965)
16. Vinogradov I.M.: Fundamentals of Number Theory. Publishing House "Lan". (2009)
17. Pachev, U.M.: On algebra and arithmetic of binomial and Gaussian coefficients. Chebyshevsky Sat. **19**(3), 257–269 (2018)



Fast Calculation of Parameters of Parallelepipedal Nets for Integration and Interpolation

N. N. Dobrovol'skii^{1,2(✉)}, N. M. Dobrovol'skii¹, Yu. A. Basalov¹,
and E. D. Rebrov¹

¹ Tula State Lev Tolstoy Pedagogical University, Tula 125 Lenin Avenue,
300026, Russia

nikolai.dobrovolsky@gmail.com

² Tula State University, Tula 92 Lenin Avenue, 300012, Russia

<https://www.mathnet.ru/eng/person77189>

Abstract. This work is devoted to the construction and analysis of search algorithms for two-dimensional parallelepipedal nets with the best parameters. The paper presents algorithms for computing the hyperbolic parameter, Bykovskii sums, and the H -function. Sequences of the best lattices are compared for each parameter. The connection between the lattices parameters (N, a) and the value of the hyperbolic parameter is investigated. In particular, the relationship between the value of a hyperbolic parameter and the representation of the number a/N as a continued fraction is shown. For lattices with an increasing hyperbolic parameter, a conjecture is put forward about the structure of the continued fraction representation of the number a/N .

Keywords: Euler brackets · continued fractions · lattice · hyperbolic lattice parameter · quadrature formula · parallelepipedal nets

1 Introduction

This work is devoted to the construction of a sequence of two-dimensional parallelepipedal nets with the best parameters.

Two-dimensional parallelepipedal nets

$$M(a, N) = \left(\frac{1}{N}, \left\{ \frac{ak}{N} \right\} \right) \quad (k = 0, \dots, N - 1) \quad (1)$$

are used for numerical solution of problems of integration, interpolation, solution of differential and integral equations. These issues are well described in classical works by N. M. Korobov [1, 2], N. S. Bakhvalov [3], K. I. Babenko [4], Hua Loken and Wan Yuan [5] and others.

The quality of parallelepipedal nets is estimated either through the hyperbolic lattice parameter $\Lambda(a, N)$ of the linear comparison $m_1 + am_2 \equiv 0 \pmod N$

$$q(\Lambda(a, N)) = \min(\overline{m}_1, \overline{m}_2), \quad \overline{x} = \max\{1, x\}, \quad (2)$$

or through Bykovsky sums [6] in the formulation [7]

$$S(B(\Lambda(a, N)), \alpha) = \sum_{(x_1, x_2) \in B(\Lambda(a, N))} \frac{1}{(\bar{x}_1 \bar{x}_2)^\alpha}, \tag{3}$$

where α is the smoothness parameter, $B(\Lambda(a, N))$ is the set of relative minima of $\Lambda(a, N)$,

or through the integration error of the boundary functions. For $\alpha = 2$ this is

$$h(x, y) = 3^2(1 - 2\{x\})^2(1 - 2\{y\})^2, \tag{4}$$

for even $\alpha \leq 4$, these are normalized Bernoulli polynomials.

The monograph [8] analyzes in detail the representation

$$H(a, N) = \sum_{(x, y) \in M(a, N)} h(x, y) \tag{5}$$

through the continued fraction of a number

$$\frac{a}{N} = \{0; q_1, q_2, \dots, q_l\}, \quad (a, N) = 1,$$

$$H(a, N) = 1 + \frac{4}{Q_l^2} \left(10 + 5l + \sum_{\lambda=1}^l q_\lambda^2 - \frac{3(P_l^2 + Q_{l-1}^2)}{Q_l^2} + \frac{1}{Q_l} \left(2 \sum_{\lambda=1}^l q_\lambda (Q_\lambda T_{l, \lambda+1} + Q_{\lambda-2} T_{l, \lambda+1} + Q_{\lambda-2} T_{l, \lambda-1}) - 10 \sum_{\lambda=1}^{l-1} Q_{\lambda-1} T_{l, \lambda+1} \right) \right),$$

where q_k are partial quotients, P_k, Q_k are numerators and denominators of convergents ($Q_{-1} = 0, Q_0 = 1$), $T_{k, \nu} = [q_{\nu+1}, \dots, q_k]_{(k-\nu)}$ — Euler bracket (see [9]).

In [7], $q(\Lambda(a, N))$ and $S(B(\Lambda(a, N)))$ are also expressed in terms of the continued fraction.

$$q(\Lambda(a, N)) = \min_{0 \leq m \leq l-1} [q_{m+2}, \dots, q_l]_{(n-m-1)} Q_m, \tag{6}$$

$$S(B(\Lambda(a, N))) = \sum_{m=0}^{l-1} \frac{1}{([q_{m+2}, \dots, q_l]_{(n-m-1)} Q_m)^\alpha}. \tag{7}$$

2 Algorithms for Constructing the Bykovsky Set and Calculating the H-function

The Euler brackets are defined by the recurrence relation

$$[]_{(-1)} = 0, []_{(0)} = 1, [q_1, \dots, q_l]_{(l)} = q_n [q_1, \dots, q_{l-1}]_{(l-1)} + [q_1, \dots, q_{l-2}]_{(l-2)}. \tag{8}$$

The numerators and denominators of convergents have a simple expression through Euler brackets

$$P_l = [q_2, \dots, q_l]_{(l-1)}, \quad Q_l = [q_1, \dots, q_l]_{(l)}. \quad (9)$$

Then we construct the Bykovsky set

$$B(A(a, N)) = \{((-1)^m [q_{m+2}, \dots, q_l]_{(n-m-1)}, Q_m) \mid m = 0, \dots, l-1\} \quad (10)$$

in a loop for $O(l) = O(\ln(N))$ operations.

The library of algorithms, as well as all the results of numerical experiments, are published in POIVS TMK [11].

2.1 Construction of the Bykovsky Set

Input: $q[l] = \{q_0 = 0, q_1, \dots, q_l\}$ — an array of incomplete quotients.

Output: $B[2, l]$ — array of Bykovsky set elements.

Array $Q[l+2]$ is filled with denominators of suitable fractions

$$Q[0] = Q_{-1} = 0, \quad Q[1] = Q_0 = 1, \quad Q[2] = Q_1, \quad \dots, \quad Q[l+1] = Q_l.$$

Array $T[l+2]$ is filled with Euler brackets

$$T[0] = []_{(-1)} = 0, \quad T[1] = []_{(0)} = 1, \quad T[2] = [q_l]_{(1)}, \\ T[3] = [q_l, q_{l-1}]_{(2)}, \dots, T[l+1] = [q_l, \dots, q_1]_{(l)}.$$

Algorithm 1. Construction of the Bykovsky set

$Q[0] = 0, \quad Q[1] = 1$

$T[0] = 0, \quad T[1] = 1$

for $i = 2, i < l + 2, i = i + 1$ **do**

$Q[i] = q[i-1]Q[i-1] + Q[i-2]$

$T[i] = q[2+l-i]T[i-1] + T[i-2]$

end for

for $i = 0, i < l, i = i + 1$ **do**

$B[0, i] = (-1)^i T[l-i]$

$B[1, i] = Q[i+1]$

end for

For simplicity of presentation, the construction is carried out in 2 cycles instead of one.

Note that after constructing the Bykovskii set, the hyperbolic parameter and the Bykovskii sum are calculated in $O(l) = O(\ln(N))$ operations.

Algorithm 2. Calculation of the hyperbolic parameter

```

 $qA = B[0, 0]B[1, 0]$ 
for  $i = 1, i < l, i = i + 1$  do
    if  $qA > |B[0, i]|B[1, i]$  then
         $qA = |B[0, i]|B[1, i]$ 
    end if
end for

```

2.2 Calculation of the Hyperbolic Parameter

Input: $B[2, l]$ — array of Bykovsky set elements.

Output: qA is a hyperbolic parameter.

Note that to speed up the calculation of the hyperbolic parameter, one can construct the Bykovsky set without sign alternation.

2.3 Calculation of the Bykovsky Sums

Input: $B[2, l]$ — array of Bykovsky set elements, α — smoothness parameter.

Output: SB is a hyperbolic parameter.

Algorithm 3. Calculation of the Bykovsky sum

```

 $SB = 0$ 
for  $i = 0, i < l, i = i + 1$  do
     $SB = SB + 1/(|B[0, i]|B[1, i])^\alpha$ 
end for

```

Note that to speed up the calculation of the Bykovsky sum, one can construct the Bykovsky set without alternating signs.

2.4 H-function Calculation

Input: $q[l] = \{q_0 = 0, q_1, \dots, q_l\}$ — an array of incomplete quotients.

Output: H — the value of the function $H(a, N)$ given by the equality (5).

Array $Q[1+2]$ is filled with denominators of suitable fractions

$$Q[0] = Q_{-1} = 0, \quad Q[1] = Q_0 = 1, \quad Q[2] = Q_1, \quad \dots, \quad Q[l+1] = Q_l.$$

Array $T[1+2]$ is filled with Euler brackets

$$T[0] = \llbracket_{(-1)} = 0, \quad T[1] = \llbracket_{(0)} = 1, \quad T[2] = [q_l]_{(1)},$$

$$T[3] = [q_l, q_{l-1}]_{(2)}, \dots, T[l+1] = [q_l, \dots, q_1]_{(l)}.$$

Note that P_l is $T[l]$, and $T_{l,\nu} = T[l - \nu + 1]$.

Algorithm 4. Function $H(a, N)$ evaluation

```

 $Q[0] = 0, Q[1] = 1$ 
 $T[0] = 0, T[1] = 1$ 
for  $i = 2, i < l + 2, i = i + 1$  do
     $Q[i] = q[i - 1]Q[i - 1] + Q[i - 2]$ 
     $T[i] = q[2 + l - i]T[i - 1] + T[i - 2]$ 
end for
 $H = 0, S = 0$ 
for  $i = 0, i < l - 1, i = i + 1$  do
     $H = H + Q[i + 1]T[l - i - 1]$ 
end for
 $H = 10H$ 
for  $i = 0, i < l, i = i + 1$  do
     $S = S + q[i + 1](Q[i + 2]T[l - i - 1] + Q[i]T[l - i - 1] + Q[i]T[l - i + 1])$ 
end for
 $S = 2S$ 
 $H = S - H$ 
 $H = H/Q[l + 1]$ 
 $H = H + 10 + 5l - 3(T[l]^2 + Q[l]^2)/Q[l + 1]^2$ 
for  $i = 0, i < l - 1, i = i + 1$  do
     $H = H + q[i + 1]^2$ 
end for
 $H = 4H/(5Q[l + 1]^2) + 1$ 

```

3 Increasing Sequence of Hyperbolic Parameters

The algorithm for calculating the hyperbolic parameter $q(\Lambda(a, N))$ in $O(\ln(N))$ operations allows us to construct the following sequence by enumeration

$$\begin{aligned}
 GP_MAX &= \{GP(a_1, N_1) = q(\Lambda(3, 8)) < GP(a_2, N_2) < \dots < GP(a_k, N_k) \dots\}, \\
 \forall N < N_k, a < N, (a, N) = 1 : GP(a_k, N_k) &= q(\Lambda(a_k, N_k)) > q(\Lambda(a, N)), \\
 \forall a < a_k : GP(a_k, N_k) &= q(\Lambda(a_k, N_k)) > q(\Lambda(a, N_k)).
 \end{aligned}$$

To reduce the enumeration, we use the properties of the hyperbolic parameter

$$q(\Lambda(a, N)) = q(\Lambda(N - a, N)), \quad q(\Lambda(a, N)) \leq a. \tag{11}$$

Thus, we reduce the enumeration of a_k to the set $\{GP(a_{k-1}, N_{k-1}), \dots, [N/2]\}$.

The interval [8, 1346269] contains 1413 members of the sequence GP_MAX . The best of them with respect to the ratio of the hyperbolic parameter to the lattice modulus is the subsequence that defines the lattices

$$\Lambda(F_{n-2}, F_n), \quad q(\Lambda(F_{n-2}, F_n)) = F_n, \tag{12}$$

where F_n — n Fibonacci number. Moreover, F_{n-2}/F_n are convergents to the quadratic irrationality $(3 - \sqrt{5})/2$. The use of this sequence as optimal coefficients was first proposed by N. S. Bakhvalov in 1959 [3].

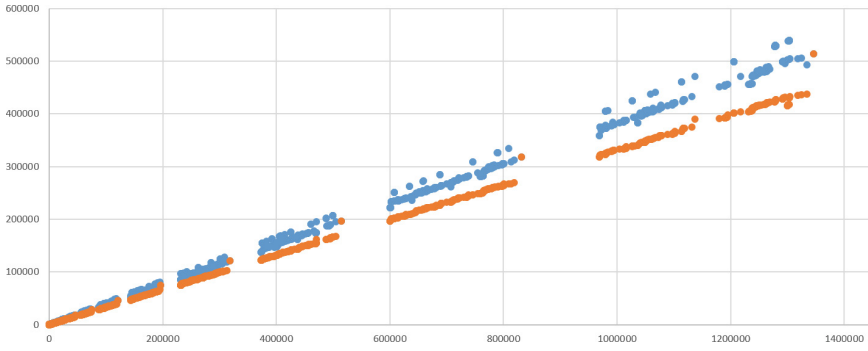


Fig. 1. The first 1413 members of the GP_MAX sequence, along the horizontal axis N_k , in blue, the optimal coefficient a_k , in orange — $GP(a_k, N_k)$.

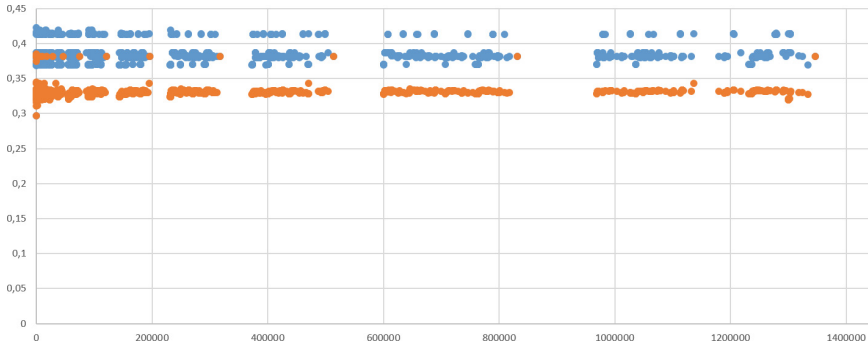


Fig. 2. The first 1413 terms of the sequence GP_MAX , along the horizontal axis N_k , in blue the ratio of the optimal coefficient to the modulus of the lattice a_k/N_k , in orange — $GP(a_k, N_k)/N_k$.

Also, other subsequences of convergent fractions of quadratic irrationalities are distinguished by the form of the continued fraction a_k/N_k . Continued fractions of the form $\{0; 2, 2, 2, 2\}$ define a sequence of convergents converging to $\sqrt{2} - 1$.

4 Decreasing Sequence of Bykovsky Sums

Using the calculation algorithm $S(B(\Lambda(a, N)), \alpha)$ enumeration a from 1 to $[N/2]$, construct the following sequence

$$\begin{aligned}
 SB_MIN_2 = \{ & SB(a_1, N_1) = S(B(\Lambda(3, 8)), 2) > SB(a_2, N_2) > \dots > \\
 & > SB(a_k, N_k) > \dots \}, \\
 \forall N < N_k, a < N, (a, N) = 1 : & SB(a_k, N_k) = S(B(\Lambda(a_k, N_k)), 2) <
 \end{aligned}$$

$$< S(B(\Lambda(a, N)), 2),$$

$$\forall a < a_k : SB(a_k, N_k) = S(B(\Lambda(a_k, N_k)), 2) < S(B(\Lambda(a, N_k)), 2).$$

The interval [8, 450000] contains 4216 members of the sequence *SB_MIN_2*.

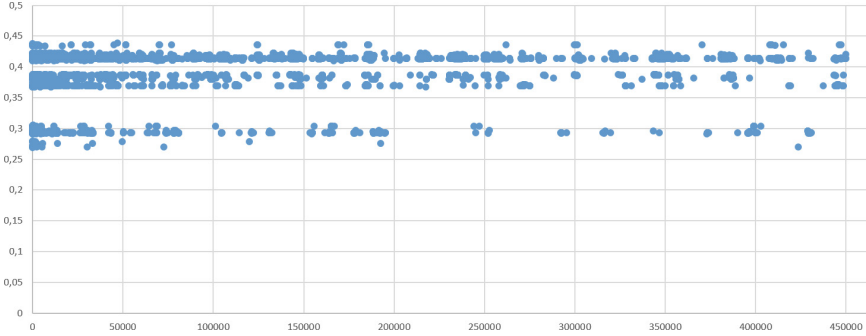


Fig. 3. The first 4216 members of the sequence *SB_MIN_2*, along the horizontal axis N_k , along the vertical ratio a_k/N_k .

It is easy to see that the sequences *SB_MIN_2* and *GP_MAX* do not match. This is illustrated by the following example:

$$N = 229, \quad A = 95,$$

$$B(\Lambda(95, 229)) = \left\{ \begin{array}{cccccc} 95 & -39 & 17 & -5 & 2 & -1 \\ 1 & 2 & 5 & 12 & 41 & 94 \end{array} \right\},$$

$$q(\Lambda(95, 229)) = 60, \quad S(B(\Lambda(95, 229)), 2) = 0.00095324\dots$$

$$N = 233, \quad A = 89,$$

$$B(\Lambda(89, 233)) = \left\{ \begin{array}{ccccccc} 89 & -55 & 34 & -21 & 13 & -8 & 5 & -3 & 2 & -1 \\ 1 & 2 & 3 & 5 & 8 & 13 & 21 & 34 & 55 & 89 \end{array} \right\},$$

$$q(\Lambda(89, 233)) = 89, \quad S(B(\Lambda(89, 233)), 2) = 0.000976334\dots$$

5 Decreasing Sequence of H Functions

Using the algorithm for calculating $H(a, N)$ by iterating a from 1 to $[N/2]$, we construct the following sequence

$$H_MIN = \{H(a_1, N_1) = H(3, 8), 2) > H(a_2, N_2) > \dots > H(a_k, N_k) > \dots\}$$

$$\forall N < N_k, \quad a < N, \quad (a, N) = 1 : H(a_k, N_k) < H(a, N)$$

$$\forall a < a_k : H(a_k, N_k) < H(a, N_k).$$

The interval [8, 450000] contains 7888 members of the sequence *H_MIN*.

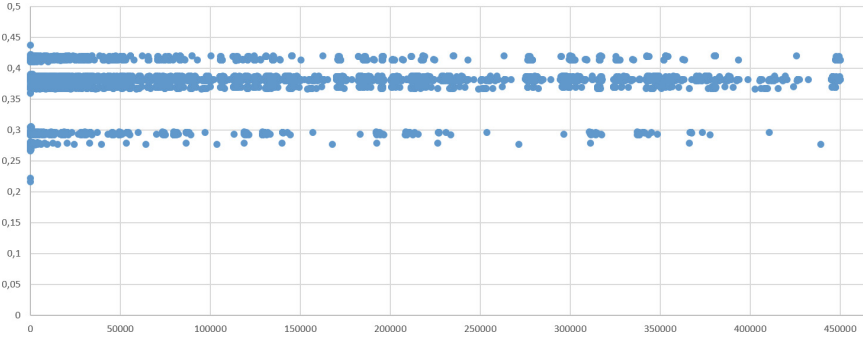


Fig. 4. The first 7888 members of the sequence H_MIN , along the horizontal axis N_k , along the vertical ratio a_k/N_k .

It is easy to see that the sequences H_MIN and GP_MAX do not match. This is illustrated by the following example:

$$\begin{aligned}
 N &= 271443, \quad A = 75025, \\
 q(\Lambda(75025, 271443)) &= 75025, \quad S(B(\Lambda(75025, 271443)), 2) = 10^{-9} \cdot 1.7489 \dots \\
 H(75025, 271443) - 1 &= 10^{-9} \cdot 1.6626 \dots \\
 N &= 271684, \quad A = 103291, \\
 q(\Lambda(103291, 271684)) &= 90208, \quad S(B(\Lambda(103291, 271684)), 2) = 10^{-9} \cdot 1.6456 \dots \\
 H(103291, 271684) - 1 &= 10^{-9} \cdot 1.7085 \dots
 \end{aligned}$$

6 Construction of Sequences by the Form of a Continued Fraction

The formula for the value of the hyperbolic lattice parameter allows us to prove two important estimates for the hyperbolic parameter.

Lemma 1. *If among the incomplete quotients $\frac{a}{N} = \{q_0; q_1, \dots, q_n\}$ there is an incomplete quotient $q_m \geq c$, then for the hyperbolic parameter $q(\Lambda(a, N))$*
 $q(\Lambda(a, N)) \leq \frac{N}{c}$.

Proof. Indeed, according to the formula for the hyperbolic lattice parameter, we have:

$$q(\Lambda(a, N)) \leq Q_{m-1}[q_{m+1}, \dots, q_n]_{(n-m)}.$$

It is well known (see [10]) the following identity

$$\begin{aligned}
 N &= [q_1, \dots, q_n]_{(n)} \\
 &= [q_1, \dots, q_m]_{(m)} [q_{m+1}, \dots, q_n]_{(n-m)} + [q_1, \dots, q_{m-1}]_{(m-1)} [q_{m+2}, \dots, q_n]_{(n-m-1)}.
 \end{aligned}$$

It follows from it that

$$\begin{aligned} \Lambda(a, N) &\leq Q_{m-1}[q_{m+1}, \dots, q_n]_{(n-m)} \\ &= \frac{NQ_{m-1}[q_{m+1}, \dots, q_n]_{(n-m)}}{[q_1, \dots, q_m]_{(m)}[q_{m+1}, \dots, q_n]_{(n-m)} + [q_1, \dots, q_{m-1}]_{(m-1)}[q_{m+2}, \dots, q_n]_{(n-m-1)}} \\ &= \frac{N}{\frac{[q_1, \dots, q_m]_{(m)}}{Q_{m-1}} + \frac{[q_{m+2}, \dots, q_n]_{(n-m-1)}}{[q_{m+1}, \dots, q_n]_{(n-m)}}} \leq \frac{N}{\frac{Q_{m-1}q_m + Q_{m-2}}{Q_{m-1}}} = \frac{N}{q_m + \frac{Q_{m-2}}{Q_{m-1}}} \leq \frac{N}{c}. \end{aligned}$$

Lemma 2. *If $1 \leq a < N$, $(a, N) = 1$ and the hyperbolic parameter $q(\Lambda(a, N))$ satisfies the inequality $q(\Lambda(a, N)) > \frac{4N}{13}$, then $q_0 = 0$ and all other incomplete partial expansions into continued fractions $\frac{a}{N} = \{q_0; q_1, \dots, q_n\}$ take values only 1 and 2.*

Proof. Indeed, it follows from the condition and formula for the hyperbolic parameter that

$$Q_m[q_{m+2}, \dots, q_n]_{(n-m-1)} > \frac{4N}{13} \quad (m = 0, \dots, n - 1).$$

Because

$$Q_m[q_{m+2}, \dots, q_n]_{(n-m-1)} = \frac{N}{\frac{Q_{m+1}}{Q_m} + \frac{[q_{m+3}, \dots, q_n]_{(n-m-2)}}{[q_{m+2}, \dots, q_n]_{(n-m-1)}}},$$

then

$$\frac{Q_{m+1}}{Q_m} + \frac{[q_{m+3}, \dots, q_n]_{(n-m-2)}}{[q_{m+2}, \dots, q_n]_{(n-m-1)}} < 3 + \frac{1}{4}.$$

Because

$$\begin{aligned} &\frac{Q_{m+1}}{Q_m} + \frac{[q_{m+3}, \dots, q_n]_{(n-m-2)}}{[q_{m+2}, \dots, q_n]_{(n-m-1)}} \\ &= q_{m+1} + \frac{1}{q_m + \frac{1}{\dots + \frac{1}{q_2 + \frac{1}{q_1}}}} + \frac{1}{q_{m+2} + \frac{1}{\dots + \frac{1}{q_{n-1} + \frac{1}{q_n}}}}, \end{aligned}$$

then $\max_{m=1, \dots, n} q_m \leq 3$. Let $\max_{m=1, \dots, n} q_m = 3$, then

$$\begin{aligned} &\max_{m=0, \dots, n-1} \frac{Q_{m+1}}{Q_m} + \frac{[q_{m+3}, \dots, q_n]_{(n-m-2)}}{[q_{m+2}, \dots, q_n]_{(n-m-1)}} \\ &\geq \max \left(q_1 + \frac{1}{q_2 + 1}, \max_{m=1, \dots, n-2} \left(q_{m+1} + \frac{1}{q_m + 1} + \frac{1}{q_{m+2} + 1} \right), \right. \\ &\quad \left. q_n + \frac{1}{q_{n-1} + 1} \right) \geq \max_{m=1, \dots, n} q_m + \frac{1}{4}. \end{aligned}$$

But this implies that $q_{m+1} < 3$ for $m = 0, \dots, n - 1$.

The lemmas proved above allow us to formulate a conjecture for constructing sequences for large N .

Hypothesis.

The sequence GP_MAX for $N \geq 377$ includes a_k, N_k such that a_k/N_k expands into a continued fraction of 1 and 2.

7 Conclusion

According to [12,13], a sequence GP_MAX of the best parallelepipedal nets for 2D Fourier interpolation is constructed, which allows solving, among other things, multidimensional acoustic problems.

The constructed sequences of parallelepipedal lattices allow us to raise the question of singling out common subsequences in them, consisting of convergent fractions to some quadratic irrationalities. The issue of identifying all such subsequences will be the subject of our subsequent work. This approach will allow us to build more efficient algorithms for finding the best optimal coefficients.

Acknowledgements. The study was supported by a grant from the Russian Foundation for Basic Research (project no. 19-41-710005_r-a) with the financial support of a grant from the government of the Tula region (agreement no. DS/306 dated November 16, 2021)

References

1. Korobov, N.M.: Teoretiko-chislovye metody v priblizhennom analize [Number-theoretic methods in approximate analysis]. Fizmat-giz, Moscow, Russia (1963)
2. Korobov, N.M.: Teoretiko-chislovye metody v priblizhennom analize [Number-theoretic methods in approximate analysis], 2nd edn. MTSNMO, Moscow, Russia (2004)
3. Bakhvalov, N.S.: On approximate computation of multiple integrals. Vestnik Moskov. Univ. Ser. I Mat. Mekh. **4**, 3–18 (1959)
4. Babenko, K.I.: Osnovy chislennogo analiza [Fundamentals of numerical analysis]. Nauka, Moscow, Russia (1986)
5. Loo Keng, H., Yuan, W.: Applications of Number Theory to Numerical Analysis. Springer-Verlag, Berlin (1981). <https://doi.org/10.1007/978-3-642-67829-5>
6. Bykovskii, V.A.: On the error of number-theoretic quadrature formulas. Dokl. Math. **67**(2), 175–176 (2003)
7. Kormacheva, A.N., Dobrovol'skii, N.N., Rebrova, I.Y., Dobrovol'skii, N.M.: On the hyperbolic parameter of a two-dimensional lattice of comparisons. Chebyshevskii Sbornik. **22**(4), 168–182 (2021)
8. Vronskaya, G.T., Dobrovol'skii, N.N.: Deviations of flat grids. Monograph edited by N. M. Dobrovol'skii. Tula. 193 (2012)
9. Gauss, K.F.: Proceedings on the theory of numbers. In: Translation of B. B. Demyanov, general edition I. M. Vinogradova, comments B. N. Delone., p. 978. Publishing House of the USSR Academy of Sciences, Moscow (1959)
10. Sushkevich, A.K.: Number theory - Kharkiv: From the Kharkiv State University named after A.M. Gorky, 204 (1956)
11. Fast calculation of parameters of parallelepipedal nets for integration and interpolation. POIVS (2022). <http://poivs.tspu.ru/ru/Count/TMK/HFunctionCalc>
12. Dobrovol'skii, N.M., Yesayan, A.R., Andreeva, O.V., Zaitseva, N.V.: Multidimensional number-theoretic Fourier interpolation [Mnogomernaya teoretiko-chislovaya Fur'e interpoliatsiya]. Chebyshevskii sbornik **5**(1), 122–143 (2004)
13. Dobrovol'skii, N.N., Skobel'tsyn, S.A., Tolokonnikov, L.A., Larin, N.V.: About application of number-theoretic grids in problems of acoustics. Chebyshevskii sbornik **22**(3), 368–382 (2021)



Modeling of the Adjustable DC Voltage Source for Industrial Greenhouse Lighting Systems

Vladimir Samoylenko[✉] , Vladimir Fedorenko , and Nikolay Kuchеров 

North-Caucasus Federal University, Stavropol, Russian Federation
vvsamoilenko@ncfu.ru

Abstract. The article considers issues related to the implementation of energy-efficient power supply modes for high-pressure discharge lamps (HID lamps). In comparison with the standard AC power supply mode, the proposed DC power supply approach with a cyclic change in the polarity of the supply voltage allows to increase the light output by up to 20%. The purpose of the article is to develop a mathematical model of a regulated DC source necessary for controlling the light flux during operation. The developed model reflects the dependence of the effective value of the load voltage on the ratio of the load resistance to the internal resistance of the power supply, the conditional frequency and the fill capacity factor when powered from a single-half-period rectifier. The presented model makes it possible to optimize the mode of changing the electrical parameters of the electronic ballast when regulating the light flux of high-pressure discharge lamps. The model was tested by simulation modeling in the Multisim environment, which showed the high efficiency of the obtained dependencies. The use of the developed mathematical tools in the design of electronic ballast devices will make it possible to effectively control the light output. This mode can be used for lighting industrial greenhouse complexes.

Keywords: HID lamps · electronic ballast · optical radiation · light dimming systems · multisim

1 Introduction

High-pressure gas-discharge lamps are one of the most popular light sources in industrial greenhouses [1, 2]. The power supply mode of these lamps from the AC mains of industrial frequency through a reactive ballast is characterized by a pulsation of the light flux due to the extinction and re-ignition of the gas discharge in each half-cycle of the supply voltage. The power supply of lamps through frequency converters (102...105 Hz), which are part of the electronic ballast equipment, partially or completely eliminates these phenomena. But there are disadvantages, among which it should be noted the danger of acoustic resonance, leading to the destruction of the lamp, and powerful electromagnetic interference. It is of scientific and practical interest to study the power supply modes of HID lamp from DC sources to increase the light output of these light sources. [3].

The correct operation of gas-discharge lamps requires the inclusion in their power circuit of electronic ballast equipment designed to ensure ignition, build and normal operation of the lamp. Ballasts are divided into standard electromagnetic [4] and electronic devices [5–7].

The traditional HID lamp power supply system from an alternating voltage network of 220V 50 Hz contains a current-limiting ballast (inductive, capacitive or combined), connected in series with a lamp, and an ignition device that generates high-voltage pulses to ignite the discharge. Such a HID lamp power supply mode is characterized by a pulsation of the light flux due to the extinction and re-ignition of the gas discharge in each half-cycle of the supply voltage, which in itself is a kind of “stress” for the lamp itself [8].

The above-described disadvantages of the electromagnetic ballast significantly increase operating costs and do not allow to fully disclose all the possibilities of lighting using discharge lamps. In addition, the Directive of the European Commission No. 2000/55/EC suppress the sale and use of electromagnetic ballast for the purpose of widespread introduction of electronic ballasts in the EU countries [9]. Electronic ballast containing a frequency converter is of interest, the output of such ballast is connected to the discharge lamp through the first transformer winding connected in series. A capacitor is connected in parallel to the discharge lamp [10]. The disadvantage of the proposed technical solution is low efficiency due to the high current load of the frequency converter. In addition, a resonant frequency is used at starting, which leads to prolonged 3–4-fold overcurrent of power semiconductor devices and reduces the reliability of the device.

The authors [11] propose a method for powering gas-discharge lamps, implemented by a device containing a rectifier with a doubling of the rectified voltage and a capacitive ballast. The disadvantage of this method is its low efficiency when powered by a direct current of the HID lamp, since in this case there is a so-called phenomenon of gas discharge plasma stratification in the lamp burner, accompanied by a significant decrease in the light flux.

The influence of the power supply mode from a DC source was investigated by Tai-Her Yang [12], who proposed a method for powering gas discharge lamps, implemented by a device containing a rectifier power board with a falling volt-ampere characteristic connected to an industrial network, connected to a polarity reversal switch controlled by a time-setting timer, consisting in the fact that the gas-discharge lamp is switched on in a DC circuit and periodically change the polarity of the supply voltage to the opposite according to fixed time settings. The use of a rectifier power supply with a falling volt-ampere characteristic makes it possible to abandon the use of a ballast element in the sodium discharge lamp power supply circuit. However, this method of supplying gas-discharge lamps has a similar disadvantage, because it does not eliminate the possibility of long-term, i.e. commensurate with the switching period of the supply voltage, modes of operation of the sodium discharge lamp during plasma stratification.

The transformer-free AC limiter for lamp power, proposed in the patent Lu C.-C. [10], consists of a bridge rectifier on 4 diodes, the diagonal of the alternating current of which, through a capacitor with a capacity proportional to the load current, is connected to the supply network, and the lamp is connected to the diagonal of the DC rectifier, and to

reduce light pulsations, an inductive choke is connected in series with the capacitor. The disadvantage of this scheme is that the gas discharge lamp serves as an active nonlinear element, therefore distortions in the form of higher harmonic components occur in the circuit. In this scheme, the most active, having up to 25% content, is the third harmonic, which is in antiphase with the main current wave, which is expressed in the appearance of current pauses, a decrease in the light output of the lamp. The current passing through the burner of the gas discharge lamp immediately after ignition exceeds the steady-state operating current of the lamp by almost twice, therefore, increased depreciation of the electrodes is observed due to the phenomenon of cataphoresis.

The above-mentioned studies have shown that when the HID lamp is powered by direct current, it is possible to increase the light output of the light sources used. However, the technical solutions discussed above do not allow to efficiently produce HID lamp power. In this connection, the purpose of this article is to develop a mathematical model of a regulated DC source and test it by simulation modeling in the Multisim environment.

2 Materials and Methods

In accordance with the goal for the development and verification of the mathematical model, the decomposition of tasks was carried out: a mathematical model based on operator equations was developed and its verification was carried out in the Multisim simulation environment.

2.1 Mathematical Model of an Adjustable DC Source for High-Pressure Discharge Lamps

The DC power source in question for the HID lamp (Fig. 1) must perform a number of functions, among which it should be noted:

- output voltage regulation;
- ensuring the ignition of the discharge at the time of lamp start-up;
- obtaining a falling voltage characteristic that allows to reduce or completely eliminate the resistance of the ballast.

One of the main tasks in the design of this source is to choose the value of the capacitance of the capacitor C (Fig. 1). It should be taken into account that the smaller this capacitance, the greater the voltage ripple coefficient on the lamp, reaching 100% in the absence of a capacitor. In this mode, the HID lamp will not work in a DC circuit, but a unipolar pulsating with a doubled frequency with the appearance of non-current pauses. On the other hand, if the capacitance of the capacitor C is significant, then the voltage ripple on the lamp will almost disappear. But at the same time, in steady-state mode, the value of the alternating current flowing through the choke coil will also decrease, and it loses the properties of the reactive ballast. Accordingly, the task of finding the optimal value of capacity C is relevant.

To begin with, we simplify the task and present the process of single-half-period detection Fig. 2 (a). As a rule, information is given in the literature for cases of detecting harmonic signals or a rectangular wave (meander). In our case, rectangular pulses with

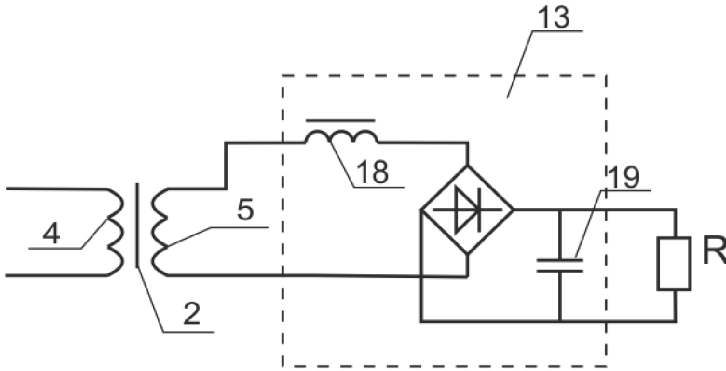


Fig. 1. HID lamp power supply model

amplitude E , duration t_f , and repetition period T are detected. Such a signal is produced by a high-frequency inverter, which serves as a power source for a single-half-period rectifier. Therefore, obtaining a refined model that allows expressing the effective value of the load voltage through the parameters of the circuit and the detected signal is of interest from the standpoint of the goal. There are two nonlinear elements in this scheme: a diode VD and a gas discharge lamp represented by a load resistance R . Taking into account these circumstances, the solution of the task becomes difficult. Therefore, in the first approximation, we assume a linear load.

In the study of diode circuits operating in the detection mode of large amplitudes, separate consideration of the processes of charge and discharge of capacitor C is allowed, i.e. the well-known method of conjugation of intervals is used for piecewise linear approximation of the volt-ampere characteristic of a diode by two linear segments.

In this case, the charge and discharge of the reservoir capacitor C can be carried out using the circuits shown in Fig. 2 (b) and Fig. 2 (c), respectively, where the functions of the diode are assigned to the SA key, which, with a frequency $f = \frac{1}{T}$, connects a source with a constant EMF E and an internal resistance R_B , including the direct resistance of the diode, for a time t_f . This approach allows us to use the law of commutation in the study of the scheme in question. When SA is closed, the capacitor C is charged via R_B , and the voltage $U_1(t)$ on its plates increases exponentially (Fig. 3). When SA is opened, the capacitor C is discharged through R , and the voltage on it decreases by another exponent $U_2(t)$. The reverse resistance of the diode can be neglected, because in this case its value is many times higher than the resistance R , which makes it possible to consider the SA key ideal.

In the transient process, at the moments of time $nT + t_f$, where $n = 0, 1, 2, \dots$, i.e. at the moments of opening SA, the voltage on the capacitor $U_1(t) = U_1(nT + t_f)$ reaches the highest U_{max} value, which determines non-zero initial conditions for the subsequent discharge process in Fig. 3 (a). At the moments of time $t = nT$, i.e. at the moments of SA closure, the voltage $U_2(t) = U_2(nT)$ on the capacitor was the lowest and its value U_{min} will determine the initial conditions for the subsequent charge of the capacitor when SA is closed.

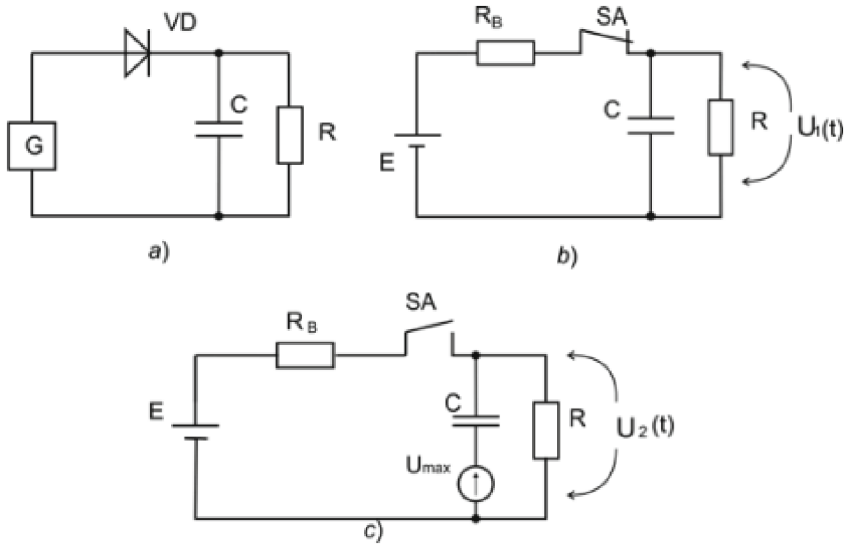


Fig. 2. Single-half-period detector and equivalent circuits: a) the basic circuit of a single-half-period detector; b) an equivalent circuit during the charge of the filter capacitor C; c) an equivalent circuit during its discharge.

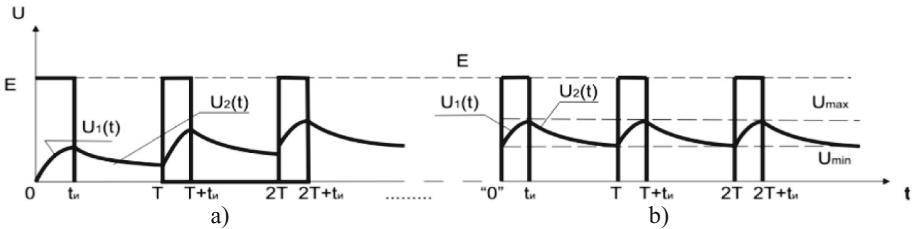


Fig. 3. Transient (a) and time-independent process (b) processes

In a particular case, if the capacitor is completely discharged, then the transient process immediately becomes time-independent, since each subsequent charge of the capacitor will begin at zero initial conditions. Theoretically, the transition process (especially exponential) lasts indefinitely. Practically, the time of this process ends quickly, and a steady-state mode occurs for which:

$$\begin{aligned}
 U_1(nT + t_l) &= U_2(nT + t_l) = U_{\max}, \\
 U_1(nT) &= U_2(nT) = U_{\min}.
 \end{aligned}
 \tag{1}$$

In this case, the values of U_{\min} and U_{\max} will become non-zero initial conditions for the charge and, accordingly, the discharge of the capacitor. Let us conditionally transfer the beginning of the countdown of time t to the new coordinate “O” Fig. 3 (b) from which the process of charging the capacitor begins in the already time-independent mode. To find the function $U_1(t)$, it is easier to use the operator method for solving linear

differential equations with non-zero initial conditions, since the transfer function of a four-pole formed with a closed SA Fig. 2 (b):

$$W(p) = \frac{U_1(p)}{E(p)} = \frac{k}{k\tau_z p + 1} \quad (2)$$

where τ_z is the charge time constant ($\tau_z = R_B C$); $k = \frac{1}{1+\alpha}$ while $\alpha = \frac{R_B}{R}$.

If we know the transfer function of the link under study, then we can immediately record the Laplace image of the desired solution, i.e. the transition function $U_1(p)$ under the action of a step-like action: $E(p) = \frac{E}{p}$:

$$U_1(p) = \frac{k(p)E(p)}{D(p)} + \frac{G(p)}{D(p)} \quad (3)$$

where $D(p)$ is the proper operator of a quadrupole, i.e. the denominator of its transfer function; $K(p)$ is the impact operator; $G(p)$ is a polynomial of p reflecting the influence of non-zero initial values $U_1(0)$ and its derivatives $U'_1(0)$, $U''_1(0)$, ...

In our case, taking into account (1) and (2):

$$D(p) = k\tau_z p; \quad k(p) = k; \quad E(p) = \frac{E}{p};$$

$$U_1(0) = U_{\min}; \quad U'(0) = U''_1(0) = \dots U^n(0) = 0$$

$$G(p) = k\tau_z U_{\min}.$$

Then:

$$U_1(p) = \frac{kE}{(k\tau_z p + 1)p} + \frac{U_{\min} k\tau_z}{k\tau_z p + 1}. \quad (4)$$

To find the original of the transition function in the image (3), the corresponding Laplace transform tables can be used:

$$L^{-1}[U_1(p)] = U_1(t) = kE \left(1 - e^{-\frac{t}{k\tau_z}}\right) + U_{\min} e^{-\frac{t}{k\tau_z}}. \quad (5)$$

In the resulting solution (5), the initial value of U_{\min} is unknown, which in turn can be found from condition (1) for $n = 1$, i.e. for $t = T$:

$$U_{\min} = U_2(nT) = U_2(T).$$

The function $U_2(t)$ itself is found from the discharge condition of the capacitor C through the resistance R in the two-pole (Fig. 2, c) when the key SA is opened at time t_1 :

$$U_2(t) = U_{\max} e^{-\frac{t-t_1}{\tau_p}}; \quad (6)$$

where τ_p is a discharge time constant ($\tau_p = RC$).

$$\text{if } t = T : U_2(T) = U_{\min} = U_{\max} A, \quad (7)$$

where, for the compactness of the record we accept: $A = e^{-\frac{t_1}{\tau_p}}$.

As $U_{max} = U_1(t_1)$ taking into account (5):

where also, for the compactness of the record we accept

$$U_{max} = kE(1 - B) + U_{min}B, \quad (8)$$

$$B = e^{-\frac{t_1}{k\tau_z}}.$$

Substituting (7) into (6), we find U_{min} :

$$U_{min} = kE \frac{(1 - B)A}{1 - AB}. \quad (9)$$

Substituting (8) into (7), we find $U_1(t)$:

$$U_1(t) = kE \left[1 - \frac{1 - A}{1 - AB} e^{-\frac{t}{k\tau_z}} \right] \quad (10)$$

We define $U_2(t)$:

$$U_2(t) = kE \left[1 - \frac{(1 - A)B}{1 - AB} \right] e^{-\frac{t-t_1}{\tau_p}}. \quad (11)$$

Let us find the effective value of the U_{DC} of the rectified voltage on the load R .

It is obvious that:

$$U_{DC} = U_{DC1} + U_{DC2},$$

where U_{DC1} and U_{DC2} are the components caused by the signals $U_1(t)$ and $U_2(t)$, respectively. It is known that the effective value of the periodic function $U(t)$ is found by the formula:

$$U_{DC} = \sqrt{\frac{1}{T} \int_0^T [U(t)]^2 dt} \quad (12)$$

Since in our case the signal $U_1(t)$ acts only in the interval $0-t_1$, and the signal $U_2(t)$ follows it from t_1 to the end of the period T , then the limits of integration in (12) should be chosen accordingly. Then:

$$U_{DC1} = \sqrt{\frac{1}{T} \int_0^{t_1} [U_1(t)]^2 dt} \quad \text{and} \quad (13)$$

$$U_{DC2} = \sqrt{\frac{1}{T} \int_{t_1}^T [U_2(t)]^2 dt} \quad (14)$$

Substituting (10) into (13), (11) into (14) and entering dimensionless parameters: $\beta = \omega RC$ is conditional frequency, $\gamma = \frac{t_f}{T}$ is fill capacity factor, $\Delta = \frac{U_{DC}}{T}$ is detection efficiency and, taking into account the previously accepted parameter $\alpha = \frac{R_B}{R}$, we obtain:

$$\Delta_1 = \frac{U_{DC1}}{E} = \frac{1}{1 + \alpha} \sqrt{\gamma - \frac{\alpha\beta(1-A)(1-B)}{\pi(1+\alpha)(1-AB)} \left[1 - \frac{(1-A)(1-B)}{4(1-AB)} \right]} \quad (15)$$

$$\Delta_2 = \frac{U_{DC2}}{E} = \frac{1}{1 + \alpha} \left[1 - \frac{(1-A)B}{1-AB} \right] \sqrt{\frac{\beta}{4\pi} (1-A^2)} \quad (16)$$

It is obvious that generalizing the rectification efficiency Δ will be defined as the sum of (15) and (16), i.e.

$$\Delta = \Delta_1 + \Delta_2$$

As the rectifier in the circuit (Fig. 1) is made according to a push-pull circuit, then in (15) and (16) the value of β should be doubled.

Graphical simulation results are presented in Results (Fig. 5).

2.2 Modeling of the rectifier unit in NI Multisim 10.1

Using a technique based on replacing the nonlinear element of the gas discharge high-pressure sodium lamp with a linear active resistance $R_L = 20$ ohms, a simulation of this device was carried out (Fig. 4) in the NI Multisim 10.1 program.

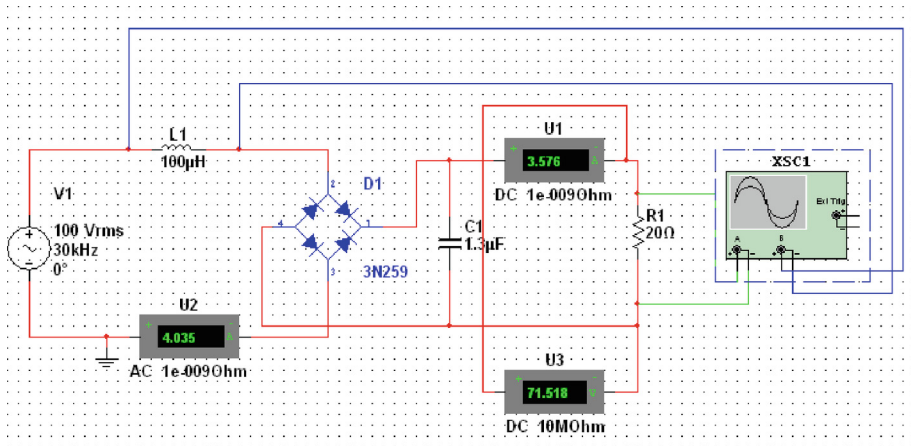


Fig. 4. Schematic diagram of an experimental installation in the NI Multisim 10.1 environment

The resulting waveforms are presented in the Results section.

3 Results

Let us make an approximate estimation of the possible limits of the change of the dimensionless parameters α , β and γ accepted in (15) and (16). In greenhouses, gas discharge high-pressure sodium lamps with a capacity of $400 \div 600$ watts are usually used. Therefore, the value of R will be $30 \div 10$ ohms. The power dissipated in the electronic keys of inverters is $5\text{--}10$ watts. Therefore, it is possible to take $\alpha = 0.01 \div 0.03$. Figure 5.

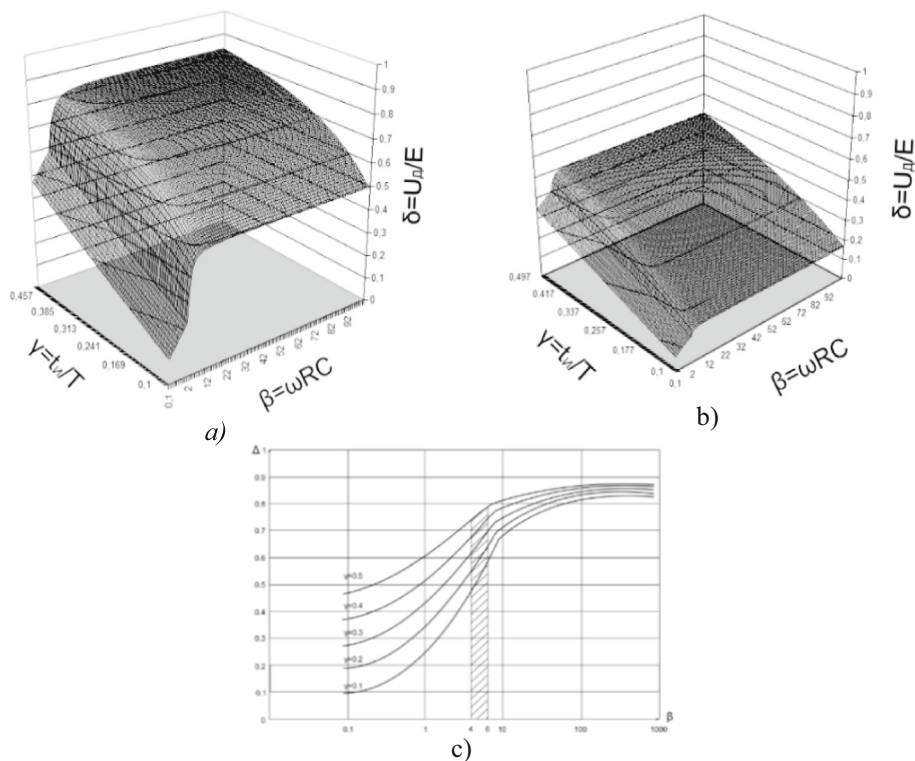


Fig. 5. The detection efficiency Δ depends on the ratio of the load resistance to the internal resistance of the power supply α ; the conditional frequency β and the fill capacity factor γ : a) $\Delta = f(\beta, \gamma)$ at $\alpha = 0.1$; b) $\Delta = f(\beta, \gamma)$ at $\alpha = 0.5$; c) $\Delta = f(\beta)$ at $\alpha = 0.1$; $\gamma = 0.1 \dots 0.5$

The coefficient γ in the case of power supply of a rectifier about a single-stroke rectifier can vary from 0.1 to 0.9. If the inverter is push-pull, it is obvious that $0.1 < \gamma < 0.5$.

The conditional frequency in the reference is usually given within $1 < \beta < 1000$. Taking into account these limitations, graphs of Δ dependences on the parameters α , β and γ are constructed (Fig. 6, 7, 8).

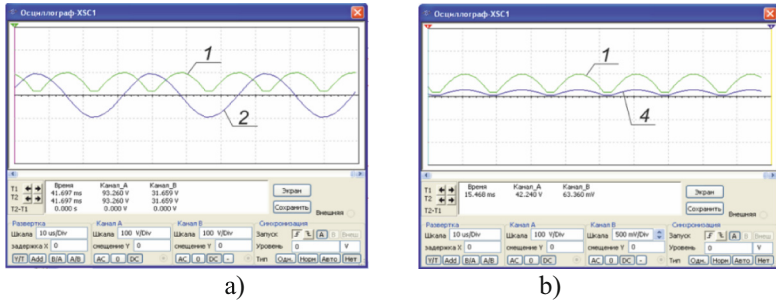


Fig. 6. The results of studies in the absence of a capacitor $C1$: a) voltage waveforms on the choke (2) and load (1); b) voltage waveforms (1) and load current (4)

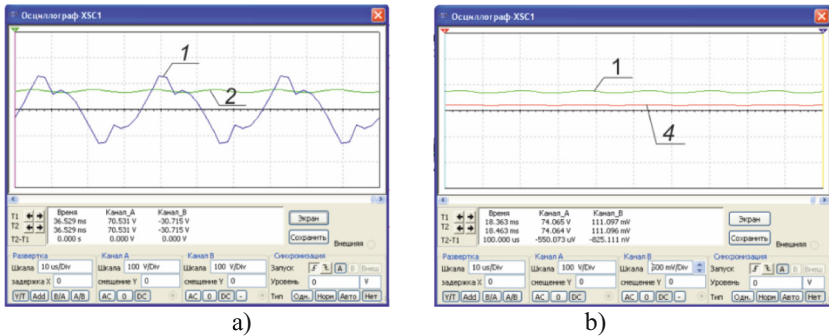


Fig. 7. The results of the studies at $C1 = 1.3 \mu\text{F}$, $wRLS = 4.9$: a) voltage waveforms on the throttle (1) and load (2); b) voltage waveforms (1) and load current (4)



Fig. 8. The results of studies at $C1 = 20 \mu\text{F}$, $wRLS = 75$: a) voltage waveforms on the choke (1) and load (2); b) voltage waveforms (1) and load current (4)

The analysis of the data obtained as a result of theoretical calculations and modeling allows us to draw the following conclusions:

- The efficiency of AC voltage detection is a function of α , β and γ .
- The current value of the output voltage can be controlled using the conditional frequency β or the fill capacity factor γ . Fine-tuning of the output voltage should be performed using the master generator 6, which controls the parameter γ and feeds the inverter 5.
- Since $\beta = \omega RC$, the detection efficiency also depends on the value of the filter capacitance, and in the range of $0.1 \leq \beta \leq 10$ this dependence is most significant.
- Given the frequency of the supply voltage ($f = 30$ kHz) and the equivalent resistance of the lamp ($R = 40$ ohms), the value of C can be calculated by selecting the average value of $\omega RC = 5$ in this range, we get the value of $C = 1.32$ μF (for comparison, in an industrial frequency network, the same function would be performed by a capacitor with a capacity of 800 μF).

4 Conclusion

Based on the results of the research and the goal, a mathematical model of a regulated DC source was developed, which reflects the relationship of the electrical parameters of the supply of high-pressure discharge lamps. The obtained dependences of the effective value of the load voltage on the ratio of the load resistance to the internal resistance of the power supply, the conditional frequency and the fill capacity factor allowed us to make the optimal choice of a capacitive ballast to limit the current flowing through the HID lamp. To test the model's operability, simulation modeling was carried out in the Multisim environment, which confirmed the theoretical conclusions obtained. The presented scientific results can be recommended for use in the development and modeling of Electronic ballasts that realize the supply of gas-discharge lamps with direct current with a cyclic change in voltage polarity.

Acknowledgments. The research was supported by the Russian Science Foundation Grant №22–71-10046, <https://rscf.ru/en/project/22-71-10046/>.





References

1. Badgery-Parker, J.: Light in the greenhouse, NSW Agric (1999)
2. Blanchard, E., Runkle, M.: Manipulating light in the greenhouse, GPN (2009)
3. Fisher, C., Donnelly, P.: Evaluating supplemental light for your greenhouse, Ohio Florist's. (2010)
4. Molina, J., Sainz, L., Monjo, L.: Model of discharge lamps with saturated magnetic ballast and non-square arc voltage. *Electr. Power Syst. Res.* **104**, 42–51 (2013). <https://doi.org/10.1016/J.EPSR.2013.06.003>
5. Alonso, J.M.: Electronic ballasts. *Power Electron. Handb.*, 573–599 (2011). <https://doi.org/10.1016/B978-0-12-382036-5.00022-7>
6. Alonso, J.M.: Electronic ballasts. *Power Electron. Handb.*, 685–710 (2018). <https://doi.org/10.1016/B978-0-12-811407-0.00023-4>

7. Gunawardana, A., Wickramarachchi, N.: Energy saving lamps and electronic ballasts. *Power Electron. Des. Handb.*, 229–242 (1998). <https://doi.org/10.1016/B978-075067073-9/50010-9>
8. Flesch, P.: *Light and Light Sources*. Springer, Berlin, Heidelberg (2007). <https://doi.org/10.1007/978-3-540-32685-4>
9. Directive on energy efficiency requirements for ballasts for fluorescent lighting №2000/55/EC of the European Parliament and of the Council of 18.09.2000, n.d
10. Lu, C.-C.: Ultra-high voltage impulse generator, US6522087B1 (2001)
11. Cheng, C.A., Cheng, H.L., Ku, C.W., Yang, F.L.: A novel low-frequency square-wave-driven electronic ballast for automotive HID lamps. *Comput. Math. with Appl.* **64**, 1409–1419 (2012). <https://doi.org/10.1016/J.CAMWA.2012.03.088>
12. Yang, T.-H.: Device for periodically alternating bulb polarities of a DC fluorescent lighting system, 5072160 (1991)



Modeling of Temperature Contribution to the Interphase Energy of the Faces of Cadmium Crystals at the Boundary with Organic Liquids

Aslan Apekov¹ (✉) , Irina Shebzukhova² , Liana Khamukova¹ ,
and Zaur Kokov² 

¹ North Caucasus Center for Mathematical Research, North Caucasus Federal University, Stavropol, Russia

aslkbksu@yandex.ru

² Kabardino-Balkarian State University, Nalchik, Russia

Abstract. Of great interest is the study of the influence of various factors on the surface characteristics of metallic crystals in contact with organic matter. The features of this interface allow you to create materials with practically useful properties used in various devices. A significant contribution to the surface characteristics of metals is made by free electrons, which at room temperature have sufficient energy of thermal motion to leave the atom and move freely in the crystal lattice. Such electrons do not experience collisions with ion cores and do not deviate from rectilinear motion at distances much larger than the constant lattice, since the ions are located in a regular periodic lattice in which electron waves propagate freely. They also rarely scatter on other conduction electrons due to the Pauli principle. A set of non-interacting free electrons in metals form an electronic liquid. In this paper, a modified version of the electron – statistical method for calculating the temperature contribution to the interfacial energy of metals at the boundary with nonpolar organic liquids is developed. The dependence of the temperature contribution to the interphase energy on the dielectric constant of the liquid and the orientation of the metal crystal are obtained.

Keywords: interphase energy · temperature contribution · cadmium · electron statistical theory

1 Introduction

The determination of surface characteristics at the interface of phases is of great interest since the properties of the solid surface play a crucial role in many technological and natural processes. The number of theoretical and experimental studies devoted to the influence of various factors on the magnitude and behavior of the interfacial characteristics of metal crystals is increasing every year. Recently, the boundary between metal and organic matter has been intensively studied. Organometallic structures are widely

used in practice [1]. The features of the metal-organic interface make it possible to create materials with practically useful properties used in energy storage, electronics [2], catalysis [3], magnetism, nonlinear optics, etc. [4–10], storage and separation of gases [11, 12]. However, it should be noted that the interfacial energy of metals at the boundary with organic liquids and its orientation dependence have not been sufficiently studied [13–20].

2 Materials and Methods

Consider a model of a metal-organic liquid system in which metal ions are immersed in an electronic liquid. The coordinate axis is drawn perpendicular to the interface and is directed towards the organic liquid. The physical interface is drawn relative to the surface ions in such a way that all positive ions of the solid metal belong entirely to the inner region occupied by the metal lattice. The course of the electron density and potential at the flat interface of the crystal face – dielectric liquid is found from the solution of the Thomas-Fermi equation taking into account the macroscopic permittivity of an organic liquid.

$$\frac{d^2V(x)}{dx^2} = 4\pi e\gamma \left(V^{3/2}(x) - V_i^{3/2} \right) \quad \text{when } x \leq 0, \quad (1)$$

$$\frac{d^2V(x)}{dx^2} = \frac{4\pi e\gamma}{\varepsilon_0} V^{3/2}(x) \quad \text{when } x > 0, \quad (2)$$

where eV_i – Fermi boundary energy, e – electron charge, $\gamma = 2^{3/2} / 3\pi^2 e^2 a_0$, a_0 – radius of the first Bohr orbit of a hydrogen atom.

Equations (1) and (2) are crosslinked on the interface $x = 0$ by the conditions of continuity of potential V and its derivative $\frac{dV}{dx}$. In addition, the potential is equal to the Fermi potential in the depth of the metal and zero in the depth of the dielectric liquid.

$$V = V_i \quad \text{when } x = -\infty,$$

$$V = 0V \quad \text{when } x = +\infty.$$

To solve Eqs. (1) and (2), we proceed to a dimensionless potential $\chi(\beta) = \frac{V(x)}{V_i}$ and dimensionless coordinate $\beta = \frac{x}{s}$, and also reduce these equations to a dimensionless form by assuming $s^2 4\pi e\gamma V_i^{1/2} = 1$ (here s is a linear parameter that leads the Thomas–Fermi equation to a dimensionless form). Then Eqs. (1) and (2) will take the form:

$$\chi''(\beta) = \chi^{3/2}(\beta) - 1 \quad \text{when } \beta \leq 0, \quad (3)$$

$$\chi''(\beta) = \frac{1}{\varepsilon_0} \chi^{3/2}(\beta) \quad \text{when } \beta > 0, \quad (4)$$

Equations (3) and (4) are solved under the following boundary conditions:

$$\chi(\beta) = 0 \quad \text{when } \beta = +\infty;$$

$$\chi(\beta) = 1 \quad \text{when } \beta = -\infty;$$

$$\chi'(\beta) = 0 \quad \text{when } \beta = \pm\infty.$$

The dimensionless potential, which determines the course of the potential and electron density at the metal-organic liquid boundary, for the inner and outer regions of the metal at the boundary with a nonpolar organic liquid are obtained [21] in the form:

$$\chi_i(\beta, \varepsilon) = 1 - \frac{1 - \chi(0, \varepsilon)}{(1 + \beta/b)^n} \quad \text{when } \beta \leq 0, \quad (5)$$

$$\chi_e(\beta, \varepsilon) = \frac{\chi(0, \varepsilon)}{(1 + \beta/b)^4} \quad \text{when } \beta \geq 0, \quad (6)$$

where $b = \frac{2\sqrt{5\varepsilon}}{\chi^{1/4}(0, \varepsilon)}$, $n = \frac{4\chi(0, \varepsilon)}{1 - \chi(0, \varepsilon)}$ and $\chi(0, \varepsilon)$ - is the dimensionless potential at the physical interface, depends on the dielectric constant of the liquid ε .

The temperature contribution to the interfacial energy, as well as to the surface energy [22], is determined by the ionic and electronic components of the metal

$$\Delta f_{\omega 12}^{(T)} = \Delta f_{\omega 12}^{(Ti)} + \Delta f_{\omega 12}^{(Te)}. \quad (7)$$

Ionic component

$$\Delta f_{\omega 12}^{(Ti)} = \Delta f_{\omega 12}^{(Th)} + \Delta f_{\omega 12}^{(Ta)} \quad (8)$$

consists of harmonic $\Delta f_{\omega 12}^{(Th)}$ and anharmonic $\Delta f_{\omega 12}^{(Ta)}$ parts. The free energy of the vibrational motion of metal ions at $T \gg \theta$ taking into account anharmonicity, is determined by the formula

$$F(\infty) = -3kT \ln \frac{kT}{\hbar\omega(\infty)} - 3kT^2 g(\infty), \quad (9)$$

where $g = 5k\beta^2/6m\omega^6(\infty)$; $\beta = -1/2(d^3E/dr^3)_{r=a}$ is the first anharmonicity coefficient associated with the thermal coefficient of linear expansion α_l by the ratio $\alpha_l = k\beta/am^2\omega^4(\infty)$; $a = 2\bar{R}$ is the average distance between ions (\bar{R} is the equilibrium radius of the elementary ball; θ is the Debye temperature).

The free energy of oscillations on the j -plane of the transition layer parallel to the Gibbs interface, will be

$$F(x_j) = -3kT \ln \frac{kT}{\hbar\omega_j(\varepsilon)} - 3kT^2 g(\varepsilon). \quad (10)$$

The temperature contribution of the metal ion subsystem to the interphase energy is determined by the expression

$$\Delta f_{\omega 12}^{(Ti)}(hkil) = \sum_j [F(x_j) - F(\infty)] n^{(j)}(hkil), \quad (11)$$

or, moving from summation to integration,

$$\Delta f_{\omega 12}^{(Ti)}(hkil) = n(hkil) \int_{-\infty}^{x_{\Gamma}(\varepsilon)} [F(x) - F(\infty)] \cdot dx, \quad (12)$$

where $n(hkil)$ is the number of particles per 1 m^2 of the metal crystal face.

Substituting (9) and (10) into (12) and expressing $\omega(\varepsilon)/\omega(\infty)$ and $g(\varepsilon)/g(\infty)$ through dimensionless potentials (5) and (6) we obtain

$$\frac{\Delta f_{\omega 12}^{(Ti)}}{n(hkil)} = - \left[0, 9kT \int_{-\infty}^{\beta_{\Gamma}(\varepsilon)} \left(1 - \frac{\beta}{b}\right)^{-6} d\beta + 3, 6 \frac{A}{N_A} \left(\frac{k}{\hbar} \alpha_l \bar{R} \theta T\right)^2 \cdot \int_{-\infty}^{\beta_{\Gamma}(\varepsilon)} \left(1 - \frac{\beta}{b}\right)^{-6} d\beta \right] \quad (13)$$

After performing a simple integration and taking into account the change in the number of particles per 1 m^2 of the surface due to the expansion of the metal in formula (13), we obtain the following expression for the temperature coefficient of the interfacial energy of the metal at the boundary with the organic liquid, due to the ionic component of the metal

$$\frac{df_{\omega 12}^{(Ti)}(hkil)}{dT} = - \frac{n(hkil)}{\left(1 - \frac{\beta_{\Gamma}(\varepsilon)}{b}\right)^5} \left\{ 0, 18 + \frac{1, 44 \alpha_l^2 \bar{R}^2 \theta^2 k^2 AT}{N_A \hbar} \right\}, \quad (14)$$

where D is the density of the metal.

The temperature contribution to the interfacial energy due to the electronic subsystem of the metal is calculated according to [22] taking into account the dielectric constant of the liquid. At a temperature much lower than the degeneracy temperature of the electron gas, the free energy per particle is equal to

$$F \cong \frac{3}{5} \mu_0 - \frac{\pi^2 k^2 T^2}{4 \mu_0}, \quad (15)$$

where $\mu_0 = eV_i = (5/3)k_k \rho^{2/3}(\infty)$ is the Fermi boundary energy.

Consequently, the temperature blurring of the Fermi boundary determines the additional free energy of the electron

$$\Delta F = - \frac{\pi^2 k^2 T^2}{4 \mu_0}. \quad (16)$$

The volume density of this additional energy in any plane j of the transition layer of the metal at the boundary with the organic liquid

$$\varpi_p = - \frac{\pi^2 k^2 T^2}{4 \mu_0(x, \varepsilon)} \rho(x, \varepsilon) = - \frac{\pi^2 k^2 T^2}{4 \mu_0(x, \varepsilon)} \rho^{2/3}(\infty) \rho^{1/3}(x, \varepsilon), \quad (17)$$

where $\rho(x, \varepsilon)$ is the volume density of the electron gas. Therefore, the excess free energy associated with the blurring of the Fermi boundary in the inner region of the metal will

give the following contribution to the interfacial energy of the metal at the boundary with the organic liquid:

$$\Delta f_{\omega 12}^{i(Te)} = -\frac{\pi^2 k^2 T^2}{4\mu_0} \rho^{2/3}(\infty) \int_{-\infty}^{x_{\Gamma}(\varepsilon)} \left[\rho_i^{1/3}(x, \varepsilon) - \rho_i^{1/3}(\infty) \right] dx. \quad (18)$$

The outer part of the electron density distribution ($x > x_{\Gamma}$) will contribute to the interphase energy

$$\Delta f_{\omega 12}^{e(Te)} = -\frac{\pi^2 k^2 T^2}{4\mu_0} \rho^{2/3}(\infty) \left\{ \int_{x_{\Gamma}(\varepsilon)}^0 \left[\rho_i^{1/3}(x, \varepsilon) - \rho_e^{1/3}(\infty) \right] dx + \int_0^{\infty} \left[\rho_e^{1/3}(x, \varepsilon) - \rho_e^{1/3}(\infty) \right] dx \right\}. \quad (19)$$

Summing up (18) and (19) and passing to the dimensionless coordinate $\beta = x/s$ and functions (5) and (6), we obtain the full contribution to the interfacial energy of the metal due to the temperature dependence of the Fermi energy:

$$\Delta f_{\omega 12}^{(Te)} = -\frac{\pi^2 k^2 T^2}{4\mu_0} s \rho(\infty) \times \left\{ \int_{-\infty}^{\beta_{\Gamma}(\varepsilon)} \left[\chi_i^{1/2}(\beta, \varepsilon) - 1 \right] d\beta + \int_{\beta_{\Gamma}(\varepsilon)}^0 \left[\chi_i^{1/2}(\beta, \varepsilon) - \frac{\rho_e(\infty)}{\rho_i(\infty)} \right] d\beta + \int_0^{\infty} \left[\chi_e^{1/2}(\beta, \varepsilon) - \frac{\rho_e(\beta, \varepsilon)}{\rho_i(\infty)} \right] d\beta \right\} \quad (20)$$

or, in view of the fact that $\rho_i(\infty) = z/\Omega$ and $\rho_e(\infty) = 0$, we have:

$$\Delta f_{\omega 12}^{(Te)}(hkil) = -\frac{\pi^2 k^2 T^2 z}{4\mu_0} \left\{ \int_{-\infty}^{\beta_{\Gamma}(\varepsilon)} \left[\chi_i^{1/2}(\beta, \varepsilon) - 1 \right] d\beta + \int_{\beta_{\Gamma}(\varepsilon)}^0 \chi_i^{1/2}(\beta, \varepsilon) d\beta + \int_0^{\infty} \chi_e^{1/2}(\beta, \varepsilon) d\beta \right\} n(hkil). \quad (21)$$

Here z is the average number of free electrons per metal atom; $\Omega = \frac{A}{DN}$ is the volume of an elementary ball. By decomposing into a series $\chi_i^{1/2}(\beta, \varepsilon)$ and integrating (21), we find the contribution to the temperature coefficient of the interphase energy due to the blurring of the Fermi level:

$$\frac{df_{\omega 12}^{(Te)}(hkil)}{dT} = -\frac{\pi^2 k^2 T^2 z b}{2\mu_0} \left\{ \frac{3(1 - \chi(0, \varepsilon))}{10 \left(1 - \frac{\beta_{\Gamma}(\varepsilon)}{b}\right)^5} - \frac{b(1 - \chi(0, \varepsilon))}{10} + \chi^{1/2}(0, \varepsilon) \right\} n(hkil). \quad (22)$$

The contribution to the interphase energy of the temperature coefficient for the metal crystal – organic liquid boundary can finally be represented as

$$\frac{df_{\omega 12}^{(T)}(hkil)}{dT} = \frac{df_{\omega 12}^{(Ti)}(hkil)}{dT} + \frac{df_{\omega 12}^{(Te)}(hkil)}{dT}. \quad (23)$$

3 Results

The temperature coefficient of the interfacial energy of a metal at the boundary with an organic liquid is due to the anharmonicity of ion oscillations and the associated expansion of the metal and a change in ion energy, a change in the nature of ion oscillations in the transition layer due to the presence of density gradient of an electron liquid and the blurring of Fermi energy. The main contribution is due to the anharmonicity of ion oscillations. According to the formula (21), the temperature contributions to the interfacial energy of the faces of a cadmium crystal having a hexagonal tightly packed structure at the boundary with 11 nonpolar organic liquids (pentane, hexane, heptane, octane, decane, nonane, n-xylene, benzene, m-xylene, toluene, o-xylene) at a temperature of 293 K are calculated (Fig. 1).

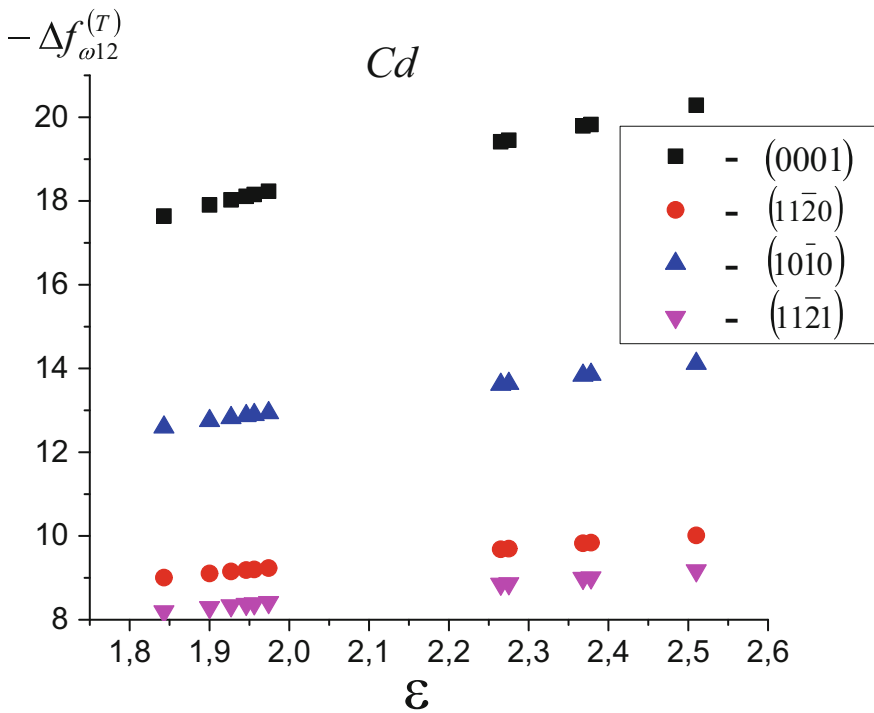


Fig. 1. Temperature contribution to the interphase energy for cadmium faces at the boundary with organic liquids.

4 Discussion

The proposed method makes it possible to establish the orientation dependence of the temperature coefficient of the interphase energy.

The magnitude of the temperature contribution to the interfacial energy of metals is due to the ionic component of the metal and the temperature blur of the Fermi level. The main contribution is due to the anharmonicity of the vibrations of the ionic component of the metal. With an increase in the permittivity of a nonpolar organic liquid, the temperature coefficient of the interfacial energy of the metal increases.

As can be seen from the graph, the temperature contributions are negative, and with an increase in the permittivity of the organic liquid, they increase in modulus. That is, they reduce the interfacial energy. The value of the temperature contribution to the interphase energy of crystals is 3–8% of the interphase energy for different faces.

The calculated values of temperature contributions to the interphase energy of the faces of cadmium crystals at the boundary with 11 nonpolar organic liquids show the following orientation dependence $\Delta f_{\omega 12}^{(T)}(0001) > \Delta f_{\omega 12}^{(T)}(10\bar{1}0) > \Delta f_{\omega 12}^{(T)}(11\bar{2}0) > \Delta f_{\omega 12}^{(T)}(11\bar{2}1)$.

5 Conclusion

In this paper, within the framework of a modified version of the electron–statistical theory, the temperature contribution to the interfacial energy of a cadmium crystal at the boundary with organic liquids from the dielectric constant ϵ of the liquid is obtained. The obtained values are in good agreement with experimental data for simple metals [9] and show that the faces with a higher reticular density have the greatest value of the temperature contribution to the interfacial energy.

Acknowledgments. The work is supported by North-Caucasus Center for Mathematical Research under agreement №. 075-02-2022-892 with the Ministry of Science and Higher Education of the Russian Federation.

References

1. Czaja, A.U., Trukhan, N., Müller, U.: Industrial applications of metal–organic frameworks. *Chem. Soc. Rev.* **38**(5), 1284–1293 (2009)
2. Liu, Z., Kobayashi, M., Paul, B.C., Bao, Z., Nishi, Y.: Contact engineering for organic semiconductor devices via Fermi level depinning at the metal–organic interface. *Phys. Rev. B* **82**(3), 035311 (2010)
3. Seo, J., et al.: A homochiral metal–organic porous material for enantioselective separation and catalysis. *Nature* **404**, 982–986 (2000)
4. Butova, V.V., Soldatov, M.A., Guda, A.A., Lomachenko, K.A., Lamberti, C.: Metal–organic frameworks: structure, properties, methods of synthesis and characterization. *Russ. Chem. Rev.* **85**(3), 280–307 (2016)
5. Stroppa, A., Barone, P., Jain, P., Perez-Mato, J.M., Picozziet, S.: Hybrid improper ferroelectricity in a multiferroic and magnetoelectric metal–organic framework. *Adv. Mater.* **25**(16), 2284–2290 (2013)
6. Ferrey, G.: Hybrid porous solids: past, present, future. *Chem. Soc. Rev.* **37**(1), 191–214 (2008)
7. Gibbons, N., Baumberg, J.: Optical minibands in metallodielectric superlattices. *Phys. Rev. B* **85**(16), 165422 (2012)

8. Bogdanov, A.A., Suris, R.A.: Effect of the anisotropy of a conducting layer on the dispersion law of electromagnetic waves in layered metal-dielectric structures. *JETP Lett.* **96**, 49–55 (2012)
9. Zadumkin, S.N.: A new version of the statistical electronic theory of surface tension of metals. *Fiz. Met. Metalloved.* **11**(3), 331–346 (1961)
10. Apekov, A.M., Shebzukhova, I.G.: Polarization correction to the interfacial energy of faces of alkali metal crystals at the borders with a nonpolar organic liquid. *Bull. Russ. Acad. Sci. Phys.* **82**(7), 789–792 (2018). <https://doi.org/10.3103/S1062873818070067>
11. Mendoza-Cortes, J.L., Han, S.S., Goddard, W.A.: High H₂ uptake in Li-, Na-, K-metalated covalent organic frameworks and metal organic frameworks at 298 K. *Phys. Chem. A* **116**(6), 1621–1631 (2012)
12. Furukawa, H., et al.: Ultrahigh porosity in metal-organic frameworks. *Science* **329**, 424–428 (2010)
13. Shebzukhova, I.G., Apekov, A.M.: Contribution of dispersion interaction of s-spheres into the interface energy of a-Li and a-Na crystals bounding to non-polar organic liquids. *Phys. Chem. Aspects Study Clusters Nanostruct. Nanomater.* **9**, 518–521 (2017)
14. Apekov, A.M., Shebzukhova, I.G.: Temperature contribution to the interfacial energy of the crystals faces of Sc, α -Ti and α -Co at the boundary with the organic liquids. *Phys. Chem. Aspects Study Clusters Nanostruct. Nanomater.* **8**, 19–25 (2016)
15. Shebzukhova, I.G., Apekov, A.M., Khokonov, K.B.: Orientation dependence of the interfacial energies of chromium and α -iron crystals at boundaries with nonpolar organic liquids. *Bull. Russ. Acad. Sci. Phys.* **81**(5), 605–607 (2017). <https://doi.org/10.3103/S1062873817050173>
16. Apekov, A.M., Shebzukhova, I.G.: Interface energy of crystal faces of IIA-type metals at boundaries with nonpolar organic liquids, allowing for dispersion and polarization corrections. *Bull. Russ. Acad. Sci. Phys.* **83**(6), 760–763 (2019)
17. Shebzukhova, I.G., Apekov, A.M., Khokonov, K.B.: Anisotropy of the interface energy of IA and IB metals at a boundary with organic liquids. *Bull. Russ. Acad. Sci. Phys.* **80**(6), 657–659 (2016). <https://doi.org/10.3103/S1062873816060307>
18. Shebzukhova, I.G., Apekov, A.M., Khokonov, K.B.: Interface energy of faces of manganese and vanadium crystals at boundaries with organic liquids. *Bull. Russ. Acad. Sci. Phys.* **79**(6), 749–751 (2015). <https://doi.org/10.3103/S1062873815060295>
19. Shebzukhova, I.G., Apekov, A.M., Khokonov, K.B.: Effect of an organic liquid on the surface energy of scandium and titanium. *Bull. Russ. Acad. Sci. Phys.* **78**(8), 804–806 (2014). <https://doi.org/10.3103/S1062873814080334>
20. Apekov, A.M., Shebzukhova, I.G.: Of the facets at the boundary between calcium/barium crystals and nonpolar organic liquids. *Phys. Chem. Aspects Study Clusters Nanostruct. Nanomater.* **10**, 20–26 (2018)
21. Apekov, A.M., Shebzukhova, I.G.: Orientational dependence of the interphase energy of low temperature modification of titanium at the boundary with an organic liquid. *Phys. Chem. Aspects Study Clusters Nanostruct. Nanomater.* **14**, 17–23 (2022)
22. Zadumkin, S.N., Pugachevich, P.P.: Temperature dependence of surface tension of metals. *Dokl. Akad. Nauk SSSR.* **146**(6), 1363–1366 (1962)

Information and Computation Systems for Distributed Environments



Mathematical Concept of a Model for Processing Metadata of Employee's Psycho-States for Identifying Him as an Internal Violator (Insider)

I. V. Mandritsa¹(✉), V. I. Petrenko¹, O. V. Mandritsa², and T. V. Minkina¹

¹ Department of Organization and Technologies Defense of Information, North-Caucases Federal University, Institute of Digital Development, Stavropol, Russia
imandritsa@ncfu.ru

² Faculty of Regional Development, Head of the Department of Regional Economics, Branch of RTU MIREA in Stavropol, Stavropol, Russia

Abstract. The identification of an internal violator is becoming more and more relevant with the development of information processing technologies, the security of technical and software complexes in information systems is increasing, and the person remains the most vulnerable place. An employee who processes information and has full access to it becomes the target of an attacker, as well as he himself may be an intruder and become a source of leakage of confidential information. The analysis of data about the employee allows us to make assumptions about his emotional state, willingness to honestly do the job. The collected data requires analysis according to their type, and this is what the work is dedicated to. Data from video surveillance cameras is analyzed to identify the time of the employee's appearance and his location during working hours. The metadata obtained is intended to create a model of an internal violator, followed by an assessment of the threat level emanating from the employee.

Keywords: Internal intruder · Mathematical model · Metadata collection · Probability of state

1 Introduction

The purpose of this article is to investigate how to extract metadata about an employee from his environment in order to then mathematically calculate the probability of the main causes and consequences of employee fraud, focusing on his dynamic changes in the face when he performs his functional tasks at work, at the company and is involved in the stages and stages of the business process. The authors note the fact that metadata collected from the external environment somehow affects the behavior of an employee of the company and introduces him into an imbalance in the performance of work functions (he more often conflicts with colleagues, more often skips or performs less productively the amount of work and reduces its quality) [1].

The compliance of the company (considering the employment relationship) with the employees, is clearly the most appreciated characteristic by the employees, and our participants. After this characteristic, the objectives of the company seem to be the most valued by the employees, which reveal concern to comply with their obligations to the company, along with their belief in the fulfillment of the employment relationship by the organization. This leads us to admit that there is a greater belief of the employees in the fulfillment of the physical contract than in the fulfillment of the psychological contract. On the opposite, the most negative aspect highlighted in the company is the concern with its employees, that is, the employees think that the company does not care as much about them as they would like, or how they think that the company should be concerned, what causes a low satisfaction among employees.

Findings also indicate that both the company's interest in knowing the life of the employee and the fulfillment of the financial obligations of the company, have a low impact on the employees, not causing a considerable impression. The company's weak concern with its human resources proves the weak internal branding of the organization [2].

Many Prior study of science has found that a company with high employees' integrity (absent Internal intruder) and cultivate integrity culture within an organization impacted on business efficiency and information security company, it also positively impacted the financial performance over the years. Besides, Rosli, Abd Aziz, Mohd, and Said, (2015) claimed that an organization with high integrity might have the potential to contribute towards competitive advantage and improve public trust and transparency in all its activities. The result of many studies show that the practices of the integrity system is affected by the leadership quality, but the result of the internal control system of employer showed a mixed relationship with the practices of accountability. Their study reveals that an integrity system can help an organization to enhance its accountability to the various group of stakeholders. The focus has now turned to the competency, integrity, and the attractiveness of a company in protecting public trust. Thus, various benefits will be gained by the organization if integrity becomes the focal point of the organization. The evidence of a psychological contract breach may be a first indicator of a possible future turnover, very common in IT corporations. The psychological contract may play the role of a formal governance particularly in the case of IT outsourcing (Lioliou et al., 2014). Thus, literature acknowledges the particular relevance of the psychological contract to IT consultancy corporations while at the same time the additional challenges of internal branding due to the typical outsourcing role of employees. They always feeling that they have not longest contract with company [3, 4].

Having studied the circumstances of conscientiousness of behavior and performance of their functions by an employee, the authors of this article came to the conclusion that in order to help the company's management, it is necessary to create a mechanism for collecting metadata about an employee from the perspective of the external environment of his life (family, relatives and their impact on him) as well as within the company's team (its conflict-free or vice versa, a change in tolerance in business-in the process of the company).

The improvement of information processing technologies entails a reduction in the possibilities for unauthorized receipt of confidential information by an external attacker,

while an employee with access to confidential information may become a source of leakage of valuable information, through malicious intent or careless handling of the information processing system. Both in the case of unintentional mistakes in handling the company's information resources, which is the result of emotional fluctuations: apathy, fatigue, bad mood, and in the case of deliberate transfer of information to third parties, hostile actions against the welfare of the company are affected by the emotional state of a person. That is, different degrees of negative actions of an employee of the company are affected by various emotional deviations caused by a reaction to external factors of the same nature. Knowing the set of these factors and the degree of their impact, it is possible to conduct a presumptive analysis of the emotional state and, as a consequence, the possible behavior of the subject, for this purpose, a model of the internal violator is compiled, reflecting the spectrum of negative emotional states of a person. Metadata for the model is taken from open sources.

2 Materials and Methods

Metadata is information about other information, or data related to additional information about the content or object. Metadata reveals information about the features and properties that characterize any entities, allowing you to automatically search for and manage them in large information flows [5].

Next, in Fig. 1, we present the concept of collecting metadata about an employee from open and open sources of the information space. Accordingly, we get the maximum amount of damage from the types of threats that may arise from an employee, and the amount of the costs of information protection measures for the developed measures to face both external and internal threats. At the same time, the main focus of the information security specialist will be a new model for building a personnel management system at the stages of the business process and, accordingly, the amounts of possible damage to these stages.

For a more detailed analysis of the risks to a participant in the business process of becoming an insider of a competitor company, the authors developed a concept for studying the internal threats of business information leakage from insider personnel, based on its metadata related to the external environment, as shown [6]. As it can be seen from Fig. 1, the concept describes three types of collected information about a participant to a business process:

- information about the work of an employee within the company;
- information from closed sources (CSINT);
- information from open sources (OSINT).

Each type of information collection combines a group of search aspects that should be emphasized when identifying the threat of internal leakage of information from personnel. Thus, the authors distinguish between two types of risks - ($\sum R(CSINT)$) and ($\sum R(OSINT)$) that are subject to a scientifically based measurement method, for the purpose of early detection of risks from an internal insider [7–9].

There is a set of tools for programs to interact with each other, called the API (Application Programming Interface), with which it is possible to get information about the

participants of the business process using the data of an application. As mentioned above, business information can be information about the business process of the company, as well as data about the participants in this business process, especially customers who bring income (benefit).

If such information is publicly available on the Internet, for example, in social networks, it is possible to find, structure, and analyze such information through the use of API technology (whether an employee has problems). Metadata creation can take place both manually and automatically. From various sources of both open and “leaks” of information from closed sources (debtors’ databases, databases on offenses on the roads, etc.), flows of data about a person, his material support, possible problems with the law are formed. From social networks there is a stream of data on the worldview, political and social views of a person. You can also use internal data of firm and video surveillance, indicating the movements of a person inside the object.

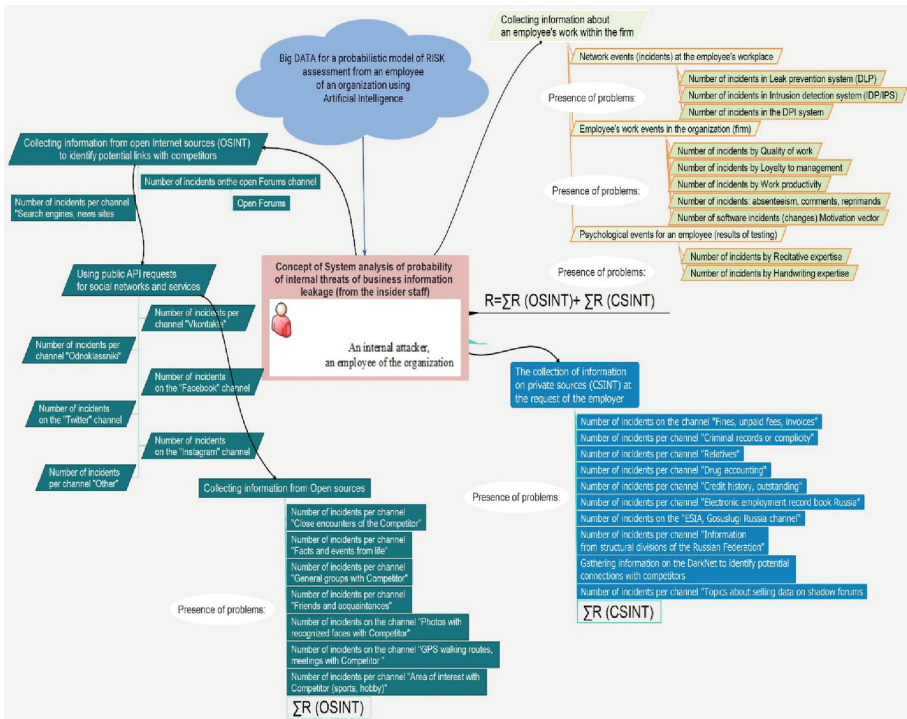


Fig. 1. Concept of collecting metadata about the worker for the mathematical model of probability of internal threats (risks of leaks of business information.)

The simplest and most accessible method for collecting metadata about a worker is the APIs-requests, that are a technology, an architectural style focused on using the HTTP protocol as a transport protocol when generating requests to the server and responses returned by this server. If the information about the participants of the business process is

located, for example, in a social network, then it can be likely obtained from the servers used by this social network via its API - leakage (from the insider staff).

An example is shown in Fig. 2. The authors also believe that threats to a commercial organization can, in principle, be divided into two main types: external and internal). The task of any criminal (as an attacker inside the company or a cyber fraudster from the external environment of the company) will be to «complicate» or «reset» the business information of the company between its stages of production and sale of products or services by introducing chaos, disorienting the employees of the company, violating the integrity of the aggregate information that discusses the entire business process. In turn, the amount of damages will depend on the recovery time or slow motion «information flow» between employees, and also the time of the financial flows between the client and the departments of the executors of the order, from the perspective of possible cyber-attacks aiming at «losing», «bankrupting», «resetting» the transaction itself and creating a «direct financial» damage from subsequent alterations, or claims (return receipts) on the client side in the form of deviations from the approved parameters of the business process.

```
1 import requests
2
3 token = '1f036ad11f036ad11f036ad1461f7521c911f031f036ad17f1ed244d3bbc59f2880e49c'
4 version = 5.89
5 user_ids = '290490593'
6 fields = 'photo_50,verified'
7 name_case = 'nom'
8
9
10 response = requests.get('https://api.vk.com/method/users.get',
11                          params={
12                              'access.token': token,
13                              'v': version,
14                              'user_ids': user_ids,
15                              'fields': fields,
16                              'name_case': name_case
17                          })
18
19 data = response.json()
20 print(1)
```

Fig. 2. Program source code using the API requests for collecting metadata of employer.

Common factors that are considered when making threat prediction models are sociodemographic factors such as age, gender, education, migration background and ethnicity, religious affiliation, marital status, household, employment, and income. For more accurate predictions we propose taking psychological factors like social learning into account as well.

Fig. 3 goes on to present the concept of a mathematical model for processing collected metadata about worker X_i for the purpose of calculating the probabilities of his behavior in two types: the firm's "Own (employee)" and "Alien" employee of the firm.

The model includes the collection of metadata about the worker in two types. Two types of "Alien" and "Own" coexist in one employee at once, but in the mode of triggering

an externally internal emotional state from external factors. The composition of the desired indicators and model factors.

- $X_1^n = \{1, 2, \dots, n\}$ – an employee of the company from the staffing table and who gained access to the company’s trade secrets;
- $S_j(X_i)_m^{ext}$ – probabilistic stable state of external motivation of the employee S_i (after exposure from the outside to someone else’s threat, when the employee has moved to an unstable state S_j - and ready for “malicious intent”, t. a. to “Winnings”»), through observable conditionally measurable indicators (incidents) per employee = $\sum QX_i(ERIF) + \sum QX_i(IEFI)$.

As a result, we will get conditionally constant values of natural values of facts (0-no, 1-yes) for two streams of information (next Table 1).

Let us stipulate that simultaneously with this flow of ERFI, the “Intra-emotional flow of threat fixation information” for the employee will be monitored, namely the observed facts and events from his private life: conflicts and quarrels in the family, illnesses, alimony, his outstanding loans, mortgage debts, fines, interest on microloans, car accidents, etc. the facts of influencing the motivation of the employee through subsequent changes in his state “Employee” - “Internal violator”. Abbreviated IEFI (Insider the Emotional Flow of Information fixing the threat).

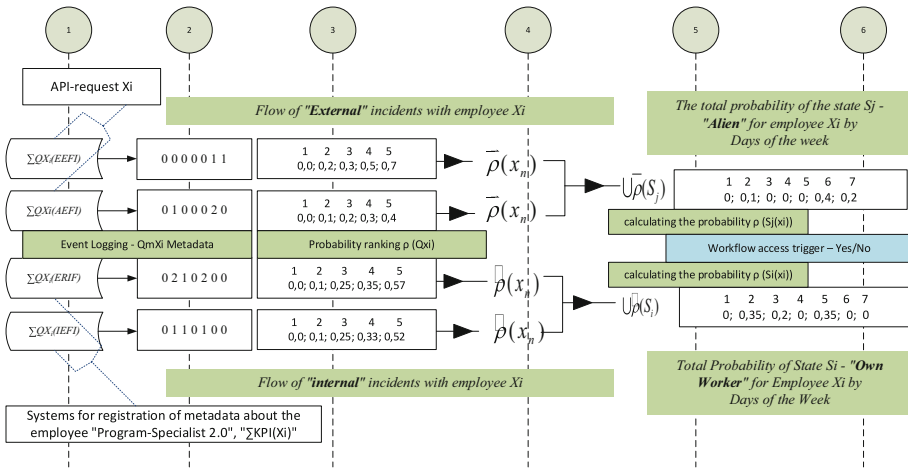


Fig. 3. The concept of a mathematical model of collection metadata of employer.

Social learning theory is a theory of learning process and social behavior which proposes that new behaviors can be acquired by observing and imitating others. According to social learning theory, people engage in crime because of their association with others who engage in crime. Their criminal behavior is reinforced and they learn beliefs that are favorable to crime [10]. They essentially have criminal models that they associate with. As a consequence, these individuals come to view crime as something that is desirable, or at least justifiable in certain situations. Learning criminal or deviant behavior is the

same as learning to engage in conforming behavior: it is done through association with or exposure to others.

At this stage of work, we propose the following concept of two scientific hypotheses – who is such an employee for the purposes of controlling access to the company’s trade secrets [11]. The first hypothesis (assumption) states that our object-worker is defined as an employee with a “public face” (in society, in a collective), as a predetermined “unstable deterministic system” of a set of indicators (factors) describing his current psycho-emotional (the first flow of metadata) and physical-economic (the second flow of metadata) states, and accordingly is interpreted by us as a “stable” or “unstable” state of his motivation to “become or not to become” an internal violator of the firm [12].

The second hypothesis (assumption), predetermines which of the two independent streams of metadata proposed according to the first hypothesis, about the confirmed (or unconfirmed) facts of its developing “crisis” in the ethics of behavior (conflicts, empathy, selfishness, etc.) and in labor indicators (productivity, quality, simple, etc.), should have a different philosophy of comprehension that is converted into different levels of ranking the probabilistic state of stability and instability of the object. The fact is that the internal environment of functioning in the enterprise “requires” from our object of research the fulfillment of a set of indicators of functioning in the company, but which reflect its “some” current in development “intermediate” state of motivation and satisfaction, which for the research model is a certain “rank” of the state [13].

3 Results and Discussion

Further, Table 1 shows the method of collecting and processing information flows about the employee - metadata such as “External resonance flow of information” - ERIF and “Intra-emotional flow of information” - IEFI for the employee Xi.

With the use API from multiple web services and social media applications we can determine whether an employee could become a threat or not. By looking at what kind of people they follow or if any of their close friends and family members have committed any criminal offences for instance can tell us if the employee is more inclined to deviant behavior. Information compiled from this factor as well as others will analyzed and put into a mathematical model which will calculate the probability of the employee being a threat. Should the probability be higher than a certain threshold which is determined by the head of information security then the employee is locked out of the organizations system.

One confirmed event returns a value of 0 in 1 (Table 1), which is important for calculating the probability, which means that there is a fact of a certain proportion (weight) of the information and psychological impact of this flow of information on Xi- of the employee until the subsequent confirmation of the fact. After all the steps of collecting metadata for the mathematical model, at the end of the algorithm it will be necessary to calculate the probabilistic final states “Own” - or “Alien” - for employee $S_i S_j$ Xi from the number of $Q(X_i)$ incidents affecting him “from the outside”.

Table 1. Metadata processing on an example “External resonance flow of information” - ERIF and “Intra-emotional flow of information” - IEFI.

№ pp	for each Xi employee with access to commercial information TIME OF DATA COLLECTION	One week of life of employee Xi (frequency of observations 2 times a day)						
		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
1	2	3	4	5	6	7	8	9
1	TOTAL Machine information gathering API-enquiries - Daily = [AMOUNTS(ERIF) + AMOUNTS(IEFI)] (string 2 + string 10)	0	1	0	0	0	3	1
2	External resonance flow of information AMOUNTS(ERIF) (string 3; string 9)	0	0	0	0	0	1	1
3	Other people's fines	0	0	0	0	0	0	0
4	Other people's debts	0	0	0	0	0	0	0
5	Other people's loans	0	0	0	0	0	0	0
6	Other people's diseases	0	0	0	0	0	1	1
7	Other people's threats	0	0	0	0	0	0	0
8	Other people's accidents and damages	0	0	0	0	0	0	0

(continued)

Table 1. (continued)

№ pp	for each Xi employee with access to commercial information TIME OF DATA COLLECTION	One week of life of employee Xi (frequency of observations 2 times a day)						
		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
9	Other people's problems from connections	0	0	0	0	0	0	0
10	Intra-emotional flow of information AMOUNTS(IEFI) (string 11: string 17)	0	1	0	0	0	2	0
11	Own fines	0	0	0	0	0	0	0
12	Your debts	0	0	0	0	0	0	0
13	Your credits	0	0	0	0	0	0	0
14	His illnesses	0	0	0	0	0	0	0
15	Your quarrels in the family	0	1	0	0	0	1	0
16	Own accidents and damages	0	0	0	0	0	0	0
17	Your Depression	0	0	0	0	0	1	0

The previously presented hypothesis for calculating the full probability of the observed impact of incidents on the employee’s condition is “Alien” according to the generalized Bayes formula $\rho_j S_j(X_i)_{m}^{ext} X_i$ [14] (1).

$$P\left(\frac{D_{Xi}}{K^{Sj}}\right) = P(D_{Xi}) * P\left(\frac{K^{Sj}}{D_{Xi}}\right) / P(K^{Sj}) \tag{1}$$

where is:

$P\left(\frac{D_{Xi}}{K^{Sj}}\right)$ – obtained full probability by state S_j – « Alien»;

$P(D_{Xi})$ – preliminary marginal probability (threshold hypothesis) of employee dissatisfaction X_i .

By converting the probabilities from the incidents observed from closed and open sources confirming the impact of the i -threat on the X_i -employee from the first ERIF data stream and the second IEFI data stream into their cumulative probability – that there is a current value – “Alien”, we will compare it with the full Bayesian probability according to the formula (2). Observations are carried out 1 time per day in 1 week. $(x^{ERIF}_i) (x^{IEFI}_i) \cup \bar{\rho}(S_j)$

$$\bar{p} = \frac{1}{7} * \sum_1^7 (0, 0 + 0, 0 + 0, 0 + 0, 0 + 0, 0 + 0, 2 + 0, 2) +$$

$$\frac{1}{7} * \sum_1^7 (0, 0 + 0, 1 + 0, 0 + 0, 0 + 0, 0 + 0, 2 + 0, 0) = 0, 057 + 0, 042 = 0, 1 \tag{2}$$

Comparing, we get two responses from the employee’s access trigger to the firm’s business process for this employee X_i : “Access is open” or “Access is denied”. The authors on a conditional example of the company and taking as the studied object of the employee - the “office manager” for which the hypothetical threshold of the dissatisfaction coefficient was set $D = (0.6)$, and the resulting probability from two streams = 0.1, we get the next complete Bayes probability (3) and (4), which will confirm or refute the reliability of the values obtained $\cup \bar{\rho}(S_j)$.

$$P_A(H_1) = \frac{0, 5 * 0, 6}{0, 5 * 0, 56 + 0, 55 * 0, 1} = \frac{0, 3}{0, 45} = 0, 77 \tag{3}$$

$$P_A(H_2) = \frac{0, 5 * 0, 1}{0, 5 * 0, 56 + 0, 55 * 0, 1} = \frac{0, 05}{0, 35} = 0, 18 \tag{4}$$

That is, the state of “Alien” for employee X_i is reliable by 77%. Therefore, the trigger of the company’s information security system will be confidently defined for the employee as “Access is allowed”. Next, we will repeat the same procedure and calculations for the third and fourth data stream for the X_i employee to calculate the probability that he is “His” and is in the state S_i (Table 2). $S_i(X_i)_{m}^{int}$ – stable state of internal motivation of the employee (positive state of the motive for work “Svoy”) - conditionally measured multifactorial indicators (registered events, incidents) as a registered fact of the state of the object $QX_i = \sum QX_i(ERFI) + \sum QX_i(AEFI)$.

Table 2. Information flows of data by type "External-emotional flow of information" EEFI and "Intra-emotional flow of information" of AEFI by employee Xi.

№ ss	for each Xi employee with access to commercial information	One week of life of employee Xi (frequency of observations 1 time per week)						
		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
	TIME OF DATA COLLECTION							
1	2	3	4	5	6	7	8	9
1	SUMMARY Manual collection of information EXPERT requests – once a week = [AMOUNTS(ERIF) + AMOUNTS(EEFI)] (string 2 + string 10)	0	3	2	0	3	0	0
2	External-emotional flow of information AMOUNTS(EEFI) (string 3: string 8)	0	2	1	0	2	0	0
3	Facts of conflicts with colleagues	0	1	1	0	0	0	0
4	Facts of quarrels and aggression	0	0	0	0	0	0	0
5	Facts of burnout and weak motivation	0	0	0	0	1	0	0
6	Facts of swearing at the management	0	1	0	0	0	0	0
7	Facts of excessive curiosity, boredom	0	0	0	0	0	0	0
8	Facts of empathy and hatred	0	0	0	0	1	0	0

(continued)

Table 2. (continued)

№ ss	for each X_i employee with access to commercial information	One week of life of employee X_i (frequency of observations 1 time per week)						
		Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
	TIME OF DATA COLLECTION							
9	Intra-emotional flow of information AMOUNTS(AEFD) (string 10: string 14)	0	1	1	0	1	0	0
10	Facts of truancy and laziness	0	0	0	0	0	0	0
11	Facts of a decrease in the quality of work	0	0	1	0	1	0	0
12	Facts of negligence and mistakes and lies	0	0	0	0	0	0	0
13	Facts of a decline in labor productivity	0	0	0	0	0	0	0
14	Facts of theft and damage	0	1	0	0	0	0	0

When trying to assess the probability, experts note that the only way to obtain objective values of the parameter “Probability of threat realization” is to accumulate statistics on incidents, which will determine the number of implementations of threats of a certain type on an information asset of a certain type [14]. However, in the domestic practice of activities for the accumulation of statistical data, insufficient attention is paid and it is currently absent on the flows of information indicated by us. Table 3 then summarizes the probability of the Office Manager employee’s state probability calculation across four metadata streams, and Fig. 4 and 5 shows the probabilities of the employee’s states by day.

Table 3. Rank approach to assessing the degree of probability of transition from the state – "Own" to the state – "Alien" (on the example of an employee – "Office Manager" of the enterprise) $S_i S_j$.

Probabilities for Employee Data Stream X_i - Office Manager	constant probability S_j				
<i>number of incidents</i>	0	1	2	3	4
External resonance flow of information AMOUNTS(ERIF)	0,0	0,2	0,3	0,5	0,7
levels are set on the assumption that external threats (incidents) affect strongly and constantly, and vice versa					
Intra-emotional flow of information AMOUNTS(IEFI)	0,0	0,1	0,2	0,3	0,4
levels are established on the assumption that the inner personality by the age of 40 is satisfied to be resistant to threats (incidents) and their impact is not prolonged					
	variable probability S_i				
number of incidents	0	1	2	3	4
External-emotional flow of information AMOUNTS(EEFI)	0,00	0,1	0,25	0,35	0,57
levels are established on the assumption that the external manifestation of personality emotions by the age of 40 is satisfied with the rapid but also retreating to threats (incidents) and the impact depends on the duration of external contacts					
Activity-economic flow of information AMOUNTS(AEFI)	0,00	0,1	0,25	0,33	0,52
levels are set on the assumption that the business activity of an individual by the age of 40 is satisfied with the unstable to threats (incidents) and the impact is dangerous					

As can be seen from the Fig. 4, the maximum impact on the probabilistic states of the employee is exerted by the “Active-emotional flow of information” and “External-emotional flow of information”.

The data is in four streams of metadata and forms the two States of Alien and Own, in the following Fig. 5. The most “dangerous” from the point of view of the company’s security information system - “External resonance” flow did not affect the object of research.

As can be seen from Fig. 5, the state of “Alien” is observed in an employee on weekends and has a stronger impact on the state of his motivation as an employee, which is confirmed by the state of “Own”. That is, at work in a team, he has the “restraining” forces of the stability of the psycho-emotional state of being a conscientious worker. All

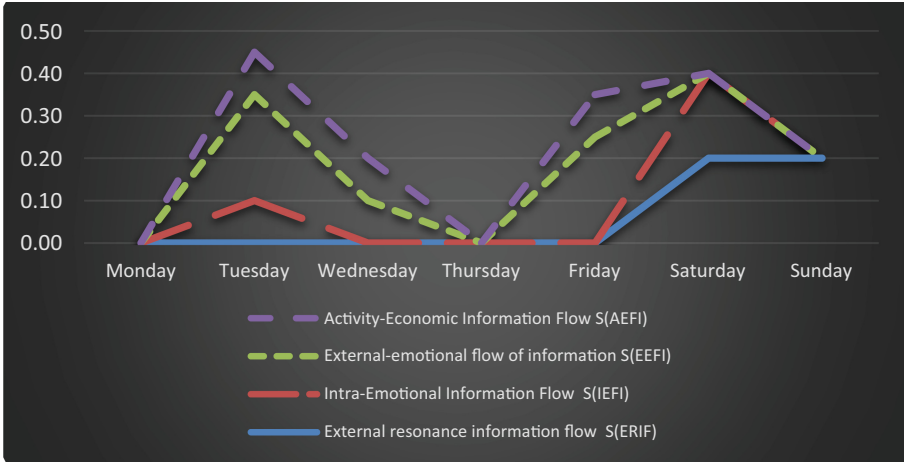


Fig. 4. The result of the effects of information flows on the Office Manager employee for 7 days of observations of the automated model.

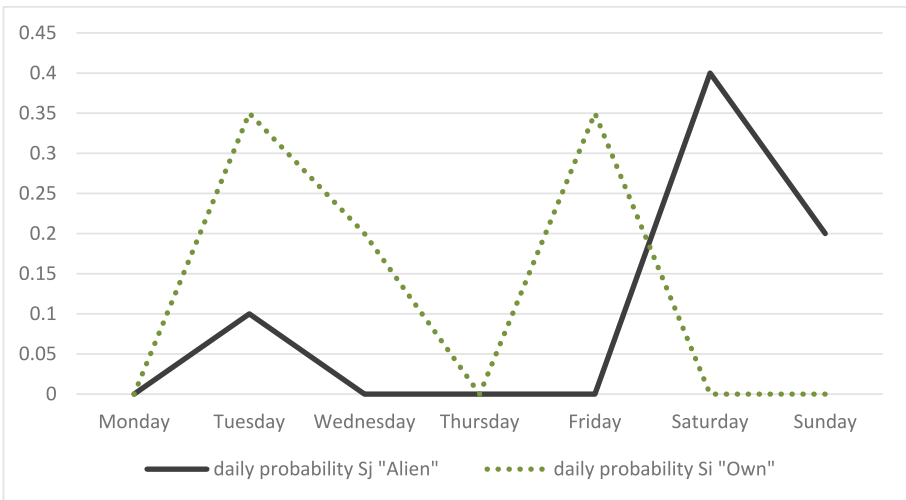


Fig. 5. Probabilities of the states "Own" and "Alien" for the office-manager employee for 7 days of observations of the semi-automated mathematical model.

values of probabilities did not exceed the established maximum probabilistic threshold of “malicious intent” - 0.77.

Therefore, based on the results of the collection and processing of metadata for the employee “Office Manager”, we have an employee to whom the management of the enterprise and the information security department can “trust” a trade secret for a period - a week ahead. The next measurement of information flows will either refute or confirm the status of this employee - you can verify the trade secret and be admitted to the business process of the company.

4 Conclusion

The article discusses the scale of metadata that can be offered to the information security department to collect about each employee, and also presents the concept of metadata processing using the Bayes formula and presents the results of calculating the probability of one of the employees - an office manager, how the final probability of his weekly state is calculated - either as a type of "Own", or transitional and the dangerous condition of a "Stranger" who, as having access to a trade secret, is already capable of criminal actions. (leakage, theft of information or negligence).

The main factors of metadata that need to be collected and taken into account from open and closed sources of information about each employee at the enterprise, for all employees with access to trade secrets and the business process of the company, are presented and deciphered, which allowed us to propose the concept of a multifactorial mathematical model for identifying an internal violator based on measuring the probability of risks of a transitional state for each employee of the company. From the "Own" type to the "Alien" type as an internal threat to trade secrets from the personnel of the enterprise, as well as the algorithm of employee admission inside the access system to the company's trade secrets. The appearance of an internal violator is a multifactorial, in fact, dynamic process that has its own typology and a set of corresponding threats that lead to such a phenomenon as an Internal Violator (often referred to as an insider), who, being under the influence of resonant factors, becomes a threat of leakage of the company's trade secrets.






References

1. Omar, M., Nawawi, A., Puteh Salin, A.S.A.: The causes, impact and prevention of employee fraud: a case study of an automotive company. *J. Financ. Crime* **23**(4), 1012–1027 (2016). <https://doi.org/10.1108/JFC-04-2015-0020>
2. Oliveira, A., Moro, S., P., Torres: Psychological contract, internal branding and employee turnover in an it company. *Acad. J. Interdisc. Stud.* **8**(1), 9–18 (2019). <https://doi.org/10.2478/ajis-2019-0001>
3. Alam, M., Said, J., Abd Aziz, M.: Role of integrity system, internal control system and leadership practices on the accountability practices in the public sectors of Malaysia. *Soc. Responsib. J.* **15**(7), 955–976 (2018). <https://doi.org/10.1108/SRJ-03-2017-0051>
4. Engelbrecht, A., Heine, G., Mahembe, B.: Integrity, ethical leadership, trust and work engagement. *Leadersh. Organ. Dev. J.*, **38**(3), 368–379 (2017). <https://doi.org/10.1108/LODJ-11-2015-0237>
5. Jaakson, K., Masso, J., Vadi, M.: The drivers and moderators for dishonest behavior in the service sector. In: Vissak, T., Vadi, M. (Ed.) (Dis) Honesty in Management Advanced Series in Management, Vol. 10, Emerald Group Publishing Limited, Bingley, pp. 169–193 (2013). [https://doi.org/10.1108/S1877-6361\(2013\)0000010012](https://doi.org/10.1108/S1877-6361(2013)0000010012)
6. FIRMEX: Spear phishing: who's getting caught? <https://www.firmex.com/resources/infographics/spear-phishing-whos-getting-caught/>. Retrieved 09 March 2021
7. Toapanta, M., Mafla, E., Benavides, B., Huilcpi, D.: Approach to mitigate the cyber-environment risks of a technology platform, march. In: The International Conference on Information and Computer Technologies (ICICT-2020) USA (2020). <https://doi.org/10.1109/ICICT50521.2020.00069>

8. Moussa Dioubate, B., Nurul Molok, N.A., Talib, S., Osman, Md., Ta, A.: Risk assessment model for organizational information security. *ARPJ J. Eng. Appl. Sci.* **10**, 23, 2006–2015 (ARPJ) (2015). ISSN 1819–6608
9. Suhartana, M., Pardamean, B., Soewito, B.: Modeling of risk factors in determining network security level. *Int. J. Secur. Appl.* **8**(3), 193 (2014)
10. Mandritsa, I.V., et al.: Study of risks of business information by stages of the business process of organization in the collection: problems of information security of socio-economic systems. In: VII All-Russian Scientific and Practical Conference With International Participation. Simferopol, pp. 20–29 (2021)
11. Martyanov, E.A.: The possibility of identifying an insider by statistical methods. *Syst. Means Inf.* **27**(2), 41–47 (2017)
12. Medvedev, V.I., Larina, E.A.: Struggle with internal threats. Identifying an Insider - “Current Accounting”, February (2014)
13. Shcheglov, A.Y., Shcheglov, K.A.: Mathematical models and methods of formal design of information systems protection systems. Textbook. – St. Petersburg: ITMO University 93 (2015)
14. Jean, D., Alben, T., Deqiang, H.: Total belief theorem and generalized Bayes’ theorem. In: 21st International Conference on Information Fusion (Fusion 2018), Cambridge, United Kingdom. Jul (2018).



Analysis of Influence of Byzantine Robots with Random Behaviour Strategy on Collective Decision-Making in Swarms

V. I. Petrenko , F. B. Tebueva , S. S. Ryabtsev  ^(✉), V. O. Antonov ,
and I.V Struchkov 

North-Caucasus Federal University, Stavropol, Russia
nalfartorn@yandex.ru

Abstract. The development of a decentralized type of management for military use, searching, exploration and monitoring actualizes issues of providing information security in swarm robotics systems. One of the unique and complex problems of information security in swarm robotics is the possibility of negative influence of Byzantine robots on collective decision-making, by voting during consensus for false alternatives. Existing studies are focused on the consideration by the Byzantine robots with the strategy of the behavior “against most” in the tasks with a binary choice. However, practical tasks often have a greater number of alternatives to the choice, and Byzantine robots are a greater diversity of the impact strategies on consensus achievement. This work is devoted to the study of the influence of Byzantine robots with a random strategy of behavior if there are two to five alternatives to the process of collective decision-making in swarm. The purpose of the study is to identify the laws of the influence of Byzantine robots with a random strategy of behavior on the effectiveness of rotary robotic systems. Modeling the operation of the swarm robotics system in the presence of Byzantine robots with random behavior and an assessment was made to reduce the effectiveness of consensus. The practical significance of the study lies in the laws received and information that can be used in the development and justification of the effectiveness of information security systems in swarm robotics systems.

Keywords: Swarm Robotics · Consensus Achievement · Collective Decision-Making · Information Security · Byzantine Robot

1 Introduction

Currently, more and more practical applications in areas requiring the coverage of large spaces and parallel performing tasks comes with a digital robotics. Swarm robotics is the study of how large number of relatively simple physically embodied agents can be designed such that a desired collective behavior emerges from the local interactions among agents and between the agents and the environment [1]. The current state of rock robotics, according to the study [2], can be characterized by the first civilian applications of swarm robotics system (SRS) in accurate farming, infrastructure inspection

and maintenance, and military applications mainly use unmanned aerial vehicles for sharing information and support. The features of SRS, allowing to achieve high efficiency in the specified areas are: Scalability, Members Simplicity, Local Interactions, Distributed Control Topology, Mostly Homogenous, Autonomy, Cooperation, Awareness, Coordination [3].

An important role is played by the possibilities of collective behavior: navigation behaviors, spatially organizing behaviors, collective decision-making (CDM) [4]. Nevertheless, SRS studies are largely laboratory and often do not take into account possible problems associated with the provision of information security (IS) of the CDM process. At the same time, CDM and the achievement of consensus (CA) are critical processes in SRS. This is due to the fact that. Any kind of an autonomous robot has to make a decision at some time. A decision is the selection between possible actions. CDM is the cognitive process of selecting an action from a number of alternatives. The perfect decision is an optimal decision which means that no alternative decision results in a better outcome [5].

In the investigation, a breakdown or cyber is a situation in which part of robots can be malicious. Under malicious robots in the current study uses the concept set out in studies [6, 7] – Byzantine Robot (BR), as a general term for describing robots, which demonstrate unintended or inconsistent behavior, regardless of the main reason, as an umbrella term to describe robots that show unintended or inconsistent behavior, independent of the underlying cause. Based on Byzantine fault-tolerance and the Byzantine Generals Problem) [8]. In practical applications of one or more BR, it can be sufficient to disrupt the work of the whole Roy, for example, by performing imposing inappropriate alternatives during CA at CDM.

In modern studies, the most often in the quality of BR is considered as work voting against the majority. At the same time, a variety of strategies of the effects of BR on CA are possible. For example, in Article [9], it is proposed to consider 3 possible types of BR behavior:

1. The contrarians are malicious robots that always oppose to the majority of the group.
2. The wishy-washy are malicious robots that keep changing their opinion every control loop. These robots ignore any information from the environment and the neighbors and just introduce a sort of noise into the swarm communications.
3. The sect is an organized group of zealots, which are malicious robots that ignore any information from the environment and the neighbors and keep communicating a constant opinion for a (possibly) inferior option.

However, in practice, a combination of completely different strategies of malicious behavior is possible, which actualizes research on the influence of various BR strategies on CDM. This paper is devoted to the study of the influence of BR with RBS. In this case, BR does not take into account the enforced environmental information and votes for a random opinion when choosing an alternative during CA. This study was analyzed experimentally, in the ARGOS simulation environment [10], the following questions:

1. What is the effect on the efficiency of BR with RBS on CA in SRS.
2. How it changes the effect of BR with RBS on the effectiveness of SRS with an increase in the number of signs of the external environment.

3. How it changes the effect of BR with RBS on the efficiency of SRS when the external environment is changed.
4. How it changes the effect of BR with RBS on the effectiveness of SRS when changing the stability of random opinion.

The purpose of the study is to identify the patterns of the effect of BR with RBS on the effectiveness of SRS. The practical significance of the study lies in the issues received and information that can be used in the development and justification of the effectiveness of IS security systems in SRS.

2 Materials and Methods

In this paper, studies aimed at studying the issues of the IS CDM process in SRS with BR with a random behavior strategy (BR with RBS) are carried out. For research, the CDM process considered the CA problem on the most common feature of the external environment (scene). Alternatives in this case serve as colors on the scene. In the overwhelming majority of modern research, such questions are considered on the binary, black and white scene, in the current study there are cases with a large number of colors: from two to five (white, black, red, blue, green). Thus, in the studies conducted, the goal of SRS was to take a collective solution and choose one of several actions on the basis of the external environment data A_i (voting for color i) with a certain number of BR with random behavior.

As a measure of measuring the complexity of the implementation of the SRS problem, the relationship between the most common color and other tiles on the scene is used, this ratio is chosen in mind the need to compare the complexity of a task with two colors with a large number of colors. The complexity is calculated by the formula:

$$Complexity = \frac{\sum_{N_1}^{N_i} P_i \setminus P_i^{max}}{P_i^{max}} / N_i - 1$$

where i – sequence color number, N_i – number of colors, P_i – interest ratio i -th colors on stage, P_i^{max} – the percentage of the most common color on the scene, $\sum_{N_1}^{N_i} P_i \setminus P_i^{max}$ – the amount of percentage ratios of all colors without P_i^{max} . In the case of an equally intended distribution of colors, as P_i^{max} you can choose any color.

If the complexity is set in such a way that it cannot be displayed with entire cells, the residue of cells is not taken into account and for convenience is painted in another color, which is not recognized by the SRS operation algorithms.

The complexity of the task can be varied by changing the ratio between the percentages of white tiles and other colors. In a simple task, the difference between the percentage of white and black tiles should be large. For example, if $P_1 = 0,72$; $P_2 = P_3 = P_4 = P_5 = 0,07$, complexity will be $\sim 0,1$. In a difficult task, on the contrary, the difference is small. For example, in the most difficult task for five colors, with their equiblibly distribution: $P_1 = P_2 = P_3 = P_4 = P_5 = 0,2$, The complexity will be 1.

For experiments, an ARGoS simulation medium was used [10]. To simulate tasks with a large number of colors and the scaling of the medium, a module was developed for ARGoS, which code which is available at reference [11]. The experiment scene is a room

bounded by 4 walls and size $S_{scene} = X \times X \text{ m}^2$, in the current experiments, the scene is square and varies from $2 \times 2 \text{ m}^2$ to $5 \times 5 \text{ m}^2$. SRS grouping consists of $N_r = 20$ robots e-puck, Moving over the surface marked with colored cells, and capable of perceiving the surface color under them, through a gray gradient. Robots have a diameter of 7 cm, a wheel platform with a maximum motion speed of 10 cm/s, RGB LED backlight, 8 approximation sensors, a surface color definition sensor, as well as a module for local exchange of information, consisting of 12 IR transceivers. Robots can communicate with each other only if the distance between robots is less than $d_n = 22 \text{ cm}$, to simulate physical limitations of SRS. The trajectory of the movement of each robot is a broken line - the robot alternates movement in direct and rotation in place.

At the beginning of the experiment, a scene of a given complexity is generated with a random arrangement of colors, robots are randomly placed inside the arena. An example of the initial configurations of the experiment of high complexity for a scene of 2 m^2 , 20 robots and 5 colors, shown in Fig. 1.

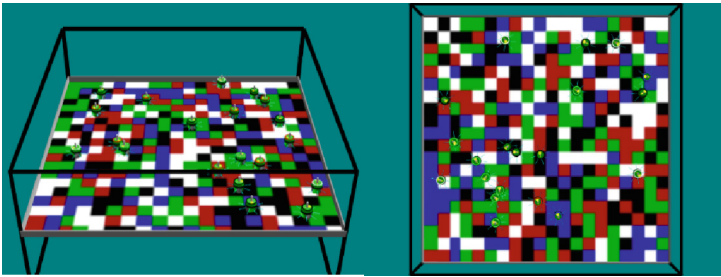


Fig. 1. Example experimental scene with 5 colors

The task of robots during the experiment is to achieve a general opinion relative to the prevalence of the number of cells of the same color if there are some BR with RBS. At the end of the successful launch, all separate SRS robots will have the same opinion corresponding to the most common color. Condition of exit from the experiment, i.e. CA condition is the reach of a quorum in 80% of the number of all robots.

The strategy of BR with random behavior, by analogy with the strategies described in chapter 6.1 [5] can be described as:

$$N_{\hat{i}j} = randN_{ij},$$

where N_{ij} , utility for each action A_i and state S_j , which is determined by the areas on the stage counted by the robots at the moment based on the ratio of colors, \hat{i} is the selected action [12].

To assess the effect of BR with RBS on efficiency SRS, the following metrics were used [6, 7]:

1. Probability of completion (E_N), i.e. the probability of making the best decision, calculated as the proportion of runs.

- Time of results completion ($T_N^{correct}$): this is the number of seconds it takes for the robots in the swarm to have a “white” opinion. The metric is calculated over all experimental runs converging to «white»; chains converging to any colors are not counted.

The influence of BR with random behavior on the known CDM methods with Ca: DC, DMMD and DMVD [13] is shown in Fig. 2. 100 runs are carried out if there is from 0 to 5 BR with random behavior (only 1,800 runs). Points in the figure, demonstrate the average value of the efficiency of SRS taking 50 experimental runs from the presence of BR with RBS. In this case, the task was viewed on two colors, as the original methods are given.

Detection of patterns of BR with RBS influence on SRS efficiency is further considered in more detail on the example of a method based on the use of a distributed registry without hash functions [14] (hereinafter DL). The influence of BR with RBS on DL under conditions similar to Fig. 2, are given separately in Fig. 3. 100 runs from 0 to 20 BR with RBS (only 2,100 runs) were performed.

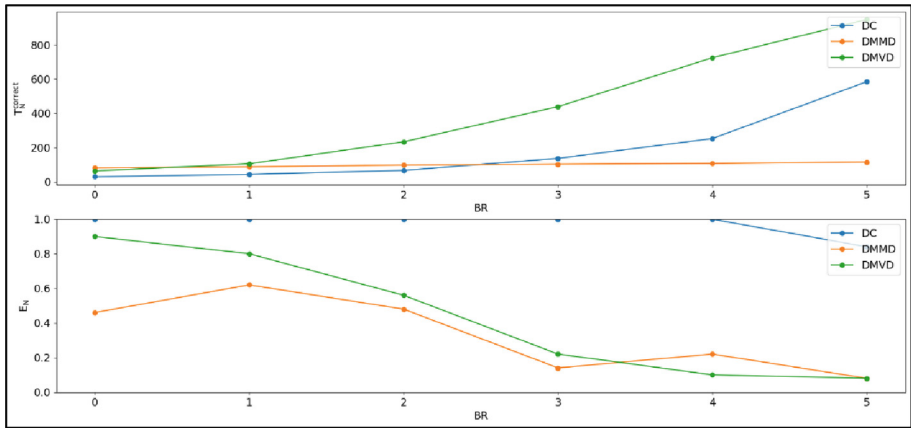


Fig. 2. Effect of 0–5 BR with RBS behavior on the effectiveness of DC, DMMD and DMVD methods, complexity 0.82

Analyzing the results obtained, it can be observed that the most resistant to the point of view of the concern (E_N) Methods to BR with RBS is the DC method, other methods tolerate a significant reduction in efficiency. Relative to the time of completion of the results ($T_N^{correct}$) It can be concluded that the presence of malicious robots with random behavior will populate the CA time in all cases. Two series of experiments were performed: with high (0.85) and low complexity (0.45). Each series is 100 of the runs to increase BR with RBS from 0 to 20 and for the number of colors from 2 to 5 (only 16,800 runs for 2 series). The results of the study are shown in Fig. 3.

Thus, the task of identifying patterns of the effect of BR with RBS on the effectiveness of SRS is relevant to further develop and construct accurate BR behavior models necessary to implement SRS protection systems. The purpose of this experiment is to assess the effect of BR with RBS on the effectiveness of SRS with an increase in the

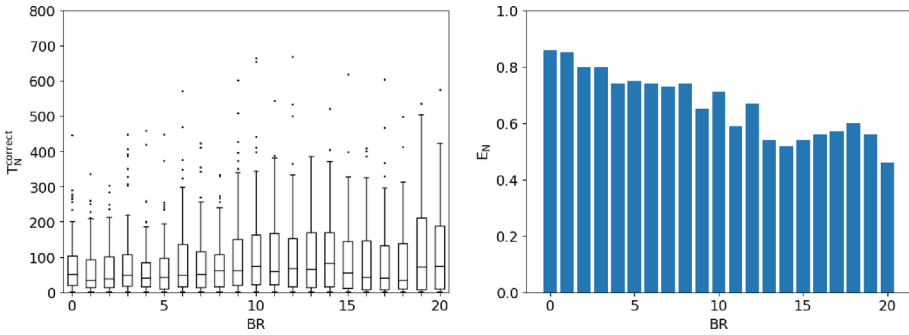


Fig. 3. The study of the effect of 0–20 BR with RBS on the effectiveness of the DL method, the complexity of 0.82

number of signs of the external environment. The hypothesis is that the influence of BR WITH RBS will be more significant with an increase in the number of alternatives, in the current case, the colors are available for choice. The points in the figure, demonstrate the average value of the efficiency of SRS taking 100 experimental runs from the presence of BR with RBS with increasing the number of alternatives available from 2 to 5.

3 Results

3.1 Influence of BR with RBS on Effectiveness of SRS with an Increase in the Number of Signs of the External Environment

The purpose of this experiment is to assess the effect of BR with RBS on the effectiveness of SRS with an increase in the number of signs of the external environment. The hypothesis is that the influence of BR with RBS will be more significant with an increase in the number of alternatives, in the current case, the colors are available for choice. Two series of experiments were performed: with high (0.85) and low complexity (0.45). Each series is 100 of the runs to increase BR with RBS from 0 to 20 and for the number of colors from 2 to 5 (only 16,800 runs for 2 series). The results of the study are shown in Fig. 4.

The points in the figure, demonstrate the average value of the efficiency of SRS taking 100 experimental runs from the presence of BR with RBS with increasing the number of alternatives available from 2 to 5. The results of the effect of BR with RBS on efficiency with an increase in the number of signs of the external environment was evaluated as a change $(E_N)_H (T_N^{correct})$ in a percentage relative to the values obtained in an experiment with 2 colors, with other things, the values are shown in Table 1.

In the easy tasks, the impact on the probability of Ca is rather weak and manifests itself only with a large number of Br, but there is a sufficiently large drop in efficiency in difficult tasks. In difficult tasks, the influence of BR when adding alternatives is significantly enhanced. In this case, the negative effect affects both the probability of CA and speed.

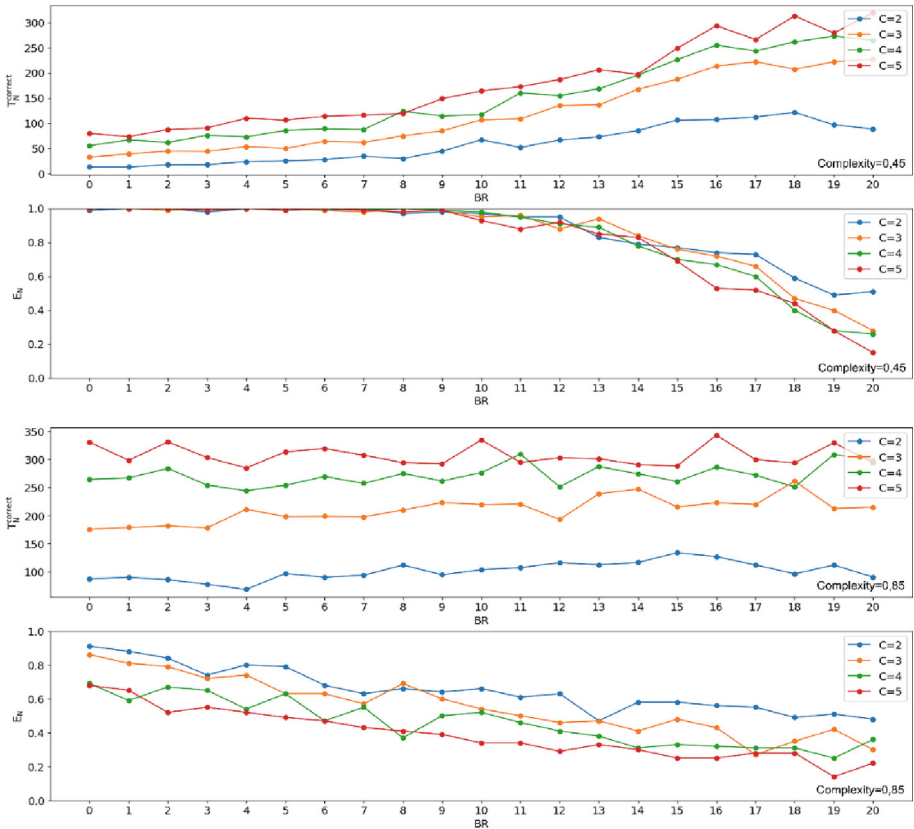


Fig. 4. Investigation of the effect of BR with RBS on the effectiveness of SRS with an increase in the number of signs of the external environment, complexity 0.45 and 0.85

Table 1. Influence of BR with RBS on the effectiveness if there is an increase in the number of signs

Complexity	Change efficiency	C = 3	C = 4	C = 5
0,85	E_N (%)	16,00147	30,495935	46,4974
	$T_N^{correct}$ (%)	111,9254	174,5997	286,5251
0,45	E_N (%)	3,9060	7,1658	11,0718
	$T_N^{correct}$ (%)	112,9402	194,0262	306,9664

3.2 Influence of BR with RBS on SRS Efficiency When Changing the External Environment

The purpose of this experiment to estimate the influence of BR with RBS on the efficiency of SRS when the external environment changes changed. The hypothesis is that the more

closely the contact of all SRS robots is, the more significant the influence of BR with RBS. As a scale of the medium, a characteristic of the scale is selected - the density of the scene of the scene by robots (M), which is characterized by the ratio of the size of the isna and the coating area of robots sensors and calculated as:

$$M = \frac{S_{scene}}{N_r * S_{d_n}},$$

where M – the density of the coating robots scene; S_{scene} – scene area; N_r – the number of robots; $S_{d_n} = \pi * d_n^2$ – square cover scene sensors 1 robot. In this way, $M \sim 1$ – means that the number of robots is enough to evenly fill the whole scene, $M < 1$ – means a high density of the scene of the scenes of robots sensors, and $M > 1$ opposite low. $M = 1.32$ conforms standard for this work of robots density: 20 robots on stage 2 m². In this experiment N_r remained constant and amounted to 20, S_{scene} changed from 1 to 5M². 100 runs are made for 0–20 BR with RBS and for 6 values M (only 12,600 experiments). The results of the study are shown in Fig. 5.

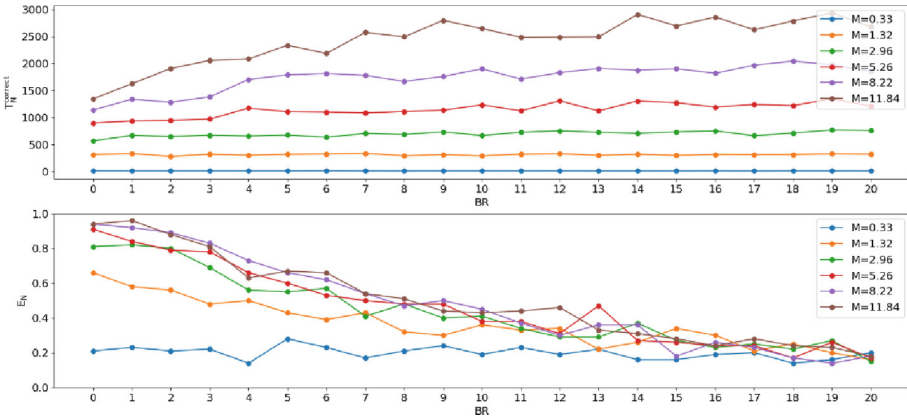


Fig. 5. Investigation of the effect of BR with RBS on the effectiveness of SRS when the external environment is changed

Points in the figure, demonstrate the average value of the efficiency of SRS taking 50 experimental runs from the presence of BR with RBS when the scene is changed, with five colors.

The results of the effect of BR with RBS on the efficiency with the scale of the medium scene was estimated as a change (E_N)и ($T_N^{correct}$)as a percentage relative to the values obtained in the experiment at $m = 1.32$, with other similar conditions, the values are shown in Table 2.

The values in the table show the change in the efficiency in percentages relative to the task with 20 robots on the scene 2×2 M². Negative values are associated with a risky CDM strategy due to which the solution is applied too quickly with errors; this suggests that with a large density of robots on a small area with a risky strategy, the likelihood of CA significantly suffers. It. With a small density of robots, opposite the probability of Ca increases, but with a large drop of speed.

Table 2. The effect of changes in the environment on the effectiveness of the CDM process of SRS in the presence of BR

Change efficiency	$M = 0.33$	$M = 2.96$	$M = 5.26$	$M = 8.22$	$M = 11.84$
E_N (%)	- 34,4484	18,27009	24,9696	26,5506	32,9998
$T_N^{correct}$ (%)	-69,6464	124,5783	270,2242	461,7412	686,9298

3.3 How to Change the Effect of BR with RBS on the Effectiveness of SRS When Changing the Stability of Random Opinion

The purpose of this experiment to estimate the influence of BR with RBS on the effectiveness of SRS with a change in the stability of a random opinion. The hypothesis is that with an increase in the stability of a random opinion, BR with RBS will increase and influence on the effectiveness of SRS. At the same time, with increasing stability, BR with RBS will be increasingly similar to BR type “Containers”. Thus, it is assumed that a hybrid strategy for holding an attack on CDM (BR with RBS and BR opponent) can have a greater effect on CDM SRS.

To describe the results, it is necessary to clarify the algorithm and the concept of sustainability of a random opinion, according to which BR with RBS works:

1. An array of numbers is initialized. The length of the array is equal to the number of colors for which BR (n). Can vote. This amount of color can be different from the number of colors presented on the scene. The first number in the array (R) is given as follows: $r_0 = rand(0, 1)$. In other words, it is simply initialized by a random real number from 0 to 1. Each next number in the array is given as follows: $r_i = r_{i-1} + rand(0, 1)$. In other words, the sum of the previous value of the array and a random number from 0 to 1.
2. Each stroke BR chooses the color as follows:
 - 2.1 Forms a random number K in the range from 0 to the maximum number in R . The maximum number in R is its last element R – this is his last element r_n . T.e. $k = rand(0, r_n)$.
 - 2.2 Next is checked in which range includes k . If a $0 \leq k \leq r_0$ – white color if $r_0 < k \leq r_1$ – color black and so on.

Thus, the stability of a random opinion is a frequency with which an array of R . If this step is 0, then it is updated only once, when generating the scene and has the highest possible stability, if 1, then every move is updated (more random Results), if 10, then R is updated once in 10 moves. Research results are shown in Fig. 6.

This experiment was carried out on a black and white field, with a complexity of 0.92 and with stability of equal to 1 (normal distribution time) 100 and 1000 for BR from 0 to 20. 3 series of experiments 100 runs with steps equal to 0 (normal distribution time) 100 and 1000. For BR from 0 to 20 (only 6,300 experiments).

The results of the effect of BR with RBS on the efficiency when changing the stability of a random opinion was estimated as a change (E_N)и ($T_N^{correct}$)as a percentage relation relative to the values obtained in the experiment at $Pace = 1$ with other similar conditions, the values are shown in Table 1. Evaluation of the influence RBS on the

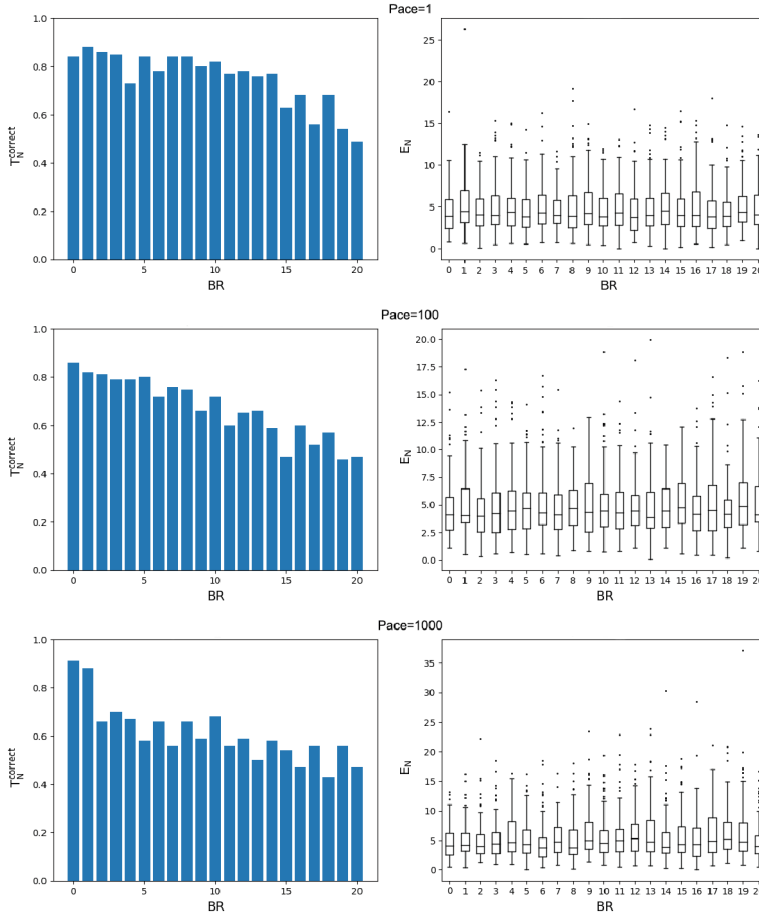


Fig. 6. The effects of BR with RBS on the effectiveness of the SRS when the step is changed, the stability of a random opinion

efficiency of SRS with a change in the stability of random opinion was carried out to relate steps 1, the obtained values are shown in Table 3.

Table 3. Influence of BR with RBS on the effectiveness of the CDM process with an increase in the stability of a random opinion

Change efficiency	$R = 100$	$R = 1000$
$E_N (\%)$	11,4636	18,8301
$T_N^{correct} (\%)$	32,91615	75,2354

There is a decrease in both CA speed and probabilities of CA, thus hybrid attack strategies may be more efficient.

4 Discussion

This paper discusses issues of IS concerning the effect of BR with SSP on the effectiveness of CDM in SRS. In the course of the study, about 39,600 experimental runs were carried out in the ARGoS simulator, in order to evaluate hypotheses: the influence of BR with RBS will be more significant with an increase in the number of alternatives; The more closely the density of all SRS robots, the more significant the influence of BR with RBS and the larger the stability of the Random opinion BR with RBS, the greater the effect on the efficiency of SRS. For research, an ARGoS supplement has been developed for modeling experiments with 3–5 colors, which can facilitate future research on CDM for situations with a large number of alternatives. According to the result, all hypotheses are confirmed, this suggests that BR is a significant problem in SRS.

Required patterns and information in practice can be used in the development and justification of the effectiveness of IS security systems and when developing the intruder models not only in SRS, but also similar systems with decentralized control.

5 Conclusion

This paper is devoted to the study of the effect of BR with RBS if there are two to five alternatives. The purpose of the study is to study the laws of the effect of BR with RBS on the effectiveness of SRS. The introduction is considered the current state of affairs in research on BR and formulate research issues of this work. The materials discuss the conditions and the scene of the experiment, settings, the method of assessing complexity, the procedure for conducting the experiment, as well as the assessment of the influence of BR with the SSP on the basic methods of CA. Also designed an addition to ARGoS allowing experiments on non-binary scenes with. The Results section contains 3 enlarged research groups in accordance with research issues. The criterion for estimating the scale of the scene is proposed - the density of the coating of the scene by robots and the concept of stability of accidental opinion - the measures of the periodicity of the random opinion BR. In total, about 39,600 experimental runs in the ARGoS modeling medium were carried out, which made it possible to obtain the numerical values of the effect of BR with RBS on CDM SRS and confirm all the extended research hypotheses.

References

1. Şahin, E.: Swarm robotics: from sources of inspiration to domains of application. *Lect. Notes Comput. Sci.* **3342**, 10–20 (2005)
2. Dorigo, M., Birattari, M., Brambilla, M.: Swarm robotics. *Scholarpedia* **9**(1), 1463 (2014)
3. Zakiev, A., Tsoy, T., Magid, E.: Swarm robotics: remarks on terminology and classification. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. vol. 11097, pp. 291–300 (2018)

4. Nedjah, N., Junior, L.S.: Review of methodologies and tasks in swarm robotics towards standardization. *Swarm Evol. Comput.* **50**, 100565 (2019)
5. Hamann, H.: *Swarm Robotics: A Formal Approach*. Springer International Publishing, Cham (2018). <https://doi.org/10.1007/978-3-319-74528-2>
6. Strobel, V., Castelló Ferrer, E., Dorigo, M.: Blockchain technology secures robot swarms: a comparison of consensus protocols and their resilience to byzantine robots. *Front. Robot. AI* **7**, 54 (2020)
7. Strobel, V., Ferrer, E.C., Dorigo, M.: Managing byzantine robots via blockchain technology in a swarm robotics collective decision making scenario: Robotics track. In: *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 541–549 (2018)
8. Lamport, L., Shostak, R., Pease, M.: The byzantine generals problem. *CM Trans. Program. Lang. Syst.* **4**(3), 382–401 (1982)
9. Canciani F., Talamali M.S., Marshall J.A.R., Reina A.: Keep calm and vote on: Swarm resiliency in collective decision making. *International Conference on Robotics and Automation*. <https://www.cl.cam.ac.uk/~asp45/icra2019/papers/Canciani.pdf>. Accessed 12 Aug 2022
10. Pinciroli, C., et al.: ARGoS: a modular, parallel, multi-engine simulator for multi-robot systems. *Swarm Intell.* **6**(4), 271–295 (2012)
11. Ncfu pmkb. (n.d.). ncfu pmkb/swarm-robotics GitLab. <https://gitlab.com/pmkb/swarm-robotics>. Accessed 11 Aug 2022
12. Tebueva, F., Ryabtsev, S., Struchkov, I.: A method of counteracting Byzantine robots with a random behavior strategy during collective design-making in swarm robotic systems. *E3S Web Conf.* **270**, 01034 (2021)
13. Hamann, H., Valentini, G., Dorigo, M.: Population coding: a new design paradigm for embodied distributed systems. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. vol. 9882, pp. 173–184 (2016)
14. Petrenko, V.I., Tebueva, F.B., Ryabtsev, S.S., Gurchinsky, M.M., Struchkov, I.V.: Consensus achievement method for a robotic swarm about the most frequently feature of an environment. *IOP Conf. Ser. Mater. Sci. Eng.* **919**(4), 042025 (2020)



Beamforming for Dense Networks-Trends and Techniques

Nabarun Chakraborty, Aradhana Misra, and Kandarpa Kumar Sarma (✉)

Department of Electronics and Communication Engineering, Gauhati University,
Guwahati 781014, Assam, India
kandarpaks@gauhati.ac.in

Abstract. Due to the proliferation of hand-held devices, such as smartphones and other wireless communication devices, spectrum and capacity in the field of wireless communication technology have become more scarce in recent decades. Long Term Evolution (LTE) and Long Term Evolution Advanced (LTE-A), often known as 4G, are the most recent addition to the rapidly expanding wireless technology as a result of the researchers' continued exploration of these technologies. In order to boost cell capacity, numerous tiny cells (femtocells, picocells) are placed next to larger macrocells in the mobile and associated networks being built around 4G and succeeding next generation technologies. The problem of interference and associated phenomena, which emerges owing to the mass roll out of smaller cells near huge macro cells, is one of the negative elements that stands as a barrier to this technology. One answer to the interference problem in dense networks is beamforming. In an effort to illustrate the interference issue in LTE/LTE-A dense networks, this study attempts to do so. As a response to the interference scenario, several beamforming techniques that help to lessen interference have been highlighted. A discussion of some of these beamforming techniques that have been previously suggested and made available in other works of literature is also included in this study.

Keywords: LTE/LTE-A · Beamforming · Microcells · Femtocells · Dense Networks · Co Tier Interference · Cross Tier Interference

1 Introduction

Unprecedented problems are being presented to developers by the rapid rise in demand for larger data rates, improved Quality of Service (QoS), and uninterrupted connection, together with the shortage of spectrum and the growing use of handheld devices like smartphones. If sufficient spectrum can be found, the need for increasing data capacity and throughput can continue to be satisfied. In popular view, it is generally understood that close proximity between the transmitter and receiver might result in greater QoS due to minimal transmission loss. Every user is constantly close to a Femto Base Station (FBS) or a Macro Base Station (MBS) because to the LTE/LTE-A system's deployment of tiny cells (femto cells, pico cells) around big cells (macro cells) [2, 3]. Due to great frequency reuse, sharing the same frequencies again, and enhanced connection

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

A. Alikhanov et al. (Eds.): APAMCS 2022, LNNS 702, pp. 217–232, 2023.

https://doi.org/10.1007/978-3-031-34127-4_21

and coverage at cell boundaries and inside settings, this deployment, also known as a heterogeneous or dense network, delivers higher connectivity, improved data rates, and better QoS [4]. The problem of interference between femto cells deployed inside a macro cellular border and between femto cells and macro cells is a harmful element that impairs the performance of network densification. In order to offer the user high QoS in an interference-limited environment, interference reduction has acquired importance [5]. Beamforming is a technique that is more effective than the traditional methods of cell splitting or cell sectoring for reducing interference [6]. This work largely focuses on beamforming techniques, and an overview of interference reduction in contemporary wireless scenarios has been given. The paper is further organized as follows Sect. 2 describes some important theoretical concepts, Sect. 3 gives an overview on femto-cells, Sect. 4 describes beamforming, Sect. 5 describes review of some works done on various types of beamformer, Sect. 6 describes the challenges faced during practical implementation of beamforming while Sect. 7 concludes the work.

2 Background

The desire to improve capacity and achieve high data rates is growing over time and doesn't appear to be going away any time soon. Higher order coding and modulation are needed for large data speeds. The majority of contemporary standards include orthogonal frequency division multiplexing (OFDM), which is characterised by high data rates and band split into smaller bands with additional use of tiny carriers to enable high data rates. Several wireless protocols, including Wi-Max, IEEE802.11a, LTE, and DVB, have adopted the OFDM technology in particular as a way to significantly improve future wireless communications. Since that OFDM is a broadband technology, it entails enormous data rates and is a significant phenomena in wireless communication. Yet, because an OFDM system's Symbol Time is shorter than the delay spread, this results in Intersymbol Interference (ISI). A specific type of multi-carrier (MC) transmission system called OFDM divides the whole bandwidth into smaller bands known as sub-bands, and each sub band contains a carrier known as a sub-carrier. This method is used to overcome ISI. In order to load data samples into sub-carriers that are orthogonal to one another, an OFDM system requires the loading of serial symbols into a serial to parallel converter.

After performing an IFFT on the symbols to create transmit samples, which are then gain parallel to serial converted and given a cyclic prefix to eliminate interblock interference, the samples are sent across a wireless channel. To recover the original samples, the reverse technique is used. Moreover, the complexity is cleverly reduced to just FFT processing and one tap scalar equalisation at the receiver [7] with the use of cyclic redundancy at the transmitter. The LTE/LTE-A systems now make considerable use of OFDM technology. Low latency and high throughput with carrier bandwidth define the LTE/LTE-A system (BW). Furthermore, it employs SC-FDMA for uplink transmission and OFDMA for downlink transmission, which significantly reduces co-channel interference. It is also capable of supporting both time division duplexing (TDD) and frequency division duplexing (FDD) on the same platform. Yet, it has been shown that in a particular setting, higher order modulation and coding methods are more vulnerable

to noise. The deployment of more channels per region, on the other hand, typically results in an increase in capacity (cell). This is made feasible by decreasing the size of each cell and so maximising channel reuse. Moreover, traditional techniques like cell sectoring and cell splitting are frequently utilised in modern wireless standards to boost system capacity.

Nonetheless, call loss in an indoor setting has been widely documented in recent years. As power loss occurs as a result of penetration losses in the walls and fading effects, increasing the power level towards an interior user is one way to address this issue. Nevertheless, focusing a lot of power on one indoor user reduces power for other users, which lowers system performance. The issue may be resolved by placing the transmitter and receiver close to one another, and LTE/LTE-A has fixed this. Incorporating various smaller cells, such as micro cells, femtocells, and pico cells with less power in the vicinity of macro cells so they can use the spectrum allotted to macro cell, is a new methodology adopted by LTE/LTE-A systems to increase system capacity, increase in the number of users per cell, as well as increase in data rate [5]. The following sections examine a few noteworthy femtocell and beamforming characteristics.

3 Femtocells

Femtocells are tiny, low-power base stations that give mobile users radio service when they're inside. Similar to a Wireless Fidelity (WiFi) router, they are placed indoors by the end user and give customers access to practically all cellular functionality. Fundamentally, a small cell radius reduces the distance between the transmitter and receiver, which results in less attenuation of the broadcast signal and better received signal strength for the receiver (RSS). Typically, the Signal to Interference Noise Ratio (SINR) is used to assess the quality of a signal at the receiver (SINR). The transmitted power from the targeted BS, the transmitted power from interference-causing transmitters, shadowing, fading, and route losses all affect the SINR. The weak interference signals are a result of the walls' penetration losses. Because to the higher frequencies employed in 3G technology's high data rate operation, attenuation there is more noticeable. These losses serve as insulation for the femtocells, allowing them to transmit with less power and still provide decent inside coverage [5, 8, 9].

3.1 Femtocells Deployment Modes

The distinguishing characteristic of an LTE/LTE-A system is the deployment of smaller cells around the edge of a larger one, increasing cell capacity. Because femtocell deployment is not done with proper network planning, that is, in a cognitive manner, and because network operators cannot control where femtocells are placed, their interference environment is much more complex than that of traditional cellular networks, which is a cause for concern. The technological difficulties are in managing interference between different femtocells installed inside a macro cellular network or between a femto cell and a macro cell [10]. There are two distinct ways that femtocells can be deployed close to macrocells [11]:

3.1.1 Separate Standalone Channel Deployment

In a distinct channel deployment option, the femto-cell network is given access to a particular channel or portion of the spectrum that the macrocell does not use. This is done to prevent femtocell and macrocell users from interfering with one another. Although this method is effective at reducing interference, it is expensive to set up and is not often favoured by operators.

3.1.2 Co-channel Deployment

The femtocell uses the same channels or spectrum that is allotted to the macrocell while in co-channel deployment mode. It is a preferred method because allocating a specific chunk of spectrum for femtocells might be costly because spectrum is a valuable and limited resource. The co channel deployment method also significantly boosts the system's total capacity, but there is a higher chance that femtocell and macrocell users would interfere with one another. Hence, interference management is a problem.

3.2 Femtocell Access Modes

Before femtocells may communicate with the current macrocell, they must be setup. Given that femtocells are low-powered small base stations that are typically user-deployed rather than network-deployed, certain facts should be noted, including the number of users who should receive the service, whether any foreign users who are not registered under a particular femtocell receive any service, and if so, for how long. There are three setting options to limit their use to a specific number of people depending on such criteria [12].

3.2.1 Open Access Mode

In open access mode, a femto cell base station (FBS) is configured to provide service to any user that enters the coverage area of that FBS. Any user who is not even configured under that certain FBS enjoys the benefit of that FBS in this mode and service is being rendered without any restriction. A certain drawback that has been reported in open access mode is that a large number of handover takes place from one femtocell to other thereby reducing the system performances. Moreover, soft handover is not permitted in femtocell scenerio, and during hard handover the amount of signalling increase which in turn increases the congestion in the network [13].

3.2.2 Closed Access Mode

In closed access mode, only a specific number of users are registered for a certain FBS and services are rendered to that specified group of users only. Under this scheme if any foreign user which is not registered under that certain FBS moves in the boundary of that FBS it does not receive any sort of service. In this approach, the hand-off is reduced but users may be denied of service if it moves out of macrocell coverage area.

3.2.3 Hybrid Access Mode

As the name suggests hybrid access mode is an amalgamation of both open access mode as well as closed access mode. Here, an user who is not registered under a certain FBS, if enters its vicinity can enjoy the services for a limited duration of time after which it shall be stripped of its services. So it is a combination of open and closed access modes. Moreover a femtocell performs identity request so that in hybrid access mode it should be able to differentiate between registered user and non-registered user in order to provide them with different level of QoS [13].

3.3 Femtocell Interference Scenario

A heterogeneous network, more commonly known as HetNet by the industry standard, is nothing but a tiered RAN cell structure that uses macrocell, and a combination of various types of smaller (micro, pico and femto) cell and several technologies together to offer wireless services. HetNets today can be stated to be a combination of radio access technologies, where the wireless standards like GSM/GPRS/CDMA/LTE/LTE-A will join hands with Wi-Fi. WiMax networks to form the multilayered heterogeneous networks [14]. The femtocell is a modern and emerging wireless technology that has been employed in LTE and LTE-A systems to offload macrocell traffic [14, 15]. But as a matter of fact, on deployment of femtocells or picocells in the boundary of a larger macrocells. The issue of interference becomes pertinent, which degrades the performance of dense networks and hence needs to be addressed. Moreover, because femtocells use licenced spectrum that wireless operators control and share with macrocell networks, reducing cross-tier interference from femtocell users at a macro-cell base station (MBS) is a need for the uplink deployment of femtocells. Interference in a heterogeneous network can take two different forms, which are explored in the next section [16].

3.3.1 Co-tier Interference Scenario

Co-tier interference is a term used to describe interference brought on by components of the same layer of the network. When two femtocells are installed inside the same macrocellular border, co-tier interference develops. A femtocell's interference is typically brought on by another femtocell nearby, hence effective cell design is required to prevent interference. Moreover, sub-channel allocation is a problem that has to be solved since interference arises when two close femtocell users utilise the same sub-channel. Thus, it is important to carefully design sub channel distribution to prevent co-tier interference. Once more, if interference causes the signal to interference plus noise ratio (SINR) to drop below a predetermined tolerable level, communication connections cannot be created, and the zone is deemed dead. In contrast to open access, co-tier interference is more severe in restricted access [5, 16, 17].

3.3.2 Cross- Tier Interference Scenario

This kind of interference is the disturbance brought on by network components from a different network tier or layer. In this instance, femtocells and macrocells placed inside the same macrocellular border experience cross-tier interference. The femtocells cover

the whole area around a macrocell and use the same portion of the frequency spectrum as the macrocells. Femtocells and macrocells must share the same radio spectrum because of spectral overlap and spectrum scarcity, which leads to cross-tier interference. In order to give better QoS to macrocells, femtocells should once again occupy a smaller frequency band, which reduces network throughput and efficiency and the locations where there will be significant interferences between femtocells and macrocells and the SINR will fall below a certain threshold thereby the zone shall be declared dead. The distance between the macrocell base station and the macrocell user devices and the unequal distribution of transmission power inside the network are the two factors that produce dead zones [16, 17]. Deploying femtocells and picocells near the edge of larger macrocells boosts system capacity and data rate and provides better and more efficient QoS, but interference still poses a challenge. The idea of beamforming has been discussed as a method for interference mitigation in order to lessen the impact of interference.

4 Beamforming

It is pretty clear from the previous section that LTE/LTE-A have adopted the idea of placing a number of smaller cells close to a large macro cell, which increases the capacity and improves interior coverage. Nonetheless, interference continues to be a barrier to network densification. In interference-prone areas, beamforming offers a solution and serves as a tool for interference cancellation. By focusing a beam in the direction of the intended user and producing nulls in all other directions, beamforming is a technique for eliminating interference in both cross-tier and co-tier interference. A downlink multi-antenna approach is beamforming. An FBS's transmitter weighs the data before transmission, generating focused beams that direct energy only at the target user while delivering a null in the other user's direction. An array of antennas may be directed in such a manner to restrict the reception of radio signals coming from specific directions using beamforming, a different name for spatial filtering. This can be done with the help of suitable analogue or digital signal processing. The beamformer is designed to combine spectral energy throughout its aperture, whereas a filter in the time domain combines energy over time. This results in a specific antenna gain in one direction while having attenuation in other directions. A processor used in conjunction with a number of sensors to offer a flexible kind of spatial filtering is another way to define a beamformer. It appears that beams are forming because of energy radiation. Nevertheless, beamforming may be used for both energy absorption and radiation [18]. Whereas beamforming denotes either radiation or the receiving of energy, beamforming indicates either radiation or the spreading of energy. As interference and the signal of interest share the same temporal band of frequencies, it is now a well-known fact that systems built to receive spatially propagating signals are frequently tainted by interference, making temporal filtering ineffective. Yet because the signal of interest and interfering noise often come from distinct geographical locations, they may be distinguished by using spatial filtering. The fundamental idea behind beam-formers is this [18, 19].

4.1 Classification of Beamformers

Due to its ability to cancel out interference in areas with dense networks that are prone to interference, beamformers have recently become extremely popular. A beamformer's goal is to provide a beam (or peak) in the direction of the intended user to whom service is to be supplied and nothing in any other direction, including those that include conflicting sources. These are some broad categories for beamformers:

4.1.1 Data Independent Beamformers

A data independent beamformer's weights are created so that regardless of the array data or data statistics, the beamformer response comes close to the intended response [18]. Data independent filter design is somewhat comparable to that of a Finite Impulse Response (FIR) filter. That is a time-tested strategy, and the method is now all but outdated. Data independent beamformers can be divided into the following categories once more:

Delay Sum Beamformer. One of the first beamformers with weights of similar size is the delay sum beamformer [19]. It is the oldest type of spatial filtering that employs an analog strategy and uses delays rather than phase shifters. In terms of narrowband signals, it is roughly equal to switched beam method, but it also has the flexibility or capacity to be used to broadband beamformer. The beam's phases shift in a certain direction as a result of the application of the delay (Fig. 1).

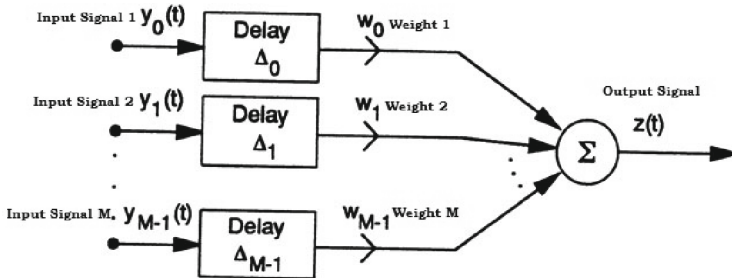


Fig. 1. A Delay Sum Beamformer

Null Steering Beamformer: Early beamformers that cancel plane waves coming from a given direction and generate a null in the response pattern in the plane wave's Direction of Arrival (DOA) were known as null steering beamformers. This procedure, which employs the idea of estimating signal arriving from a known source and positioning or directing a beam from a beamformer in the direction of that beam's arrival, has previously been described in DICANNE. The output of this step is then subtracted from each element. Lastly, the signal is approximated using sum and delay beamformers, where each element of the beamformer experiences a shift register delay. After a delay, each element's waveforms are added together and then subtracted [18, 19]. For numerous cancellations of powerful interferences, such a method was highly effective, but as the

number of interferers rises, the effectiveness of the strategy declines. This DICANNE constraint was overcome by null steering beamforming, which put nulls in the direction of interferences while directing a beam of unity response in the direction of the intended user. It is important to note that although null steering beamformers offer beams in the direction of intended users and nulls in the direction of interfering sources, they are unable to eliminate uncorrelated noise at array output (Fig. 2).

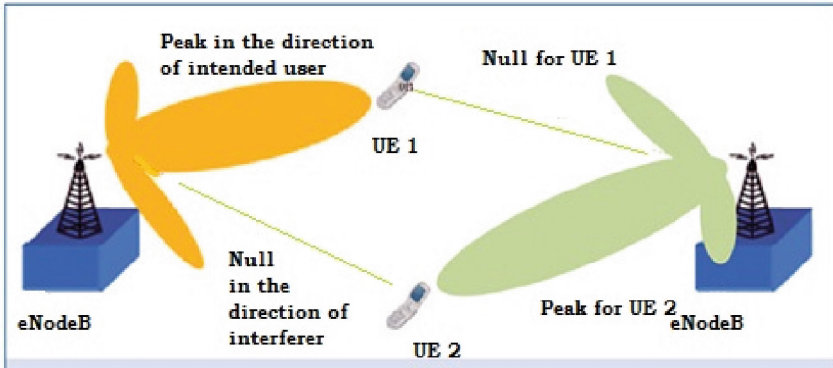


Fig. 2. Null Steering Beamformer

4.1.2 Statistically Optimum Beamformer

The weights are selected for statistically optimal beamforming based on the statistics of the data that was received at the array. Generally speaking, the objective is to improve the beamformer response such that interference signals and noise contributions to the output are as small as possible. We saw in the last section that the beamforming gain was fixed since the weights were fixed. Statistically optimal beamformers, where the weights alter or adapt to provide superior interference suppression, were proposed to enhance gain and for better suppression of interfering signal [22]. Wide-sense stationary (WSS) data are assumed, and second order statistics are known. In a word, we can state that the statistics of the data signal are used to calculate the beamforming weights. If “ x_n ” is the signal from an element and multiplied by weight w_n^* , it is feasible to determine the suitable signal using the mathematics of optimum beam- formation. It should be emphasised that the signal is multiplied by the weights’ conjugate and not by the weights themselves. Signal, interference, and AWGN are all combined in the received data. These statistically ideal beamformers are given below and can be categorised in the following ways:

Minimum Mean Squared Error (MMSE): Before describing MMSE technique we firstly define the interference and data covariance matrices as $R_n = E[nn^H]$ and $R = E[xx^H]$ respectively. The main aim of MMSE beamformer is to reduce the error with respect to reference signal. The reference signal be $d(t)$ where $\alpha = \beta d(t)$ and β is the

signal amplitude and $d(t)$, the reference signal is assumed to be known as receiver base station. The output signal $y(t)$ is required to track the reference signal $d(t)$ [20–22].

Minimum Output Energy (MOE) and Minimum Variance Distortionless Response (MVDR) Beamformer: The fundamental idea behind the MOE beam-former is to keep the gain of the array on the intended signal constant while minimising the total output energy. As a result, when gain is constant, output energy is decreased by suppressing interference, as the name implies, leading to energy minimization [20–22]. The fact that this method does not require a reference signal sets it apart from MMSE. These plans are referred to as blind beamforming plans.

Multiple Side Lobe Canceler (MSC): The multiple side lobe canceler is a well-established beamforming method [18] and is likely the first statistically ideal beam-former to be discovered. The idea is based on the design of an MSC beamformer, which features auxiliary channels in addition to a primary channel that might be an antenna with a high gain or a data independent beamformer. The primary channel antenna is very directional and is oriented in the general direction of the signal's arrival. It is expected that interfering signals enter through the side lobes of the main channel, however they can also enter through the auxiliary channel. According to the MSC beamformer's guiding concept, the auxiliary channel weights are selected such that they cancel out the interference signals from the primary channel. Such types of beamformer are quite useful when signal of interest is very weak. A few works related to General side Lobe canceler has been shown in [23, 24].

Linear Constrained Minimum Variance Beamformer (LCMV): For many complex applications, all of the beamformers outlined previously might not produce the desired results [18, 19]. There is now a need for an effective method that may be used in every circumstance. We are aware that the intended signal could always exist and might have an undetermined intensity, causing signal cancellation with the MSC. As a result, it is nearly difficult to estimate the signal and noise covariance matrix for the maximum SNR process, and the development of the reference signal method is also hampered by the lack of information about the desired signal. By imposing linear restrictions on the weight vector, LCMV beamformers overcome these drawbacks. The goal of the LCMV beamformer's design is to limit the beamformer's response such that the signal in the direction of interest may be amplified with a certain gain and phase, and the weights are therefore adjusted to minimise output variance and power under the restrictions. Such a beamformer is adaptable and durable, with the sole drawback being the laborious computation of the limited weight vector.

4.1.3 Adaptive Beamformers

Knowledge of second order statistics is necessary to determine the ideal beamformer weight vectors. These statistics are often unknown, but under the premise of ergodicity, one may determine the ideal weights using the information at hand. Statistics may also alter across time and space, as well as in response to interference that moves or changes location. Weights are often chosen via adaptive algorithms to address the problem of changing statistics [18]. Moreover, the co-relation matrix \mathbf{R} and its instantaneous value

R_N are absent in practise. As a result, it is difficult to calculate the array's ideal weights. Weights are now continually adjusted using adaptive methods to reduce inaccuracy and improve efficiency. An adjustable beamformer is shown in Fig. 3. Here are some adaptive algorithms that are discussed:

Least Mean Squares Algorithm (LMS): The LMS method, an adapted variation of the steepest descent algorithm, is regarded as the most readily implementable adaptive algorithm, however it has the downside of having weak convergence. At each iteration, the method changes the weights by estimating the quadratic surface's gradient and then shifting the weights very slightly in the opposite direction of the gradient, known as the step size [19, 21, 22]. Smaller step size improves performance and facilitates the discovery of ideal weights.

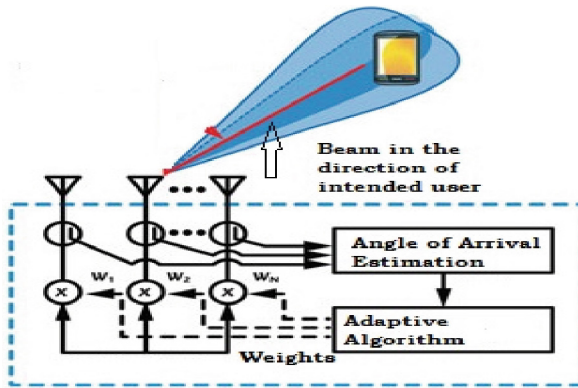


Fig. 3. Adaptive Beamformer

Recursive Least Squares (RLS): This is a different type of adaptive beamforming approach that collects and reduces error across all prior time indices in addition to minimising error for a specific time index. RLS has the benefit of converging more quickly than LMS, but its computational complexity is substantially higher than LMS. There have been a few reports on robust adaptive beamforming in [49, 50].

5 Review of Some Recent Works

Beamformers have long been regarded as a flexible tool for removing interferences. Multiple input multiple output (MIMO) systems were created to provide higher capacity, higher data rates, and better quality of service. However, because interference impairs the performance of MIMO systems as well as that of LTE/LTE-A systems, beamformers are now widely used to clear interference from these contemporary systems. Studies are currently focusing on statistically optimal as well as adaptive beamformer. In this section, a few techniques worth mentioning have been examined:

5.1 Zero-Forcing Beamforming (ZFBF)

Zero Forcing Beamforming is a spatial signal processing in multiple antenna wireless devices. In uplink, ZFBF algorithm allows a transmitter to send data to desired user together with nulling out in the direction where desired user is not present while in downlink it receives only from desired users while nulling in direction of interference users. Table 1 throws some light on various works done by various authors on ZFBF.

Table 1. Review on Zero Forcing Beamformers

Papers	Salient Features
[28]	<ul style="list-style-type: none"> • MISO channel ICI cancellation plan • Limited bandwidth partial CSI sharing across base stations • There are two forms of beamforming that have been discussed: egoistic ZFBF and altruistic ZFBF • Throughput is used as a performance metric, and it is compared to a complete CSI-based Monte-Carlo simulation
[29]	<ul style="list-style-type: none"> • Both Large Scale MIMO and Network MIMO employ the ZFBF for Orthogonality concept; comparative comparison of the two technologies • In LS-MIMO and network MIMO, the ZFBF's purpose is to reduce inter-cluster interference
[30]	<ul style="list-style-type: none"> • The goal is to reduce interference by utilising angle dimension • Measure angular information using the MUSIC algorithm, which operates on a PRB basis and then estimates DOA using ZFBF • Evaluation of the BER performance of ZFBF and traditional beamformers

5.2 Minimum Variance Distortionless Response Beamformer (MVDR)

MVDR beamforming was covered in the part before this one. It is a sort of statistically optimal beamformer, and when the MVDR beamformer weights are applied to an array's elements, they guide the sensor array's response in a predetermined direction of interest that is determined by the array's location. Its categorization is a little hazy because some academics believe that it is also an adaptable kind of beamforming. Table 2 lists the writings on MVDR beamforming approaches produced by various writers.

5.3 Linear Constrained Minimum Variance Beamformer (LCMV)

By using the idea of placing limitations along the target direction or in the direction of the arrival of the signal of interest, the LCMV beamformer, a flexible tool for interference removal, may be utilised to offer null steering. The fact that the LCMV beamformer lessens the likelihood of suppressing the signal of interest when it comes from a slightly different angle than predicted is a key aspect. Due to the LCMV beamformer's comparison of the target signal with a reference signal, beams are generated in the direction of the multipath signal that matches the target signal. The LCMV beamformer is perfect for non-line of sight (NLOS) environments and is appropriate for urban deployment since it not only lessens the effects of multipath fading and interference, but also helps

Table 2. Review on Minimum Variance Distortionless Response Beamformer

Paper	Salient Features
[32]	<ul style="list-style-type: none"> • To lessen interference, an iterative strategy of offering a null in the direction of the interferer has been proposed • The iterative approach is a non-adaptive methodology for removing macrocell to femtocell interference • Iterative null steering, optimum beam former (adaptive beam former employing MVDR approach), and constant null steering technique (non-adaptive) are compared
[33]	<ul style="list-style-type: none"> • Comparative examination of the MVDR and LCMV beamforming techniques is emphasised • The determination of weight affects both • One base station with a 4-element antenna array and a single M mobile user are taken into consideration • MVDR is not appropriate for use in cities since it cannot reduce the multipath effect, but it is more useful in rural regions • The output power level comparison performance metric reveals that MVDR exhibits better reaction than LCMV in a comparative research
[34]	<ul style="list-style-type: none"> • Focus is placed on the utilisation of the Clonal Selection Algorithm (Clonalg) of the Artificial Immune System in the Advanced MVDR Technique for Interference Cancellation (AIS) • The ULA, or uniform linear array, has been incorporated • Simulation demonstrates that the AIS aided MVDR technique has superior interference based on the output measure of received SINR

produce beams that point in the direction of the intended user. A few studies on LCMV beamforming algorithms are discussed in Table 3.

Table 3. Review on LCMV Beamforming

Paper	Salient Features
[35]	<ul style="list-style-type: none"> • A comparison of the MVDR and LCMV beamforming methods • Both beamforming algorithms depend on weight estimation • For simulation purposes, a base station with a 4-element antenna array and a single M mobile users is taken into consideration • In an urban setting, LCMV is an effective instrument for interference cancellation. It can also eliminate the multipath effect • The best option for an NLOS urban context is LCMV beamforming • When comparing output power levels, MVDR responds more quickly than LCMV beamformer.

(continued)

Table 3. (continued)

Paper	Salient Features
[36]	<ul style="list-style-type: none"> • The importance of calculating outage probabilities using an LCMV beamforming technique • Interferers are Rayleigh faded, and Rayleigh and Nakagami fading statistics are employed to describe the intended signal • The performance of wireless systems during outages using LCMV beamforming, which eliminates many significant interferers • According to numerical study, the LCMV beamforming's outage probability is highly dependent on the locations of the main interferers • The performance of LCMV beamforming and traditional beam-forming during outages is also compared, taking into account the directions in which the dominating interferers would arrive

6 Practical Challenges in Implementation of a Beamformer

Although beamformers have a wide range of applications and have shown to be adaptable interference mitigation tools, their actual implementation is still relatively challenging. The following are some of the difficulties encountered: -interferers.

- Although null steering beamforming is fairly simple to use for a single interference source, there are some implementation issues. Yet, when the number of interfering factors grows, the computational complexity increases and the performance deteriorates. Moreover, the null steering beamformer has a drawback in that it cannot eliminate uncorrelated noise. Optimal beamforming, sometimes referred to as Minimal Variance Distortionless Response (MVDR) beamforming, eliminates all of these problems.
- To reduce ICIC, In Zero-forcing beamforming partial channel state information (CSI) is required. The interchange of information leads in bandwidth usage, which is a key drawback for various beamforming techniques because gathering CSI information necessitates a robust backhaul network.
- One of the difficulties with adaptive beamforming is that in many real-world scenarios, the training data are frequently insufficient, which may result in significant performance degradation because it is impossible to form a trustworthy estimate of the covariance matrix without enough training data. The LCMV algorithm has quite complex mathematics and the co-variance matrix is not known in a realistic scenario.
- It has been theoretically and mathematically proven that the LMS algorithm has poor convergence and the computing complexity of the adaptive LMS and RLS schemes is quite high.

7 Conclusion

The continual desire for larger data rates, higher capacities, and better QoS has led the way for the creation of emerging standards to promote better, quicker, and more dependable communication. As time went on, other technologies emerged, but the quick rise in

mobile and handheld device use, spurred by spectrum shortages, prompted researchers to create a newer communication standard dubbed LTE/LTE-A. It tries to increase data rate and capacity by placing smaller cells close to bigger cells, which results in the growth of heterogeneous networks, also referred to as dense networks. Moreover, the cognitive deployment of several smaller cells increases interference, leading to the development of the beamforming idea for interference reduction. A beamformer is a spatial array that places a null in one direction and a peak in the direction of the intended user, eliminating interference from other users. There has been a thorough discussion of the numerous types of beamforming techniques, and it should be noted that the choice of beamformers may be greatly influenced by the situation at hand, the difficulty of the hardware implementation, and the data that is readily available.

References

1. Rappaport, T.S., et al.: Millimeter wave mobile communications for 5G cellular: it will work. *IEEE Access* **1**, 335–349 (2013)
2. Saquib, N., et al.: Interference management in OFDMA femtocell networks: issues and approaches. *J. IEEE Wire. Com.* **19**(3), 86–95 (2012)
3. Bartoli, G., Fantacci, R., Letaief, K.B., et al.: Beamforming for small scale deployment in LTE advanced and beyond. *J IEEE Wire. Com.* **21**(2), 50–56 (2014)
4. Baldemair, R., et al.: Ultra-dense networks in millimeter-wave frequencies. *IEEE Com. Mag.* **53**(1), 202–208 (2015)
5. Zahir, T., Arshad, K., Nakata, A., Moessner, K.: Interference management in femtocells. *IEEE Com. Surv. Tutorials* **15**(1), 293–311 (2013)
6. Emadi, M., et al.: Co channel interference cancellation by the use of iterative digital beamforming method. In: *Progress Electromagnetics Research, PIER-87*, pp. 89–103
7. de Courville, M.: Utilisation de bases orthogonales pour l’algorithmique adaptative et l’galisation des systmes multiporteuses. Dissertation, cole Nationale Suprieure des Tlcommunications (1996)
8. Islam, M.T., Ahmed, A.U., Singh, M.S.J., Ismail, M., Rahman, T.B.A., Misran, N.: Cognitive closed access femtocell application using multi-element antenna. *EURASIP J. Wirel. Commun. Netw.* **2015**(1), 1–7 (2015). <https://doi.org/10.1186/s13638-015-0314-5>
9. Jo, H.S., Mun, C., Moon, J., Yook, J.G.: Interference mitigation using uplink power control for two-tier femtocell networks. *IEEE Trans Wire Com.* **8**(10), 4906–4910 (2009)
10. Arif, M., Yameen, I.M., Matin, M.A.: Femtocell suburban deployment in LTE networks. *Int. J. Inf. Elec. Eng.* **3**(2), 88–95 (2013)
11. Letourneux, F., Corre, Y., Suteau, E., Lostanlen, Y.: 3D coverage analysis of LTE urban heterogeneous networks with dense femtocell deployments. In: *Proceedings of EURASIP J on Wire Com and Net.*, 319, pp. 1–14 (2012)
12. Perez, D.L., Valcarce, A., de la Roche, G., Zhang, J.: OFDMA femtocells: a roadmap on interference avoidance. *IEEE Com Mag.*, 1–8 (2009)
13. Hu, R.Q., Qian, Y.: *Heterogeneous Cellular Network*. Wiley IEEE Press, 1st edn., pp. 223–224 (2013). Cotanis, I., Hedlund, A.: *HetNets: Opportunities and Challenges*. An Ascom Network Testing White Paper: 1–22 (2013)
14. Mhiri, F., Sethom, K., Bouallegue, R.: A survey on interference management techniques in femtocell self-organizing networks. *J. Net Comp. Appl.* **36**, 58–65 (2013)
15. Chaudhary, K.R., Rawat, D., Madwal, E.: Interference Aware and SINR Estimation in Femtocell Networks. *J. Comp. Eng. (IOSR-JCE)* **10**(6), 64–69 (2013)

16. Ali, S., Ismail, M., Nordin, R.: Femtocell sleep mode activation based interference mitigation in two-tier networks. In: 4th Int Con Elec Engg and Informatics (ICEEI 2013), pp. 1088–1095 (2013)
17. Veen, B.D.V., Buckley, K.M.: Beamforming: A Versatile Approach to Spatial Filtering. *IEEE ASSP Mag.*, 4–24 (1988)
18. Godara, L.C.: Application of Antenna Arrays to Mobile Communications, Part II: Beam-Forming and Direction-of-Arrival Considerations. In: Proceedings of IEEE Antenna Arrays and Mob Com 85(8), pp. 1195–1245 (1997)
19. Bourdoux, A., Khaled, N.: Joint TX-RX optimisation for MIMO-SDMA based on a nullspace constraint. In: Proceedings. of IEEE Vehicular Technology Conference (VTC-02 Fall), pp. 171–174 (2002)
20. Reed, I.S., Mallett, J., Brennan, L.: Rapid convergence rate in adaptive arrays. *IEEE Trans. Aerospace Elect Syst.* **10**(6), 853–863 (1974)
21. Psaromiligkos, I., Batalama, S.: Interference-plus noise covariance matrix estimation for adaptive space-time processing of DS/CDMA signals. In: Proceedings of IEEE Fall Vehicular Technology Conference, pp. 2197–2204 (2000)
22. Huang, F., et al.: Robust partially adaptive array processing based on generalized side lobe canceller. In: IEEE Microwave Conference (APMC), Asia-Pacific (2008). <https://doi.org/10.1109/APMC.2008.4958554>
23. Lu, S., Liu, D., Sun, J.: A Distributed Adaptive GSC Beamformer over Coordinated Antenna Arrays Network for Interference Mitigation. *IEEE*, pp. 237–242 (2012)
24. Chen, S., Yao, W., Hanzo, L.: CMA and Soft Decision Directed Scheme for Semi-Blind Beamforming of QAM System. *IEEE*, pp. 1–5 (2008)
25. Li, D., Wang, Y.: Study of Smart Antenna Beam-former based on Constant Modulus Algorithm. In: IEEE 4th Int Con on Information and Computing, pp. 178–180 (2011)
26. Zhou, Z., Zeng, X., Ni, M.: Researches on low frequency source location based on partially adaptive subarray focused beamformer under finite sample-support. In: IEEE International Conference on Electronics and Optoelectronics (ICEOE 2011), pp. 403–406 (2011)
27. Lee, C., Moon, S.H.: Adaptive beamforming selection methods for inter-cell interference cancellation in multicell multiuser systems. In: Proceedings of IEEE ICC 2013 - Wireless Communications Symposium, pp. 5684–5688 (2013)
28. Hosseini, K., Yu, W., Adve, R.S.: Large-scale MIMO versus network MIMO for multicell interference mitigation. In: Proceedings of IEEE 15th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), pp. 70–74 (2014)
29. Bartoli, G., et al.: Angular interference suppression in cognitive LTE-A femtocells. In: IEEE Wireless Communications and Mobile Computing Conference (IWCMC), pp. 979–984 (2014)
30. Bartoli, G., Fantacci, R., Marabissi, D., Pucci, M.: Resource allocation schemes for cognitive LTE-A femtocells using zero forcing beamforming and users selection. In: Proceedings of Globecom-2014-Wireless Communication Symposium, pp. 3447–3452 (2014)
31. Emadi, M., Sadeghi, K.H. et al.: Co channel interference cancellation by the use of iterative digital beamforming method. *progress In Electromagnetics Research, PIER-87*, pp. 89–103 (2008)
32. Balasem, S.S., Tiong, S.K., Koh, S.P.: Beamforming algorithms technique by using MVDR and LCMV. In: Proceeding of Special section International E-Conference on Information Technology and Applications (IECITA) 2012 World Applied Programming, vol. 2(5), pp. 315–324
33. Balasem, S.S., Kiong, T.S., Paw, J.K.S., Hock, G.C.: Artificial Immune System Assisted Minimum Variance Distortionless Response Beamforming Technique for Adaptive Antenna System. *IEEE, ICTC*, pp. 938–943 (2013)

34. Li H, Yao YD, Yu J (2007) Outage Performance of Wireless Systems with LCMV Beamforming for Dominant Interferers Cancellation. In: Proceedings of IEEE Communication Society, pp. 190–195
35. Filho, D.Z., Cavalcante, C.C., Resende, L.S., Romano, J.M.T.: On Adaptive LCMV Beamforming for Multiuser Processing in Wireless Systems. SBMO/IEEE MTT-S (2007)
36. International Microwave and Optoelectronics Conference (IMOC 2007), pp. 521–525
37. Chahbi, I., Jouabe, B., Zeghlache, D.: Improving performance of ad hoc and vehicular networks using the LCMV beamformer. In: 2009 IEEE International Conference on Wireless and Mobile Computing, Networking and Communications, pp. 98–103. IEEE Computer Society (2009)
38. Hur, S., Kim, T., et al.: Millimeter wave beamforming for wireless backhaul and access in small cell networks. *IEEE Trans. Com.* **61**(10), 4391–4403 (2013)
39. Dowhuszko, A., Hmlinen, J.: Performance of transmit beamforming codebooks with separate amplitude and phase quantization. *IEEE Signal Process. Lett.* **22**(7), 813–817 (2015)
40. Park, S., Seo, W., Choi, S., Hong, D.: A beamforming codebook restriction for cross-tier interference coordination in two-tier femtocell networks. *IEEE Trans. Vehicular Technol.* **60**(4), 1651–1662 (2011)
41. Noh, J.H., Oh, S.J.: Beamforming in a multi-user cognitive radio system with partial channel state information. *IEEE Trans. Wire Com.* **12**(2), 616–625 (2013)
42. Lee, D., et al.: Coordinated multipoint transmission and reception in LTE-advanced: deployment Scenarios and Operational challenges. *IEEE Com Mag.*, 148–155 (2012)
43. <http://www.radio-electronics.com/info/cellulartelecomms/lte-long-term-evolution/4g-lte-advanced-comp-coordinated-multipoint.php>. Accessed on 29/01/2016
44. Yu, W., Kwon, T., Shin, C.: Multicell coordination via joint scheduling, beamforming and power spectrum adaptation. In: Proceedings of IEEE Infocom, pp. 2570–2578 (2011). <https://doi.org/10.1109/INFCOM.2011.5935083>
45. Lee, H.C., Oh, D.C., Lee, Y.H.: Coordinated user scheduling with transmit beamforming in the presence of inter-femtocell interference. In: Proceedings of IEEE Int Conference on Comm 2011 (2011). <https://doi.org/10.1109/icc.2011.5963055>
46. Zhu, J, Yang, H.C.: Interference control with beamforming coordination for two-tier femtocell networks and its performance analysis. In: Proceedings of IEEE Int Conference on Comm 2011 (2011). <https://doi.org/10.1109/icc.2011.5963130>
47. Nguyen, D.H.N., Le, L.B., Ngoc, T.L.: Joint multiuser downlink beamforming and admission control in heterogeneous networks. In: Proceedings of IEEE Globecom 2014 Wire Com Symposium, pp. 3653–3658 (2014). <https://doi.org/10.1109/GLO-COM.2014.7037375>
48. Lu, L., Wang, D., Liu, Y.: Joint user association power control and beamforming in HetNets via distributed SOCP. In: Proceedings of 8th Int Wireless Distributed Networks Workshop on Co-Operative and Heterogeneous Cellular Networks, pp. 358–363 (2015). <https://doi.org/10.1109/WCNCW.2015.7122581>
49. Mo, R., Quek, T.Q.S., Heath Jr., R.W.: Robust Beamforming and power control for two-tier femtocell networks. In: IEEE 73rd Vehicular Technology Conference (VTC Spring), pp. 1–5. (2011). <https://doi.org/10.1109/VETECS.2011.5956628>
50. Lin, G., Li, Y., Jin, B.: Research on the algorithms of robust adaptive beamforming. In: Proceedings of IEEE Int. Conference on Mechatronics and Automation (ICMA):751–755 (2010). <https://doi.org/10.1109/ICMA.2010.5589024>



Modified CLNet: A Neural Network Based CSI Feedback Compression Model for Massive MIMO System

Dikshita Sarma, Mrinmoy Shandilya, Aradhana Misra,
and Kandarpa Kumar Sarma^(✉)

Department of Electronics and Communication Engineering, Gauhati University,
Guwahati 781014, Assam, India
{aradhana.misra,kandarpaks}@gauhati.ac.in

Abstract. The downlink Channel State Information (CSI) is necessary for the precoder design in a massive multiple input multiple output (MIMO) system. This CSI, generally estimated at the user equipment (UE) is feedback to the base station (BS) by frequency division duplexing (FDD) which is a bandwidth consuming technique. In a massive MIMO system, the large antenna array size at BS may cause such feedback overhead to be very overwhelming. It is therefore necessary to utilize appropriate CSI compression with low computational complexity and high accuracy. This paper discusses a neural network based CSI compression method based on intrinsic properties of CSI and with a novel activation function. Named as CLNet+, the simulation results shows that the proposed method outperforms the existing CS-based and some DL-based methods.

Keywords: Massive MIMO · FDD · CSI · NMSE · Attention mechanism

1 Introduction

Massive MIMO is a key technology to increase the spectral and energy efficiency by deployment of huge antenna array at the transmitter and receiver. To fulfill the precoder design however, the BS requires the downlink CSI to be fed back to the BS from the UE side. In massive MIMO this may be a overwhelming process with large feedback overhead. With limited uplink bandwidth and transmit power, this huge feedback may prove to be a bottleneck and hence some form of CSI compression is required to adequately utilize the resources as well as convey the correct CSI to the BS.

The recent deep learning methods have made tremendous progress in the wireless communication including the CSI feedback. C. K. Wen et al. first applied deep learning to CSI limited feedback, and built a novel CSI sensing and recovery neural network (CsiNet) [1] which fits a transformation from CSI to codewords and an inverse transformation from codewords to CSI, which outperforms the

conventional methods of compressive sensing based CSI feedback and reconstruction. Qiuyu Cai, Chao Dong and Kai Niu introduced a long short-term memory (LSTM) network in the encoder and an attention mechanism [2] at the decoder, that can achieve high reconstruction rate. Jiajia Guo, Xi Yang, Chao-Kai Wen, Shi Jin, and Geoffrey Ye Li introduced deep learning-based CSI feedback and cooperative recovery framework, CoCsiNet [9] to reduce the feedback overhead, which is extremely constrained by this method. Zhenyu Liu, Lin Zhang, and Zhi Ding introduces a compression framework CQNet to jointly tackle CSI compression, codeword quantization, and recovery under the bandwidth constraint [10] It provides better feedback efficiency and reconstruction accuracy. Jiajia Guo, Chao-Kai Wen, Shi Jin, Geoffrey Ye Li presented a multiple-rate compressive sensing neural network framework to compress and quantize the CSI [4] It reduced parameter number compared with the existing methods that train and store a different neural network for a different CR. DSNLCsiNet [3] by taking advantage of non-local blocks, DS-NLCsiNet can capture long-range dependencies efficiently. The computational overhead of DS-NLCsiNet was approximately x2.5 higher than the original CsiNet. Recently, some methods start to reduce the complexity, for like BcsiNet [8]. However, their performance also degraded. So far, only channel reconstruction network CRNet [4] which extracts CSI features on multiple resolutions, outperforms CsiNet without increasing the computational complexity. This work is mainly on the inherent characteristic of CSI data and considering the limited computation resource and limited storage at UE side, and design a lightweight framework CLNet as proposed in [5], a modified version, CLNet+, has been discussed and analyzed.

The rest of the paper is organized as follows: Sect. 1 gives the introduction and the state of the art on deep learning based CSI compression techniques. Section 2 describes the massive MIMO model with the CLNet design. Section 3 discusses the results obtained and Sect. 4 concludes the work.

2 Massive MIMO System Model

The simulations in this paper is considered in a single cell massive MIMO system with N_t transmit antenna at BS and N_r receive antenna at UE, where $N_t \gg 1$ and $N_r = 1$ for simplicity of calculations. Orthogonal Frequency Division Multiplexing (OFDM) with N_c subcarriers is considered. The received signal r_n on the n^{th} sub-carrier can be expressed as,

$$r_n = h_n^H p_n s_n + w_n \quad (1)$$

where s_n is the transmitted symbol and w_n denotes the additive Gaussian noise. $h_n \in \mathbb{C}^{N_t \times 1}$ is the vector denoting the channel gain and the precoding vector is denoted as $p_n \in \mathbb{C}^{N_t \times 1}$. The CSI information of all sub-carriers can be represented by matrix $\mathbf{H} \in \mathbb{C}^{N_t \times N_c}$ such that,

$$\mathbf{H} = [\mathbf{h}_1 \ \mathbf{h}_2 \ \mathbf{h}_3 \ \dots \ \mathbf{h}_{N_c}]^H \quad (2)$$

The precoder design at the BS requires knowledge of the CSI, which is feedback from the UE. It is assumed here that the UE estimates this required CSI data through pilot based training and detailed discussion of CSI estimation is considered to be out of scope of this paper. The UE has to feedback $2N_c N_t$ parameters consisting of both real and imaginary CSI data. Since a massive MIMO system involves huge antenna array size at the BS, it causes a huge feedback overhead and hence increased bandwidth utilization.

To explore the sparsity of the channel matrix in the angular-delay domain, \mathbf{H} is subjected to 2D Discrete Fourier Transform (DFT) to convert it from the original spatial frequency domain \mathbf{H} to angular delay domain \mathbf{H}' ,

$$\mathbf{H}' = \mathbf{D}_c \mathbf{H} \mathbf{D}_t^H \quad (3)$$

where \mathbf{D}_c and \mathbf{D}_t are the DFT matrices with dimensions $N_c \times N_c$ and $N_t \times N_t$, respectively. The channel matrix \mathbf{H}' contains useful information only in the first N_a rows and the remaining row values are near zero which can be neglected without any loss of information. Thus a reduced matrix \mathbf{H}_a can be obtained from \mathbf{H}' containing the informative rows and this reduced data is put as input to the encoder to generate the codeword \mathbf{m} with compression ratio η ,

$$\mathbf{m} = f_{enc}(H_a, \Theta_{enc}) \quad (4)$$

where f_{enc} is the encoding process and Θ_{enc} , represents a set of encoder parameters.

The channel is reconstructed at the decoder as,

$$\hat{H}_a = f_{dec}(m, \Theta_{dec}) \quad (5)$$

where f_{dec} and Θ_{dec} , represents the decoding process and the set of decoder parameters. So the feedback process can be expressed as,

$$\hat{H}_a = f_{dec}(f_{enc}(H_a, \Theta_{enc}), \Theta_{dec}) \quad (6)$$

The reconstructed channel matrix \hat{H}_a should be as close as possible to H_a . Hence, the encoder decoder design should be done in a way to minimize the difference between \hat{H}_a and H_a . Hence, the parameter sets $(\Theta_{enc}, \Theta_{dec})$ may be expressed as,

$$(\hat{\Theta}_{enc}, \hat{\Theta}_{dec}) = \arg \min_{\Theta_{enc}, \Theta_{dec}} \|H_a - f_{dec}(f_{enc}(H_a, \Theta_{enc}), \Theta_{dec})\|_2^2 \quad (7)$$

The performance of the CSI feedback scheme highly depends on the compression part at the encoder. If the information loss is less in the compression, the decompression accuracy will be high.

3 Neural Network Model Description

In this section, first the CLNet as proposed in [5] is discussed followed by the improved version CLNet+ which is proved to have improved NMSE performance (Fig. 1).

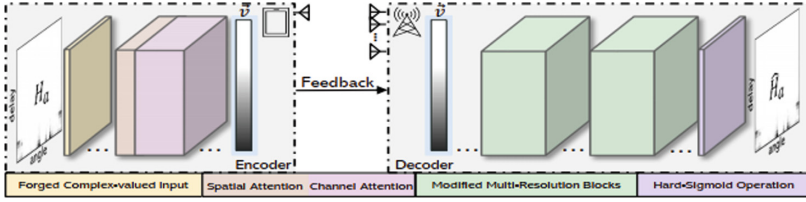


Fig. 1. architecture of clnet [5]

3.1 Forged Complex Valued Input Layer

In typical deep learning environment, the complex CSI data is separated into two parts: the real and imaginary parts and mixed convolution is applied. This however reduces the significance and properties of the complex valued CSI. To overcome this issue, the forged complex-valued input layer of CLNet [5] employs multiple 1×1 convolutional filters to encode the real and imaginary parts of each complex-valued element in \mathbf{H}_a with respective learnable weights.

$\mathbf{F}_{tr} : \mathbf{H}_a \rightarrow I$ is a convolutional transformation. $\mathbf{H}_a \in \mathbb{R}^{N_a \times N_a \times 2}$ is a 3D tensor, where the third dimension express the real and imaginary parts, and $I \in \mathbb{R}^{N_a \times N_a \times C}$, where C represents the number of convolutional filters applied to learn different weighted representations. If $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_c]$ presents the set of filter kernels, where \mathbf{f}_c is the learnable parameter of the c -th filter. The output of the transformation is $I = [\mathbf{i}_1, \mathbf{i}_2, \dots, \mathbf{i}_C]$, $\mathbf{i}_C \in \mathbb{R}^{N_a \times N_a}$, where

$$\mathbf{i}_c[m, n] = \mathbf{f}_c * \mathbf{H}_a = \sum_{i=1}^m \sum_{j=1}^n \sum_{k=1}^2 \mathbf{f}_c[i, j] \mathbf{H}_a^k[m - i, n - j] \quad (8)$$

Based on the trade-off between accuracy and model size, the model considers $C = 32$ as learnable filters. The complex valued input layer is shown in Fig. 2.

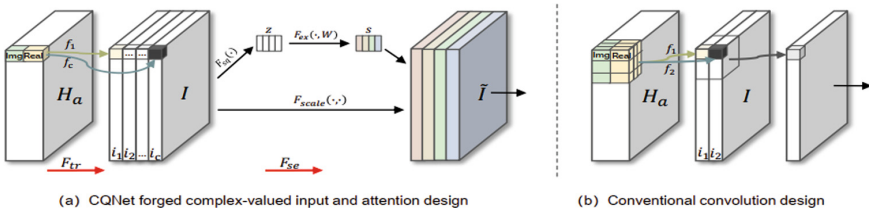


Fig. 2. Forged complex valued input and attention mechanism [5]

3.2 Attention Mechanism

CLNet in [5] introduces the SE block [6], for the channel attention to focus on the important features and suppress the unnecessary ones. The forged input layer goes through a F_{sq} transformation by global average pooling to obtain channelwise statistics descriptor, $z \in \mathbb{R}^C$

$$z_c = \mathbf{F}_{sq}(\mathbf{i}_c) = \frac{1}{N_a^2} \sum_{i=1}^{N_a} \sum_{j=1}^{N_a} \mathbf{i}_c(i, j) \quad (9)$$

where $c \in \{1, 2, \dots, C\}$. The transformation expands the receptive field to the whole angular delay domain to obtain the global statistical information. After that the channel goes through a transformation \mathbf{F}_{ex} by a gated layer with sigmoid activation to learn the non linear interaction and non mutually exclusive relationship between the channels,

$$s = \mathbf{F}_{ex}(\mathbf{z}, \mathbf{W}) = \sigma(g(\mathbf{z}, \mathbf{W})) = \sigma(\mathbf{W}_2 \delta(\mathbf{W}_1 \mathbf{z})) \quad (10)$$

where δ is the ReLU function. \mathbf{F}_{ex} transformation models the inter-channel dependencies obtain the calibrated \mathbf{s} , the attention vector that includes intra-channel and interchannel dependencies. Then each channel of I is scaled by attention value,

$$\tilde{I}_{:,i} = \mathbf{F}_{scale}(\mathbf{s}, I) = \mathbf{s}_i I_{:,i} \quad (11)$$

Spatial attention focuses on learning the places of the more informative parts across the spatial domain to tackle the cluster effect by a CBAM BLOCK [7], which uses two pooling operation average pooling and max pooling, to generate feature maps $\mathbf{F}_{avg} \in \mathbb{R}^{N_a \times N_a} \times 1$ and $\mathbf{F}_{max} \in \mathbb{R}^{N_a \times N_a} \times 1$ respectively as shown in Fig. 3.

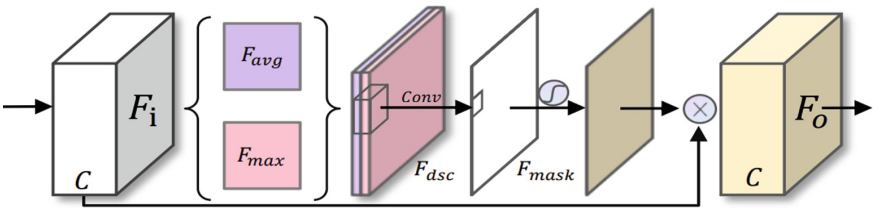


Fig. 3. Spatial attention mechanism [5]

CLNet joins the two feature maps to generate a compressed spatial feature descriptor $\mathbf{F}_{dsc} \in \mathbb{R}^{N_a \times N_a} \times 2$, and convolves it with a convolution layer to produce a 2D spatial attention mask $\mathbf{F}_{mask} \in \mathbb{R}^{N_a \times N_a} \times 1$. The Sigmoid activation

function is applied on the mask and then multiplied with the original feature maps to obtain with spatial-wise attention. CLNet uses the hard Sigmoid, its piece-wise linear analogy function,

$$h\sigma(x) = \frac{\min(\max(x + 3, 0), 6)}{6} \tag{12}$$

3.3 Modified CLNet/CLNet+

To make the decoder lightweighted, CLNet modifies the compression block[] by reducing the filter size from 1×9 to 1×3 . Further, the proposed CLNet+ modifies the CLNet structure by applying the hyperbolic tangent (tanh) operation which range is $[-1, +1]$ instead of $[0, 1]$, so that it can reduce the error and can give high reconstruction accuracy.

$$f(x) = \tanh(x) = \frac{2}{(1 + e^{-2x})} - 1 \tag{13}$$

The gradient of tanh is much higher than the sigmoid function, As shown in Fig. 4, so the error can be lowered. The architecture of CLNet+ is shown in Fig. 5.

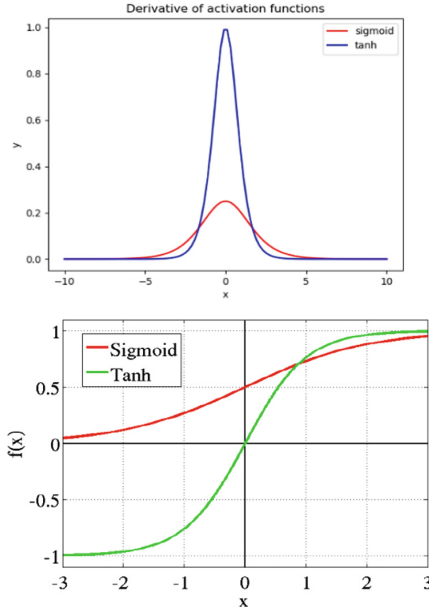


Fig. 4. Comparison graph between sigmoid and tanh

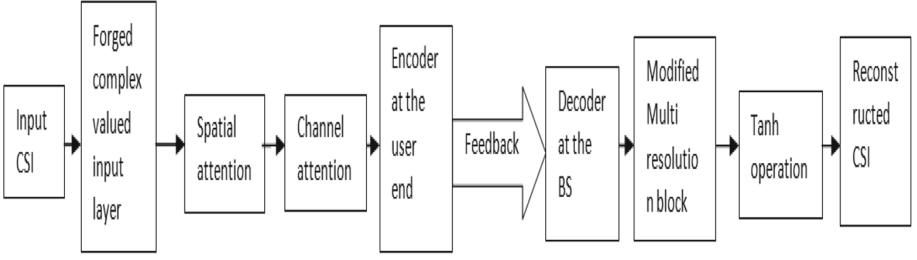


Fig. 5. Architecture of CLNet+

4 Results and Discussion

This section provides a discussion on the simulation environmental setup and a comparison of the results obtained for the proposed modified CLNet with the state of the art methods.

4.1 Dataset Consideration

The dataset used in this work is COST 2100 channel model for indoor picocellular scenario at 5.3 GHz, which is the same as used in previous works on deep learning-based Massive MIMO CSI feedback [1, 5] for a fair comparison. The CSI matrices are converted in angular delay domain \mathbf{H}_a by 2D DFT. The total 150,000 generated CSI are split into three parts, 100,000 for training, 30,000 for validation, and 20,000 for testing, respectively.

4.2 Training Scheme

CLNet is first trained with the batch size of 200, 8 workers and 100 epochs on a single NVIDIA 2080Ti GPU. In the next stage it has been trained for 1500 epochs instead of 1000 and batch size is taken as 600 instead of 200, so that we can expect better result. In the third stage CLNet+ is trained for 1500 epochs and batch size 600. To evaluate the performance, the normalized mean square error (NMSE) is measured between the original channel and the reconstructed channel as

$$NMSE = E\|\mathbf{H}_a - \hat{\mathbf{H}}_a\|_2^2 / \|\mathbf{H}_a\|_2^2 \quad (14)$$

The simulation environment for the purpose of training is as shown in Table 1.

A test run of the simulation set up with CLNet is done with 100 epochs and batch size of 200 initially to check the consistency of results and the same has been shown in Table 2. The best NMSE is obtained at compression ratio of 1/4 in the indoor scenario. Similar performance is seen with 1500 epochs and batch size of 600 in the indoor scenario as seen from the same table.

Table 1. Simulation parameters and considerations

Parameters	Values
Scenario	Indoor
Squared area length	20 m
Bandwidth	5.3 GHz
Tx antenna array size	32
Subcarriers	1024
Optimizer	Adam
Batch size	600
Learning rate	0.001

Table 2. Performance comparison of CLNet with different compression ratios, epochs and batch size

η	epochs = 100 batch = 200			epochs = 1500 batch = 600		
	NMSE	Test Loss	Test rho	NMSE	Test Loss	Test rho
1/4	-27.79	8.063e ⁻⁰⁷	0.99	-28.92	6.215e ⁻⁰⁷	0.99
1/8	-14.17	1.906e ⁻⁰⁵	0.97	-15.66	1.344e ⁻⁰⁵	0.98
1/16	-9.821	5.264e ⁻⁰⁵	0.94	-8.557	1.138e ⁻⁰⁵	0.90
1/32	-8.251	7.536e ⁻⁰⁵	0.92	-8.420	7.05e ⁻⁰⁵	0.93
1/64	-4.839	1.600e ⁻⁰⁴	0.82	-6.396	1.130e ⁻⁰⁴	0.88

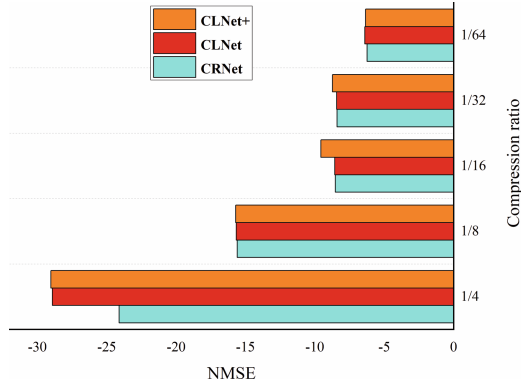
In Table 3, the NMSE performance of CLNet+ has been shown for different compression ratios and the result shows that NMSE for $\eta = 1/4$ is the best. This simulation has been done with 1500 epochs and a batch size of 600. Table 4 compares the performance of CLNet+ with state of the art models and it is seen that CLNet+ performs better than the previous CsiNet, CS-CsiNet+ and CRNet. A comparative graph representation of the NMSE for CRNet, CLNet and CLNet+ is shown in Fig. 6 which strongly recommends the use of CLNet+ proposed in this work as the compression model when the compression ratio is kept below 1/64.

Table 3. Performance of the CLNet+

Compression ratio	NMSE	Test loss	Test rho
1/4	-29.02	6.239e ⁻⁰⁷	0.99
1/8	-15.71	1.377e ⁻⁰⁵	0.98
1/16	-9.567	1.157e ⁻⁰⁴	0.92
1/32	-8.720	7.362e ⁻⁰⁵	0.94
1/64	-6.34	1.130e ⁻⁰⁴	0.88

Table 4. Comparison of CLNet+ with different state of art models

Compression ratio	Models	NMSE	Cosine similarity
1/4	CS-CsiNet [1]	-11.82	0.96
	CsiNet [1]	-17.36	0.99
	CRNet [4]	-24.10	0.97
	CLNet [5]	-28.92	0.99
	CLNet+	-29.02	0.99
1/16	CS-CsiNet [1]	-6.09	0.66
	CsiNet [1]	-8.65	0.93
	CRNet [4]	-8.52	0.89
	CLNet [5]	-8.557	0.90
	CLNet+	-9.567	0.92
1/32	CS-CsiNet [1]	-4.67	0.83
	CsiNet [1]	-6.24	0.89
	CRNet [4]	-8.40	0.90
	CLNet [5]	-8.420	0.93
	CLNet+	-8.720	0.94
1/64	CS-CsiNet [1]	-2.467	0.68
	CsiNet [1]	-5.84	0.87
	CRNet [4]	-6.23	0.87
	CLNet [5]	-6.396	0.88
	CLNet+	-6.340	0.88

**Fig. 6.** Comparison graph of NMSE of CRNet, CLNet, CLNet+

5 Conclusion

In this paper a CSI compression and recovery mechanism is used which is a modified version of CLNet. It is based on the physical properties of the CSI and its usefulness in reduction of NMSE is demonstrated. Simulation has been carried out with the modified decoder with smaller filter size and the use of a novel activation function to minimize the NMSE in an outdoor scenario. The overall performance of the CLNet+ has 5.30% higher accuracy than CRNet and 2 % higher than CLNet. As the CLNet+ is lightweighted and UE's resource is limited, it can be applied in practical scenarios.

References

1. Wen, C.-K., Shih, W.-T., Jin, S.: Deep learning for massive MIMO CSI feedback. *IEEE Wirel. Commun. Lett.* **7**(5), 748–751 (2018)
2. Cai, Q., Dong, C., Niu, K.: Attention model for massive MIMO CSI compression feedback and recovery. In: 2019 IEEE Wireless Communications and Networking Conference (WCNC). IEEE (2019)
3. Yu, X., et al.: DS-NLCsiNet: exploiting non-local neural networks for massive MIMO CSI feedback. *IEEE Commun. Lett.* **24**(12), 2790–2794 (2020)
4. Lu, Z., Wang, J., Song, J.: Multi-resolution CSI feedback with deep learning in massive MIMO system. In: ICC 2020-2020 IEEE International Conference on Communications (ICC). IEEE (2020)
5. Ji, S., Li, M.: CLNet: complex input lightweight neural network designed for massive MIMO CSI feedback. *IEEE Wirel. Commun. Lett.* **10**(10), 2318–2322 (2021)
6. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
7. Woo, S., et al.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
8. Lu, Z., Wang, J., Song, J.: Binary neural network aided CSI feedback in massive MIMO system. *IEEE Wirel. Commun. Lett.* **10**(6), 1305–1308 (2021)
9. Guo, J., et al.: DL-based CSI feedback and cooperative recovery in massive MIMO. arXiv preprint [arXiv:2003.03303](https://arxiv.org/abs/2003.03303) (2020)
10. Chen, T., et al.: Deep learning for joint channel estimation and feedback in massive MIMO systems. arXiv preprint [arXiv:2011.07242](https://arxiv.org/abs/2011.07242) (2020)



Facemask Wearing Correctness Detection Using Deep Learning Approaches

Atlanta Choudhury and Kandarpa Kumar Sarma (✉)

Department of Electronics and Communication Engineering, Gauhati University,
Guwahati 781014, Assam, India
{atlantachoudhury07, kandarpaks}@gauhati.ac.in

Abstract. The occurrence of widespread fatality, inadequate medical infrastructure and the infectious nature of the COVID-19 virus have necessitated the formulation of appropriate risk minimization methods including extensive use of technology. In the absence of effective antiviral and limited medical resources, among the measures recommended by the World Health Organization (WHO), wearing a mask is considered to be an effective non-pharmaceutical intervention that can be used to prevent the spread of the COVID-19 virus. Hence proper wearing of the mask and its effective monitoring in public places for reliable enforcing of the community level protocol, adoption of technology becomes crucial. To contribute towards communal health, this paper aims to report the design of an accurate and real-time technique that can efficiently detect non-mask faces in public and thus, suggest ways to formulate measures for enforcing proper wearing of the mask. Among many such technologies, internet of things (IoT) and artificial intelligence (AI) have emerged as viable ones as these combine a host of sensor packs, wireless communication based networking and automated decision making. Further emerging applications involving edge computing, deep learning (DL) and Deep Transfer Learning (DTL) have enabled IoT to take part in a decisive role in the health care sector and help to minimizing damages related to pandemic situations. The presented framework is based on Artificial Neural Network (ANN) tools that use hand-crafted feature samples and DL techniques like Convolutional Neural Network (CNN) and a specialized CNN called YOLO and Support Vector Machine (SVM) classifiers. Here MobileNet has been used as a baseline method which is extended by applying the concept of transfer learning to fuse high-level semantic information in multiple feature maps with samples from Real World Masked Face Dataset. In addition, we also propose a bounding box transformation to improve localization performance during mask detection. It is observed that the proposed technique achieves high accuracy (97.2%) when implemented with MobileNet.

Keywords: COVID-19 · Deep Learning (DL) · Deep Transfer Learning (DTL) · Social Distancing · face mask

1 Introduction

The spread of COVID-19 has resulted in fatalities all over the world [1]. The spread of virus can be avoided by mitigating the effect of the virus in the environment [2, 3] or preventing the virus transfer from person to person by practicing physical distance and

wearing face masks. The World Health Organization (WHO) defined physical distancing as keeping at least six feet or two meters distance from others and recommended that keeping the physical distance and wearing a face mask can significantly reduce transmission of the COVID-19 virus [4–7]. Another challenge is related to the use of technology in detecting people with or without mask so as to prevent the transmission of SARS-CoV-2 between humans. Therefore, the regulatory use of face masks and its monitoring can slow down the high spread of this virus. It provides the opportunity for extensive use of technology. In the absence of effective antiviral and limited medical resources, many measures are recommended by WHO to control the infection rate and avoid exhausting the limited medical resources. Wearing a mask is among the non-pharmaceutical intervention measures that can be used to cut the primary source of SARS-CoV2 droplets expelled by an infected individual. Regardless of discourse on medical resources and diversities in masks, all countries are mandating coverings over the nose and mouth in public. To contribute towards communal health, this paper aims to report the design of an accurate and real-time technique that can efficiently detect non-mask faces in public and thus, formulating measures for enforcing wearing of mask. Internet of things (IoT) and artificial intelligence (AI) have emerged as viable ones as these combine a host of sensor packs, wireless communication based networking and automated decision making. Further emerging applications involving edge computing, deep learning (DL) and Deep Transfer Learning (DTL) have enabled IoT to take part in a decisive role in the health care sector and help to minimizing damages related to pandemic situations. The proposed work is based on Artificial Neural Network (ANN) tools using special feature samples and DL techniques like Convolutional Neural Network (CNN). Further, it uses a specialized CNN called YOLO and Support Vector Machine (SVM) classifiers for ascertaining the robustness of the approach. The popular MobileNet trained using transfer learning serves as a baseline method. It combines high-level semantic information in multiple feature maps with samples from Real World Masked Face Dataset. In addition, we also propose a bounding box transformation to improve localization performance during mask detection. The experiments are conducted with three popular baseline models of MobileNet. It is observed that the proposed technique achieves high accuracy (97.2%) when implemented with MobileNet.

2 Block Diagram of Proposed Model

The proposed model of face mask detection is shown in Fig.1 It is constituted by face detection and region of interest (RoI) blocks together with trained classifiers that are able to discriminate between acceptable and unacceptable ways of wearing masks in public. Here, 2079 images are used to ascertain whether a person is wearing mask or not or incorrectly wearing. First, we trained the classification models i.e. ResNet-50 and MobileNet and then apply to each training sets.

We obtained a face mask dataset openly available and increased the size of dataset by applying data augmentation. This has been used to train the classifiers which are configured to perform different object detection tasks and to choose the most accurate model for face mask detection. For facemask detection, we used ResNet-50 and MobileNet. For the physical distance detection, we used a Faster R-CNN model to detect people and

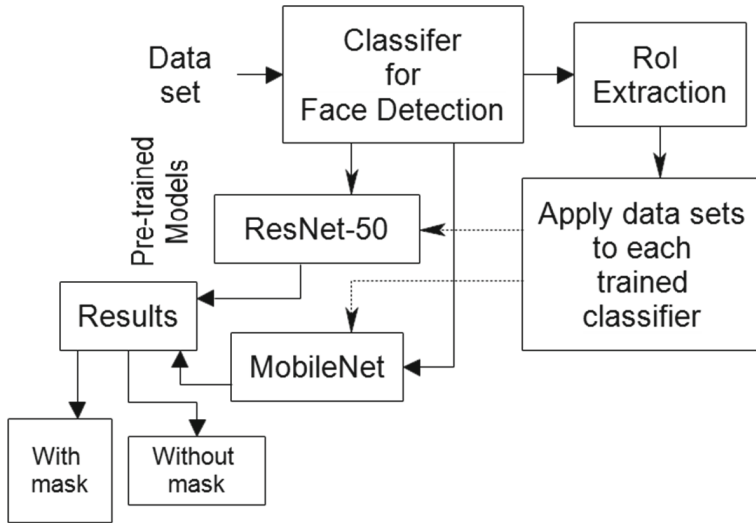


Fig. 1. Proposed block diagram of facemask detection

then used the Euclidian distance to measure the actual distance between people using the pixel numbers in the image. Subsequently, we utilized transfer learning to increase the accuracy.

2.1 About the Facemask Datasets

The Kaggle datasets are used here, which is a masked face dataset devoted mainly to improve the recognition performance of the existing face recognition technology that became popular during the COVID-19 pandemic period. The datasets contain three different variations, real-world masked face recognition dataset, simulated masked face recognition datasets, and real-world masked face verification.

A part of the dataset of face masks has been obtained from MakeML website that contains 2079 images that each image includes one or multiple normal faces with various illuminations and poses. We used 80% images (1664) for training the system and 20% (415) images are used for testing. The images are already annotated with faces with a mask, without mask and incorrect mask. The total of 2079 images were used as the face mask dataset (with mask: 690, without masks: 686, incorrect masks: 703). To avoid class imbalance and to increase the size of training dataset, we applied a combination of rotation, flipping, contrast, and brightness data augmentation techniques. This helps to avoid possible over-fitting and making the model robust in the detection of new or unseen data.

2.2 Features Identification and Extraction from Optical Facemask Image Transfer Learning

The size of the training dataset for the face mask detection is limited. Since the face mask detection model is a complex network, the use of limited data is likely to result in

an overfitted model. Transfer learning (TL) is a common technique in machine learning (ML) adopted in situations where the dataset is limited and the training process is computationally expensive. Transfer learning uses the weights of a pre-trained model on a big dataset in a similar network as the starting point in training. Here, the weights of pre-trained object detection models are used for the model training configured for the present application.

2.3 Model Selection and Implementation

It is proposed to design a system that is capable of identifying a person’s face, even if it is with or without a mask. For the system to work properly, it is necessary to use two databases: the first is for classifier training and consists of a large number of images of people who wear a face mask and others who do not. The second method is used for training the facial recognition system, and here there are people with and without the bio safety material (face mask). The input data are obtained either from an image or a video and the architecture used is MobileNet, with the aim of having a better precision and robustness.

3 Experimental Results

The performance of the proposed approaches are compared using the quantitative analysis for the evaluation. The model accuracy and model loss graphs are shown in Figs. 2, 3, 4 and 5 for ResNet-50 and MobileNet learning models respectively. Table 1 shows the results of different parameters calculated for ResNet-50 network. Table 2 shows the various parameters calculated for MobileNet. It is observed that the accuracy of these learning techniques are 92% and 97.2% respectively.

$$\text{Accuracy} = \frac{\text{No. of correctly detected}}{\text{Total no of samples}}$$

Table 1. Performance Evaluation of ResNet-50 for detecting incorrect wearing of Facemask

Sl.no	Model	Class	Precision	Recall	F1 Score	Support	Confusion Matrix	Accuracy
1	ResNet-50	0	0.99	0.80	0.88	147	[117 23 7]	92%
		1	0.85	0.97	0.90	137	[1 133 3]	92%
		2	0.93	0.99	0.96	131	[0 1 130]	92%

Using the above approach, with trained classifiers, we can detect the people with and without a mask and incorrectly wearing a mask. These two models (ResNet-50 and MobileNet) are trained and tested with images representing actual situations. During the training, it is noticed that there is an over adjustment which is due to the fact that the database is built for this stage with a few participants, although it is composed of several

Table 2. Performance Evaluation of MobileNet for detecting incorrect wearing of Facemask

Sl.no	Model	Class	Precision	Recall	F1 Score	Support	Confusion Matrix	Accuracy
1	MobileNet	0	0.92	0.87	0.91	147	[131 9 7]	97.2%
		1	0.90	0.97	0.94	137	[3 133 1]	97.2%
		2	0.94	0.90	0.92	131	[8 5 118]	97.2%

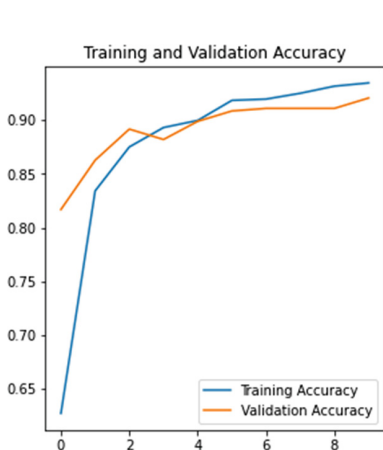


Fig. 2. Training and validation accuracy of MobileNet



Fig. 3. Training loss in MobileNet

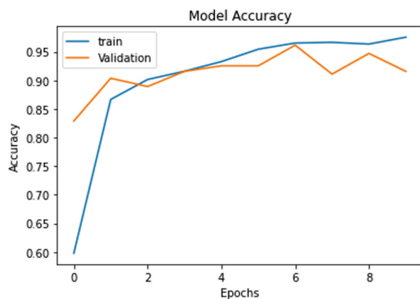


Fig. 4. Training and validation accuracy of ResNet-50

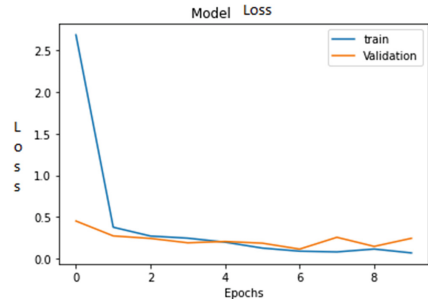


Fig. 5. Training and validation loss in ResNet-50

images, and does not exist much variability. However, the system shows potential to be used in differentiated facial recognition applications. It should be noted that if a face is not found in the database, it will be detected, but the tag “mask” or “no mask” or

“improperly wear” will be added, which refers to whether the person is wearing a mask. When subjected to testing, the average accuracy performance in identifying whether people are wearing a mask or not or incorrectly wearing obtained are 92% (ResNet-50) and 97.2% (MobileNet) (Fig. 6).



Fig. 6. Datasets of incorrect facemask

4 Conclusion

Here we have reported the design of two classifier (ResNet-50 and MobileNet) based face mask recognition system. The two systems are extensive trained and tested to work in a robust manner so that actual situations are handled efficiently. While defining whether people are wearing a mask or not or incorrectly wearing, the accuracies are 92% (ResNet-50) and 97.2% (MobileNet). One important thing to note is that the system has difficulty in the face recognition stage (with or without or not wearing mask) when the detected face presents a certain angle of inclination. This is an important aspect that needs proper addressing.


References

1. Johns Hopkins University: COVID-19 Map. Johns Hopkins Coronavirus Resource Center (2020). Accessed 30 July 2020
2. WHO: Water, sanitation, hygiene and waste management for COVID-19: technical brief. World Health Organization (2020)
3. Jahromi, R., Mogharab, V., Jahromi, H., Avazpour, A.: Synergistic effects of anionic surfactants on corona virus (SARS-CoV-2) veridical efficiency of sanitizing fluids to fight COVID-19. *Food Chem. Toxicol.* **145**, 111702, 1–4 (2020)
4. Ellis, R.: WHO Changes Stance, Says Public Should Wear Masks. <https://www.webmd.com/lung/news/20200608/who-changes-stance-says-public-should-wear-masks>. Accessed 31 July 2020
5. Feng, S., Shen, C., Xia, N., Song, W., Fan, M., Cowling, B.J.: Rational use of face masks in the COVID-19 pandemic. *Lancet Respiratory Med.* **8**(5), 434–436 (2020). Lancet Publishing Group

6. WHO: Advice on the use of masks in the context of COVID-19, 31 July 2020. <https://www.who.int/publications>
7. WHO: COVID-19 advice - Know the facts, WHO Western Pacific (2020)



An Approach to the Implementation of Nonparametric Algorithms for Controlling Multidimensional Processes in a Production Problem

Vyacheslav Zolotarev¹, Maria Lapina²(✉) , and Darya Liksonova¹

¹ Siberian State University of Science and Technology named after Academician M.F. Reshetnev, Krasnoyarsk, Russian Federation

² North Caucasus Federal University, Stavropol, Russian Federation
mlapina@ncfu.ru

Abstract. In paper, consider nonparametric algorithms for the identification and control of multidimensional discrete-continuous processes that are typical for many practical problems. The main feature of the considered multidimensional systems is the presence in them of stochastic dependences of both input and output variables. Under such conditions, the mathematical description of such objects leads to a system of implicit equations. Nonparametric algorithms for modeling and control for multidimensional systems are proposed, as well as the subsequent implementation of software-based on these algorithms in production.

Attention should be paid to the fact that the modeling of such processes is reduced to the search for the predicted values of the output variables for the known input. Moreover, for implicit equations, it is only known that one or another input variable can depend on other input and output variables that determine the state of a multidimensional system. The subsequent control of multidimensional systems is considered in conditions of insufficient a priori information about the object of research. The main component in the control problem is the determination of the reference influences, for the search of which a nonparametric algorithm is used. The essence of this algorithm is to find a common area of intersection of the values of the output variables, which will satisfy all components of the output vector. When implementing software, it is necessary to take into account information security issues, especially for three key tasks: process control, archiving (saving history) of technological parameters and working with alarms. This article provides some guidelines for implementing safeguards for archiving and alarms.

Keywords: multidimensional system · identification · control · a priori information · Information security · data protection

1 Introduction

At present, problems of identification and control of multidimensional processes of a discrete-continuous nature are of great interest. Discrete-continuous processes include processes that occur continuously in time, but the control of the output variables is carried

out at discrete moments in time. There are a lot of such processes in practice, for example, the process of oxygen-converter steel melting, where there are connections between the input and output variables since the material is loaded following certain proportions and the range of values of the output variables must correspond to the technical regulations.

A distinctive feature of multidimensional objects of a discrete-continuous nature is the presence of a stochastic dependence of the components of the vector of output variables through various channels. In this case, the mathematical description of a multidimensional object is reduced to a system of implicit stochastic equations, the parametric form of which is unknown. Therefore, it is not possible to solve such an equation using the existing methods of parametric identification.

The use of parametric methods and models requires sufficiently complete information about the object of research and can lead to a lower quality result in the case of inadequacy of the parametric description [1].

One of the areas of research for such a system is the use of nonparametric identification methods, as well as the prospective use of system analysis methods.

The management of a multidimensional object occurs under conditions of incomplete information about the object of research; therefore, the use of well-known methods will not lead to the desired result [2]. Moreover, it should be borne in mind that the system under study is multidimensional and contains unknown dependencies of the output variables. Therefore, to begin with, it is necessary to determine the setting influences, since it is impossible to choose them arbitrarily due to the unknown dependence of the output variables, and then look for control actions. To determine the reference influences, a nonparametric algorithm is used, the essence of which is to find the common area of intersection of the values of the output variables, which will satisfy all the components of the output vector.

Thus, subsequently developed software based on nonparametric algorithms is a promising direction when used in industries where there are multidimensional discrete-continuous processes. This entails the necessary accounting for information security when implementing such a system.

2 Modeling Multidimensional Systems

Consider a multidimensional object that implements the multidimensional process in Fig. 1.

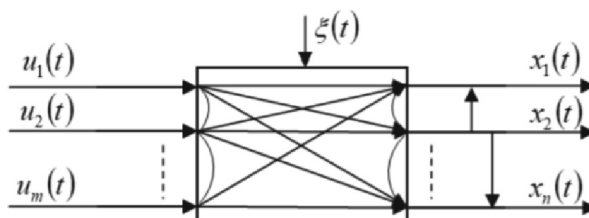


Fig. 1. Multidimensional process

In Fig. 1 the following designations are adopted: $u_1(t), u_2(t), \dots, u_k(t), \dots, u_m(t)$, $k = \overline{1, m}$, – vector components $u(t)$ input variables; $x_1(t), x_2(t), \dots, x_j(t), \dots, x_n(t)$, $j = \overline{1, n}$, – components of the vector of output variables; $\xi(t)$ – random interference acting on the object; vertical arrows indicate a stochastic dependence of the output variables; arrows inside the object show the dependence of the input variables on the output; (t) – continuous time. The measurement of all process variables is noisy.

Next, we will consider a diagram of a multidimensional process in the implementation of control and subsequent implementation in a production task. When introducing into production, it is also necessary to take into account the capabilities of the software part: from using the data exchange protocol to disaster resistance [3]. The detailed recommendations will be shown below.

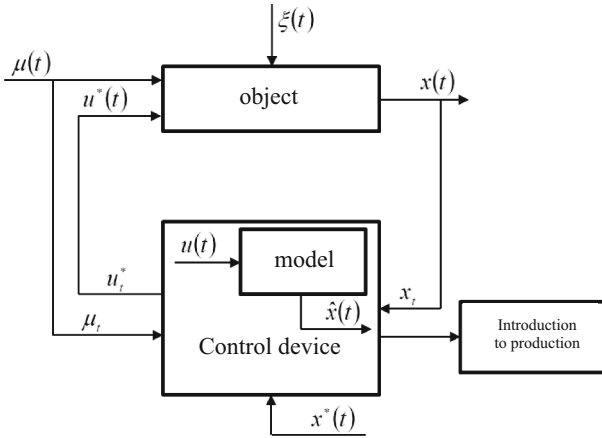


Fig. 2. Control scheme

In Fig. 2, the following designations are adopted: $u(t) = (u_1(t), u_2(t), \dots, u_m(t))$ – vector of input controlled variables of the process; $\mu(t) = (\mu_1(t), \mu_2(t), \dots, \mu_p(t))$ – vector of input uncontrolled but controlled process variables; $x(t) = (x_1(t), x_2(t), \dots, x_n(t))$ – vector of output variables of the process; $u^*(t) = (u_1^*(t), \dots, u_m^*(t))$ – vector of found input controlled process variables; $x^*(t) = (x_1(t), \dots, x_n^*(t)) \in \Omega(x^*) \subset R^n$ – vector of setting influences, $\xi(t)$ – random interference acting on the object, (t) – continuous time, u_i^* , μ_i , x_i – measurements of input and output variables at a discrete moment in time t .

We represent the mathematical description of a multidimensional object in the form of a system of implicit functions of the following form:

$$F_j(u(t), \mu(t), x(t), \xi(t)) = 0, \quad j = \overline{1, n} \tag{1}$$

which has the only solution in the field of technological process, $F_j(\cdot)$ – are not known.

For different channels of the investigated multidimensional object, the dependence of the j -th component of the output vector $x(t)$ can be represented in the form of some

dependence on certain components of the input vector $u(t)$:

$$x^{<j>} = f_j(u^{<k>}, \mu^{<p>}), \quad j = \overline{1, n}$$

This relationship is called a composite vector. A composite vector is a vector composed of the components of the vectors of input and output variables [4]. Composite vectors are written out by the researcher based on the available a priori information. In this case, the system of Eqs. (1) will take the form:

$$F_j(u^{<k>}(t), \mu^{<p>}(t), x^{<j>}(t), \xi(t)) = 0, \quad j = \overline{1, n}, \quad (2)$$

where $u^{<k>}(t)$, $\mu^{<p>}(t)$, $x^{<j>}(t)$ – compound vectors. Note that the form of Eqs. (2) continues to remain unknown and cannot be interpreted as a model of the process under study. The task is to simulate such processes [5, 6].

The system of models of the process under study can be represented in the following form:

$$\hat{F}_j(u^{<k>}(t), \mu^{<k>}(t), x^{<j>}(t), u_s, \mu_s, x_s) = 0, \quad j = \overline{1, n} \quad (3)$$

where u_s, μ_s, x_s – time vectors (data set arriving at the s -th moment), but in this case, the functions $F_j(\cdot)$ remain unknown. In the theory of identification, such problems are not only not considered, but also not posed. Usually, they follow the path of choosing a parametric structure (2), but, unfortunately, overcoming this stage is difficult due to the lack of a priori information.

Under such conditions, the estimate of the components of the vector of output variables $x(t)$ for known values of the input $u(t)$ leads to the need to solve the system of Eqs. (3). Therefore, the problem is reduced to the fact that for the given values of the input variables $u(t)$ it is necessary to solve system (3) concerning, the output variables $x(t)$ [7]. The solution scheme is reduced to the following computational procedure. First, the residuals are calculated using the formula:

$$\varepsilon_{ij} = \varepsilon_j(u^{<j>}, x^{<j>}(i), x_s, u_s), \quad j = \overline{1, n}, \quad (4)$$

where $\varepsilon_j(\cdot)$ are calculated using a nonparametric estimate of the Nadaraya-Watson regression function [8]:

$$\varepsilon_j(i) = x_j(i) - \frac{\sum_{l=1}^s x_j(l) \prod_{k=1}^{<m>} \Phi\left(\frac{u'_k - u_k[l]}{c_{su_k}}\right)}{\sum_{l=1}^s \prod_{k=1}^{<m>} \Phi\left(\frac{u'_k - u_k[l]}{c_{su_k}}\right)}, \quad j = \overline{1, n}, \quad (5)$$

where $< m >$ is the dimension of the composite vector u_k , $< m > \leq m$, in what follows, this designation is used for other variables as well. Bell-shaped functions $\Phi(\cdot)$ and the blur parameter c_{su_k} satisfy the convergence conditions. The next step is to evaluate the conditional mathematical expectation:

$$x_j = M \left\{ x_j | u^{<j>}, \varepsilon = 0 \right\}, \quad j = \overline{1, n}, \quad (6)$$

Ultimately, the forecast for each component of the vector of output variables will look like this:

$$\hat{x}_j = \frac{\sum_{i=1}^s x_j[l] \cdot \prod_{k_1=1}^{<m>} \Phi\left(\frac{u_{k_1} - u_{k_1}[l]}{c_{su}}\right) \prod_{k_2=1}^{<n>} \Phi\left(\frac{\varepsilon_{k_2}[l]}{c_{s\varepsilon}}\right)}{\sum_{i=1}^s \prod_{k_1=1}^{<m>} \Phi\left(\frac{u_{k_1} - u_{k_1}[l]}{c_{su}}\right) \prod_{k_2=1}^{<n>} \Phi\left(\frac{\varepsilon_{k_2}[l]}{c_{s\varepsilon}}\right)}, \tag{7}$$

where $j = \overline{1, n}$, and the bell-shaped functions $\Phi(\cdot)$ can be taken, for example, in the form of a triangular core.

As a result of this procedure, the values of the output variables are obtained x with input influences on the object u , which can later be used in various control systems.

3 The Task of Managing Multidimensional Processes

When considering the control problem for a multidimensional object, it is necessary to take into account the fact that the multidimensional system has unknown stochastic dependences of the output variables; in this case, it is not possible to choose arbitrary setting influences from the range of output variables. Specifying influences should be selected from the general area of intersection of the values of the output variables.

Taking this fact into account, the nonparametric procedure for calculating the reference influences will take the following form:

$$\begin{aligned} x_j^{**s} &= \frac{\sum_{i=1}^{s_{j-1}} x_j^i \prod_{j=1}^{j-1} \left(\frac{x_j^{**} - x_j^i}{c_{x_j}}\right) \prod_{k=1}^{<m>} \Phi\left(\frac{u_k - u_k^i}{c_{u_k}}\right) *}{\sum_{i=1}^{s_{j-1}} \prod_{j=1}^{j-1} \left(\frac{x_j^{**} - x_j^i}{c_{x_j}}\right) \prod_{k=1}^{<m>} \Phi\left(\frac{u_k - u_k^i}{c_{u_k}}\right) *} \tag{8} \\ & * \prod_{v=1}^{<p>} \left(\frac{\mu_v - \mu_v^i}{c_{s\mu}}\right) \\ & * \prod_{v=1}^{<p>} \left(\frac{\mu_v - \mu_v^i}{c_{s\mu}}\right) \end{aligned}$$

In general, for a multidimensional system, the calculation of control actions will look like this:

$$\begin{aligned} u_s^k &= \frac{\sum_{i=1}^s u_i^k \prod_{k=1}^{k-1} \Phi\left(\frac{u_k - u_k^i}{c_{u_k}}\right) \prod_{j=1}^n \Phi\left(\frac{x_j^{**} - x_j^i}{c_{x_j}}\right) *}{\sum_{i=1}^s \prod_{k=1}^{k-1} \Phi\left(\frac{u_k - u_k^i}{c_{u_k}}\right) \prod_{j=1}^n \Phi\left(\frac{x_j^{**} - x_j^i}{c_{x_j}}\right) *} \tag{9} \\ & * \prod_{v=1}^p \Phi\left(\frac{\mu_v - \mu_v^i}{c_{\mu_v}}\right) \\ & * \prod_{v=1}^p \Phi\left(\frac{\mu_v - \mu_v^i}{c_{\mu_v}}\right), \quad k = \overline{1, m} \end{aligned}$$

Thus, performing these procedures (8) and (9) to find the master and control actions at the output of a multidimensional system, you can get the desired result. Moreover, it should be noted that the use of nonparametric algorithms significantly reduces the time to find the desired solution, improves the accuracy of modeling and control, and significantly reduces costs [1].

4 Implementation

We will separately consider the issues of information security. Now, this is a question included in many industry standards [9]. As part of the task, as mentioned above, it is necessary to solve the issue of data protection when archiving technological parameters, controlling a technological process, and working with alarms. If the process control seems to be a complex process, which is advisable to consider separately, for example, from the standpoint of embedded encryption of programmable logic controllers [10], then some basic recommendations can be considered for alarms and archiving, including modern trends (Fig. 3).

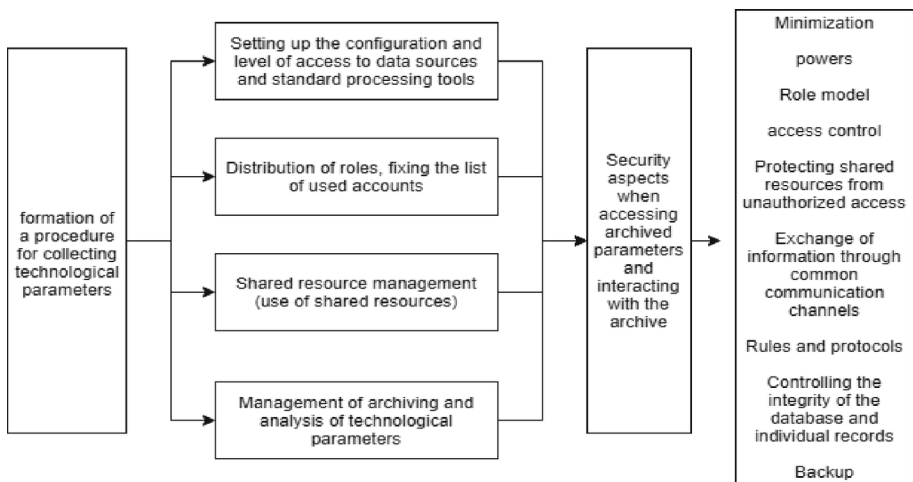


Fig. 3. Basic information security scheme

At the same time, it can be noted that archiving of technological parameters has its main purpose to provide the possibility of retrospective analysis. Thus, the key parameters are the integrity and access control to the archive, including for standard tools.

The alarm signaling has the following features: fixing the fact of a specific operator's action, timing, using data to investigate incidents. It is also stored in the archive of technological parameters, but for various reasons, it can be stored on an external server (protection against deliberate distortion by the operator, for example).

The final recommendations can be presented in the form of a diagram (Fig. 4).

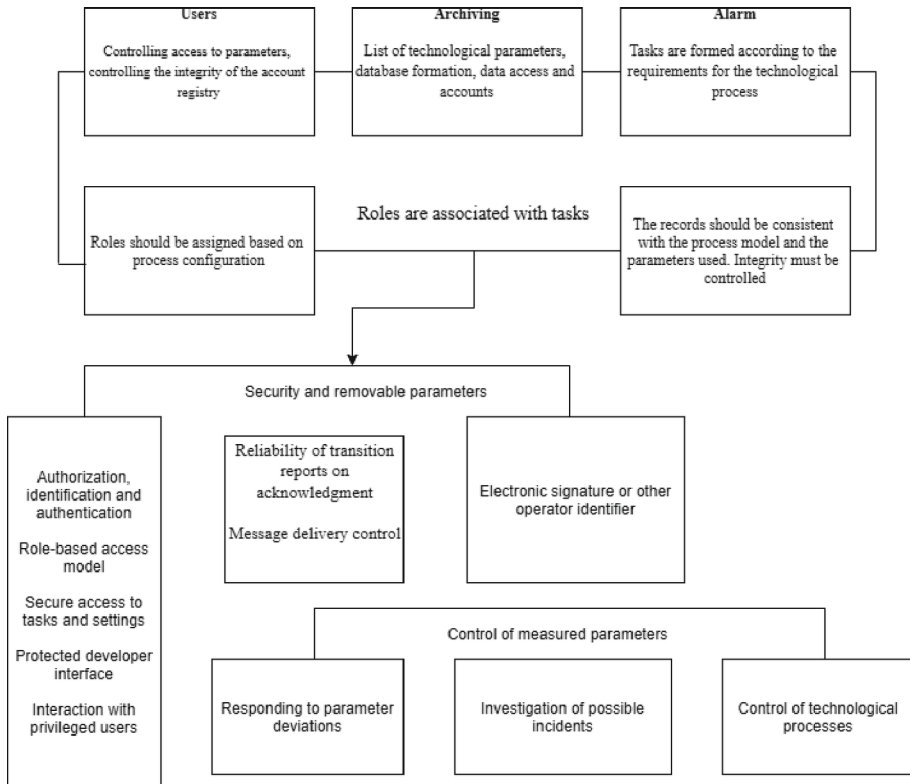


Fig. 4. Detailed recommendations for the implementation of information protection measures

The result of this implementation will be the following sequence of actions:

1. Evaluation of the measured parameters, their format, and reading method to analyze the possibility of applying protective measures.
2. Analysis of data exchange protocols to obtain information about the transmission method, format of data and headers, service information, intermediate communication devices.
3. Analysis of the data storage and processing management system. In most cases, we will talk either about working with files, including large files or about a database management system. Accordingly, information protection measures will focus either on the security of regular data processing facilities or on the protection of accounts associated with their processing. End-to-end encryption is also acceptable, including “lightweight” crypto algorithms.
4. Analysis of the order of access to data. This can be significant for both process control and data extraction and processing. What matters is the operating environment used,

as well as the capabilities of standard means of access, such as backup and protection of backup information, identification, and authentication, data encryption.

Thus, it can be noted that mathematical models of technological process control, implemented in real technological operations, must be provided with both algorithmic support and adequate information protection measures.

5 Conclusion

In the presented work, nonparametric algorithms for the identification and control of multidimensional processes that are typical for many practical problems were considered.

Separately, recommendations are given on the algorithmic and software support of the control process, as well as information protection measures implemented for the tasks of archiving technological parameters and alarms.

References

1. Vasiliev, V.A., Dobrovidov, A.V., Koshkin, G.M.: Nonparametric Estimation of Functionals from the Distributions of Stationary Sequences, Nauka, 508 p. (2004). (in Russian)
2. Pupkova, K.A., Iagupova, N.D. (eds.): Methods of the Classical and Modern Theory of Automatic Control. In 5 volumes. Vol. 2: Statistical Dynamics and Identification of Automatic Control Systems/under, 640 p. Publishing House of MSTU named after N.E. Bauman (2004). (in Russian)
3. Klaver, M., Luijff, E.: Analyzing the Cyber Risk in Critical Infrastructures. Issues on Risk Analysis for Critical Infrastructure Protection (2021). <https://doi.org/10.5772/intechopen.94917>
4. Tsyppkin, Ya.Z.: Adaptation and Training in Automatic Systems, 400 p. Nauka, Moscow (1968). (in Russian)
5. Medvedev, A.V.: To the theory of nonparametric systems. Bull. SibGAU. **5**(31), 4–9 (2010). (in Russian)
6. Medvedev, A.V., Yareshchenko, D.I.: Nonparametric modeling of T-processes in conditions of incomplete information. Inf. Technol. **10**(25), 579–584 (2019). (in Russian)
7. Medvedev, A.V.: Foundations of the Theory of Nonparametric Systems. Identification, Management, Decision Making, 732 p. Krasnoyarsk: Siberian State University named after M.F. Reshetnev (2018). (in Russian)
8. Nadaraya, E.A.: Nonparametric Estimation of the Probability Density and the Regression Curve, 194 p. Publishing House of Tbilisi University, Tbilisi (1983). (in Russian)
9. Ozarpa, C., Aydin, M., Avci, I.: International security standards for critical oil, gas, and electricity infrastructures in smart cities: a survey study. In: Ben Ahmed, M., Rakıp Kardeş, İ., Santos, D., Sergeyeva, O., Boudhir, A.A. (eds.) SCA 2020. LNCS, vol. 183, pp. 1167–1179. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-66840-2_89
10. Alves, T., Das, R., Morris, T.: Embedding encryption and machine learning intrusion prevention systems on programmable logic controllers. IEEE Embed. Syst. Lett. **10**, 1 (2018). <https://doi.org/10.1109/LES.2018.2823906>



Analysis of Neural Networks for Image Classification

Nikolay Vershkov¹(✉), Mikhail Babenko^{1,2}, Viktor Kuchukov³,
Nataliya Kuchukova¹, and Nikolay Kucherov¹

¹ North-Caucasus Federal University, Stavropol, Russia

nvershkov@ncfu.ru

² Institute for System Programming of the Russian Academy of Sciences,
Moscow, Russia

³ North-Caucasus Center for Mathematical Research, North-Caucasus
Federal University, Stavropol, Russia

Abstract. The article explores the option of using information theory's mathematical tools to model artificial neural networks. The two primary network architectures for image recognition, classification, and clustering are the feedforward network and convolutional networks. The study investigates the use of orthogonal transformations to enhance the effectiveness of neural networks and wavelet transforms in convolutional networks. The research proposes practical applications based on the theoretical findings.

Keywords: artificial neural networks · orthogonal transformations · wavelets · feature vector · convolution · correlation

1 Introduction

The construction of artificial neural networks (ANN) is based on the organization and operation principles of their biological equivalents [1]. The research on ANN has its roots in the theory of brain functioning, which was established in 1943 by W. McCulloch and W. Pitts. Their work is widely regarded as a significant contribution to this field [2]. The theory of ANN has undergone substantial development over the past 80 years, including advances in architecture and learning methods. However, it is crucial to note that the development of ANN is primarily intuitive and algorithmic rather than mathematical. Many ANN architectures have been borrowed from the biological realm [3]. The mathematical depiction of ANN received a significant advancement through the works of Kolmogorov-Arnold [4, 5] and Hecht-Nielsen [6]. These works are regarded as notable milestones in this field.

The use of information theory in the study of ANN has been relatively uncommon. Claude Shannon's groundbreaking work in information theory [7] established the basis for measuring and optimizing information transmission through communication channels, including the role of coding redundancy in improving error detection and correction. Since ANNs are essentially systems that process

information, researchers have applied the mathematical tools of information theory to investigate self-organization models [3].

The authors of this article conducted a series of studies [8–10] to examine the information processes in feedforward ANNs. These studies have shown potential benefits, including reduced redundancy, energy consumption, and training time, when considering the information processing characteristics of neural networks. The mathematical model of a neuron's ability to perform various transformations in the ANN layers enables us to analyze the input information processing methods from an information theory standpoint. The conventional approach, known as the McCulloch-Peets model [2], regards the mathematical model of a neuron as follows:

$$y_{k,l} = f \left(\sum_{i=1}^n w_i^{k,l} x_i^{k,l} \right) \quad (1)$$

where k and l are the number of layer and neuron in the layer, respectively, $y_{k,l}$ is the output of the neuron, $x_i^{k,l}$ signifies the inputs of the neuron, $w_i^{k,l}$ symbolizes the weights (synapses) of the input signals, and f is the neuron output function, which can be linear or not. Several linear transformations in information theory possess a comparable structure, such as orthogonal transformations, convolution, correlation, filtering in the frequency domain, among others. Previous research [8–10] addressed problems such as the optimal loss function, non-linear neuron characteristics, and neural network volume optimization. The goal of this article is to examine neural networks for image processing from an information theory perspective and establish general principles for building ANNs to solve specific problems. The research is entirely theoretical, and the article does not aim to experimentally validate the authors' propositions using mathematical tools of information theory.

2 Materials and Methods

2.1 The Wave Model of Feedforward ANN

According to previous studies [8], the information model of a feedforward ANN involves a multidimensional input vector $X_i = \{x_1^i, x_2^i, \dots, x_n^i\}$, which can be discretized in time and level values of some input function $x(t)$. This input value X_i is processed by each neuron in each layer of the ANN according to Eq. (1), resulting in discrete output values $Y_i = \{y_1^i, y_2^i, \dots, y_m^i\}$. The Kotelnikov theorem, also known as the Nyquist criterion, is used to discretize the functions $x(t)$ and $y(t)$ in the information model of a feedforward ANN. It should be noted that the set $\{X_i\}_{i=1,2,\dots,n}$ is not complete, which means that some input values may not be included in the training alphabet of the ANN. This is different from the decoding process in an information channel, where the alphabet of transmitted discrete messages is finite and predefined, as described by Shenon [7]. Additionally, the weight values in all neurons of the ANN are assumed to be randomly assigned before the learning process begins. When training the ANN with a teacher, the output function $y(t)$ is completely known.

ANNs are composed of an input layer that can handle X_i , an output layer with a capacity of Y_i , and one or more hidden layers. Depending on the application, ANNs can perform different tasks like image classification, clustering, and function approximation. To better understand the function being studied, the operation of the network will be analyzed in the application domain that is most appropriate.

The output layer is a critical component in ANNs for tasks such as classification or clustering. Its purpose is to assign input signals to their corresponding classes or clusters, similar to how a received signal in communication systems is observed to determine the transmitted signal [12]. However, just like in communication systems, ANNs can also experience interference, which depends on the set of input information used for classification or clustering rather than the communication channel. To mathematically describe the ANN, a transition probability $p[x(t)|y(t)]$ is used, which represents the probability of converting a received realization into the correct class or cluster. A model using additive white Gaussian noise, similar to communication theory, can be applied to the data [13]. This model is suitable when there is a large number of data in the set, such as in the MNIST database [14], which contains 60,000 records. The transition probability decreases exponentially with the square of the Euclidean distance $d^2(x, y)$ between the obtained value of X_i and the ideal representation of class Y_i given by:

$$p[x(t)|y(t)] = k \exp\left(-\frac{1}{N_0}d^2(x, y)\right), \tag{2}$$

where k is a coefficient independent of $x(t)$ and $y(t)$, N_0 is the spectral density of noise, and

$$d^2(x, y) = \int_0^T [x(t) - y(t)]^2 dt. \tag{3}$$

In some problems involving approximation and prediction, it is assumed that the signals $x(t)$ and $y(t)$ have the same period, but in the specific problem classes, they are treated as separate. For instance, in image classification problems like those found in the MNIST database, the input vector comprises 784 pixel values, and there are ten image classes. To solve these problems, it's necessary to establish a clear mapping $Y_i \leftrightarrow \tilde{X}_i$, where an observation X_i is compared to an "ideal representation" of class \tilde{X}_i , and if they are similar, it is inferred that observation X_i belongs to class Y_i . This process is expressed mathematically in the following equation:

$$(X_i \in Y_i) = \min_j d^2(X_i, \tilde{X}_j). \tag{4}$$

Opening the parentheses in Eq. (3) and replacing the representation y with $\tilde{x}(t)$, we obtain:

$$d^2(x, \tilde{x}) = \int_0^T x(t)^2 dt - 2 \int_0^T x(t) \tilde{x}(t) dt + \int_0^T \tilde{x}(t)^2 dt = \|x\|^2 - 2z + \|\tilde{x}\|^2. \tag{5}$$

The Eq. (5) involves the energy of the input realization and the cluster representation \tilde{x} denoted by $|x|^2$ and $|\tilde{x}|^2$, respectively. These values are constants when the input signal is normalized. The term z in the equation represents the correlation between the input realization x and the cluster representation \tilde{x} and can be calculated as follows:

$$z = \int_0^T x(t) \tilde{x}(t) dt. \tag{6}$$

This quantity is often referred to as the mutual energy of the two signals. Taking Eqs. (5) and (6) into account, we can represent Eq. (4) as follows:

$$(X_i \in Y_i) = \max_j z_j^i \tag{7}$$

The correlation between input signal X_i and the j -th cluster representation \tilde{X}_j is denoted by z_j^i . To prevent signal distortion, it is necessary to normalize the cluster representations as shown in Eq. (5)). When dealing with input signals of different lengths, the process is referred to as volume packing, where the average energy $\bar{E} = \frac{1}{n} \sum_i E_i = const$ is constant. If all input signals have the same length and their endpoints are on a spherical surface, it is called spherical packing.

Let's revisit Eq. (1), which forms the foundation of all neuron operations. If the weights of the output layer, denoted by $W^{k,l}$, are randomly assigned, then the vector $W^{k,l}$ acts as a multiplying interference, causing an increase in the packing volume. However, during the learning process, the weights become meaningful values determined by Eq. (7) by computing the error function and converting it into the gradient vector $W^{k,l}$. This operation is called "matched filtering" in information theory, and as the ANN's output layer is optimized during learning, it takes on the form of a matched filter. According to information theory, the condition for achieving the maximum response from a device with an impulse response is given by [15]:

$$h(t) = kx(-t). \tag{8}$$

In order to determine the weights of a neuron for a particular class Y_i , it is necessary for them to have a Hilbert-conjugate relationship with the ideal representation of class \tilde{X}_i . This implies that if the weights are set in each neuron of the output layer based on expression (8) for each class and the function $\max_i Y_i$ is used as the output layer's function, a matched filter with a dimension of m can be obtained. However, there is an issue with this proposed solution. However, there is a certain issue with this proposed solution. The correlation integral (6) can be represented in both the time and frequency formats:

$$z_i = \int x(t) \tilde{x}_i(t - \tau) d\tau = X(j\omega) \tilde{X}_i(j\omega). \tag{9}$$

Equation (1) is not suitable for calculating the correlation function in the time domain. This is because if the signals X and \tilde{X} are decomposed into an

orthogonal basis, such as the Fourier basis, all products with non-coinciding indices are set to zero, resulting in expression (1). However, if the orthogonality condition is not met, using Eq. (1) will produce correlation values (9) that contain errors. This can lead to an increase in classification errors and results that deviate from the expected outcomes.

Equation (9) indicates that the most favorable outcome could be achieved if the inputs to the output layer are orthogonal vectors. To accomplish this, a group of orthogonal functions, denoted as $\{u_n(t)\} = \{u_1(t), u_2(t), \dots, u_n(t)\}$, is utilized. These functions fulfill the criteria (10) for each pair, and they are utilized to determine the conversion coefficients.

$$\int_0^T u_i(t) u_j(t) dt = \begin{cases} a, \forall i = j \\ 0, \forall i \neq j \end{cases} \tag{10}$$

The conversion coefficients are not difficult to determine as

$$c_j = \frac{1}{a} \int_0^T x(t) u_j(t) dt, \quad j = 1, 2, \dots, m. \tag{11}$$

The original Eq. (11) was used to transform continuous images represented by $x(t)$ to the discrete space of clusters. In the context of digital image processing, the integral in Eq. (11) was substituted with a sum.

$$c_j = \frac{1}{a} \sum_{k=0}^{n-1} x_k u_j^k. \tag{12}$$

The article by Ahmed et al. [16] provides an extensive discussion of various types of orthogonal transformations that can be used for pattern recognition. These transformations are linear and establish a one-to-one correspondence between the input vector X and the output vector of coefficients C , resulting in an n -dimensional output vector. Comparing Eqs. (12) and (1), it becomes evident that they are identical. In other words, if we substitute weights w_j^k for u_j^k , the ANN layer can represent an orthogonal transformation, and the output of the layer will have values $\{c_j\}$. By representing vector \tilde{X} as an orthogonal transformation \tilde{C}_x , we obtain expression (9) in the following form:

$$Z_i = \sum_{j=0}^{n-1} X_i \tilde{X}_i = \sum_{j=0}^{n-1} x_j \tilde{x}_j. \tag{13}$$

Therefore, using an orthogonal transformation allows for the implementation of a feedforward ANN-based pattern recognition system.

In the study of the wave model of ANN [8–10], it was noted that both the standard and wave models had similar classification errors during the learning process, but they took different amounts of time to achieve this. This is because the standard learning algorithm, which primarily relies on the gradient method and error backpropagation, modifies the weights from the last layer to the first

(error backpropagation). Consequently, the decomposition functions in the first layer are selected based on the classification errors in the last layer. The key characteristic of the gradient used in ANN training is that it determines the direction in which a function $f(x)$ increases the most.

$$\nabla f(x) = \frac{df}{dx_1}e_1 + \frac{df}{dx_2}e_2 + \dots + \frac{df}{dx_n}e_n. \quad (14)$$

The error function is represented by the vector $E = (e_1, e_2, \dots, e_n)$, and the direction in which the function $f(x)$ does not increase is indicated by the opposite of the gradient. Using this information, the algorithm calculates the correction vector for the weights of the last layer and the previous layer based on the respective errors. The algorithm selects the decomposition functions of the first hidden layer, which become complex due to the nonlinearity of the neuron transfer function. This complexity was predicted by V.I. Arnold in [5].

The above examples indicate that using orthogonal transformations in artificial neural networks can enhance information processing. Such transformations allow for operations like correlation and convolution to be performed in appropriate planes, and the multiplication of elements with non-coincident indices in different planes is automatically excluded due to the orthogonal properties. Consequently, the use of orthogonal transformations can greatly reduce the computational burden required for image processing tasks in neural networks.

2.2 Wave Model of Convolutional ANN

Classification or clustering tasks are better suited for convolutional neural networks (CNN) than feedforward neural networks. CNN were proposed by Ian Lekun in 1988 and are known for their efficiency. They consist of one or more convolutional layers that use a small-sized kernel for the convolution operation. This operation reduces the size of the image, which is particularly beneficial for color images. A 3-dimensional kernel is used in this case to produce a single image on the layer output instead of 3. Typically, the convolutional layer is the first layer in the ANN structure and may be followed by pooling (subsampling) operations. However, we will not discuss this aspect in detail here. The output of the convolution operation is a feature map that can be classified using the last layer of the feedforward ANN.

Since the convolution integral is similar to the correlation integral (9), the advantages of using orthogonal transformations discussed in the previous section also apply to the convolution operation. Therefore, using an orthogonal transformation to represent the input signal and kernel can improve the efficiency and simplicity of the convolution calculation. Consequently, it is reasonable to use a layer that performs orthogonal transformations as the first layer in a typical CNN.

Linear transformations are commonly used in signal processing for information theory. Among them, subband encoding, which is a linear transformation, has several advantageous properties that are relevant to ANN theory. There are

two types of encoders based on linear transformation: transformation encoders and subband encoders [17]. The Fourier transform, which decomposes a signal into sinusoidal components, is an example of the first type, while the discrete cosine transform (DCT) and the Karhunen-Loève theorem are examples of the second type. These transformations are computed by convolving a finite-length signal with a set of basis functions, resulting in a set of coefficients that can be further processed. Most of these transformations are applied to non-overlapping signal blocks, and efficient computational algorithms have been developed for many of them [17].

Subband encoding applies several bandpass filters to the signal and then thins out the result by decimation. Each resulting signal carries information about a specific spectral component of the original signal on a particular spatial or temporal scale. There are several crucial properties to consider when encoding images using this method [17], including:

- scale and orientation;
- spatial localization;
- orthogonality;
- fast calculation algorithms.

In subband coding, orthogonality is not usually emphasized in communication theory. Instead, orthogonal transformations are used to decorrelate signal samples. While Fourier bases have good frequency localization, they lack spatial localization, which is not a problem when encoding a signal described by a Gaussian process. However, certain image features cannot be accurately represented by this model and require bases that are spatially localized. Filter blocks that are local and in space provide better decorrelation on average. The correlation between pixels decreases exponentially with distance, as shown by the equation:

$$R_l = e^{-\omega_0|\delta|}, \quad (15)$$

where δ is the distance variable. The corresponding spectral power density is

$$\Phi_l(\omega) = \frac{2\omega_0}{\omega_0^2 + (2\pi\omega)^2}. \quad (16)$$

To obtain smooth segments of the spectrum, it is necessary to accurately divide the spectrum at lower frequencies and approximately divide it at higher frequencies, as revealed by the Eq. (16). This process will generate subbands that exhibit white noise characteristics, with the variance directly proportional to the power spectrum within that range.

The Fourier transform is known to have a drawback in that it necessitates all of the time-related data of a signal in order to produce a single conversion coefficient. This leads to the time peak of the signal spreading throughout the frequency domain of the Fourier transform. To address this issue, the windowed Fourier transform is frequently utilized.

$$\Phi_x(\omega, b) = \int x(t) e^{-j\omega t} w(t-b) dt. \quad (17)$$

In this particular case, the transformation characterization involves a time window of the form $w(t - b)$. As a result, the transformation becomes time-dependent, generating a time-frequency matrix of the signal as described in [18]. By selecting the Gaussian function as the window, the inverse transformation can also be conducted using the same function.

The fixed size of the window in Eq. (17) is a major drawback, as it cannot be adapted to suit the features of the image. A wavelet transform can be used instead of the Fourier transform to overcome this limitation. The wavelet transform has the form:

$$\psi_{a,b}(t) = a^{-\frac{1}{2}} \psi\left(\frac{t-b}{a}\right). \quad (18)$$

It is evident that the basic wavelet functions are real and located at different positions in proximity to the x-axis. These wavelets are defined for a brief time interval, which is shorter than the signal period. The fundamental functions can be seen as rescaled and time-shifted versions of one another, according to Eq. (18), where b and a denote the time position and scaling factor, respectively. The direct wavelet transform can be mathematically formulated as:

$$\Phi_x(a, b) = a^{-\frac{1}{2}} \int x(t) \psi\left(\frac{t-b}{a}\right) dt. \quad (19)$$

The convolutional layer of a CNN is responsible for computing the convolution of the input signal block X with a core J of size $s \times s$, i.e.

$$C_{i,j} = \sum_{k=0}^{s-1} \sum_{l=0}^{s-1} X_{i+k,j+l} J_{k,l}. \quad (20)$$

Through the conversion of Eq. (19) for discretized signals and functions and comparing it with (20), the fundamental wavelet transform function can be depicted as the essential component of a convolutional layer. This implies that utilizing multiple fundamental functions is equivalent to applying several filters with distinct kernel sizes. Consequently, it is feasible to choose adaptable parameters for the window that accommodate the signal, enabling greater flexibility in the convolutional layer of the CNN.

The use of wavelet transforms in ANNs is not a novel concept, as it has been investigated in prior research [20]. Nonetheless, a more recent approach entails using the wavelet transform as the foundation of the convolutional layer in the initial layer of a feedforward CNN, as presented in [21]. This method is more attractive since the convolutional layer can function with several kernels simultaneously, making it possible to obtain multiple approximations within a single layer.

In communication theory, a signal can be expressed as a series of successive approximations, which can be advantageous for signal analysis. For instance, in image transmission, an initial rough version of an image can be transmitted and subsequently refined in sequence, facilitating rapid viewing of numerous images

from a database. A similar method can be employed for image recognition. If an image cannot be classified into a specific category based on the coarsest approximation, there is no need to compare it in a more precise approximation. This technique is referred to as multiscale analysis.

Multiscale analysis involves describing the space $L^2(R)$ using hierarchical nested subspaces V_m , that do not overlap, and their union results in the limit $L^2(R)$, i.e. $\dots \cup V_2 \cup V_1 \cup V_0 \cup V_{-1} V_{-2} \cup \dots$, $\bigcap_{m \in \mathbb{Z}} V_m = \{0\}$, $\bigcup_{m \in \mathbb{Z}} V_m = L^2(R)$. These subspaces have the property that any function $f(x)$ belonging to V_m will have a compressed version that belongs to V_{m-1} , i.e. $f(x) \in V_m \Leftrightarrow f(2x) \in V_{m-1}$. Additionally, there exists a function $\varphi(x) \in V_0$, whose shifted versions $\varphi_{0,m}(x) = \varphi(x - m)$ form an orthonormalized basis of space V_0 . The functions $\varphi_{n,m}(x) = 2^{-\frac{m}{2}} \varphi(2^{-m}x - n)$ form an orthonormal basis of space V_m . These basis functions are called scaling functions as they create scaled versions of functions in $L^2(R)$ [17]. Thus, a function $f(x)$ in $L^2(R)$ can be represented by its set of successive approximations $f_m(x)$ in V_m .

Therefore, it is possible to perform image analysis at various resolution or scale levels by selecting the value of m , which is known as the scale factor or level of analysis. A higher value of m results in a coarser approximation of the image, lacking in details, but allowing for identification of broader generalizations. Decreasing the scaling coefficient enables identification of finer details. In essence, $f_m(x)$ is an orthogonal projection of $f(x)$ onto V_m [17], i.e.

$$f_m(x) = \sum_n \langle \varphi_{m,n}(x), f(x) \rangle \varphi_{m,n}(x) = \sum_n c_{m,n} \varphi_{m,n}(x). \tag{21}$$

Without delving into the specifics of wavelet analysis at present, it is worth mentioning that any function $f(x)$ within the space $L^2(R)$ can be expressed as a combination of orthogonal projections. When analyzing the function up to a specific scale factor m , the function $f(x)$ can be represented as the addition of its crude approximation and various details. The Haar wavelet family, for example, offers such functionalities [18].

When employing subband transforms, the potential for constructing filter banks must be taken into account, which involve filtering followed by down-sampling [17, 19]. In a two-band filter bank, the low-frequency component provides a crude estimation of the signal without capturing intricate details, while the high-frequency component contains finer details. Depending on the particular processing objective, an ANN can utilize the low-frequency approximation to emphasize broad and smooth features, or the high-frequency component to emphasize specific details.

Utilizing wavelets as the kernel of a CNN enables the extraction and enhancement of the necessary image features. While this approach is not new in information processing and transmission theory, it is being utilized to establish an information model for CNNs. This technique not only advances our comprehension of the process of feature map generation but also simplifies the development of a lifting scheme for information processing in a multi-layer CNN.

3 Results

Using orthogonal transformations can be advantageous when working with images, irrespective of the ANN architecture employed. For instance, in feedforward ANNs, the use of orthogonal transformations can improve the efficiency of the final layer where image classification or clustering is performed. Orthogonalizing the data can enhance the accuracy of computing the correlation integral for the classified signal and ideal class representation.

Convolutional neural networks (CNNs) employ feedforward networks in their last layer, similar to traditional feedforward ANNs, which is essential for feature map classification. To enhance the efficiency of the last layer in CNNs, orthogonal transformations are utilized, as in feedforward ANNs. However, when analyzing image details, the Fourier transform (or similar ones) does not offer significant benefits. Therefore, wavelet transforms are more promising as they have localization in both frequency and time, unlike the window Fourier transform. Wavelets can also function as orthogonal transformations and enable the creation of filter banks for general and detailed image analysis based on specific criteria. This approach not only allows for general image classification, as in the case of the MNIST database, but also enables complex image classification based on specific details.

To confirm the effectiveness of the approach described above, experimental validation is necessary. The next step is to explore the wavelet transforms currently available for CNNs and their implementation in convolutional layers. It is essential to ensure that the feature maps are sufficiently detailed to enable efficient processing in subsequent layers.

Acknowledgments. This work was carried out at the North Caucasus Center for Mathematical Research within agreement no. 075-02-2022-892 with the Ministry of Science and Higher Education of the Russian Federation. The study was financially supported by the Russian Foundation for Basic Research within the framework of the scientific project No. 20-37-51004 “Effective intelligent data management system in edge, fog and cloud computing with adjustable fault tolerance and security” and Russian Federation President Grant SP-3186.2022.5.

References

1. Kruglov, V.V., Borisov, V.V.: Artificial neural networks. Theory and practice (Iskusstvennye nejronnye seti. Teoriya i praktika). — M.: Goryachaya liniya — Telekom (2002). (in Russian)
2. McCulloch, W., Pitts, W.: A logical calculus of ideas immament to nervous activity (Logicheskoe ischislenie idej, odnosyashchihsya k nervnoj aktivnosti). — Avtomaty. — M.: Izd. inostr. lit. (1956). (in Russian)
3. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice-Hall (1999)

4. Kolmogorov, A.N.: On the Representation of Continuous Functions of Several Variables as Superpositions of Continuous Functions of One Variable and Addition (O predstavlenii nepreryvnykh funktsiy neskol'kih peremennykh v vide superpozitsiy nepreryvnykh funktsiy odnogo peremennogo i slozheniya). - Dokl. AN SSSR, 1957, T. 114, vol. 5, pp. 953–956 (1957). (in Russian)
5. Arnol'd, V.I.: On the Representation of Functions of Several Variables as a Superposition of Functions of Fewer Variables (O predstavlenii funktsiy neskol'kih peremennykh v vide superpozitsii funktsiy men'shego chisla peremennykh). Mat. Prosveshchenie **3**, 41–61 (1958). (in Russian)
6. Hecht-Nielsen, R.: Neurocomputing. Addison-Wesley (1989)
7. Shannon, K.: Works on information theory and cybernetics. (Raboty po teorii informatsii i kibernetike). Izd-vo inostrannoy literatury (1963). (in Russian)
8. Vershkov, N.A., Kuchukov, V.A., Kuchukova, N.N., Babenko, M.: The wave model of artificial neural network. In: Proc. IEEE Conf. of Russian Young Researchers in Electrical and Electronic Engineering, EIConRus: Moscow, St. Petersburg 2020, pp. 542–547 (2020)
9. Vershkov N.A., Babenko M.G., Kuchukov V.A., Kuchukova N.N. Advanced supervised learning in multi-layer perceptrons to the recognition tasks based on correlation indicator. //Trudy ISP RAN/Proc. ISP RAS, vol. 33, issue 1, pp. 33–46 (2021)
10. Vershkov, N.A., Kuchukov, V.A., Kuchukova, N.N. The Theoretical Approach to the Search for a Global Extremum in the Training of Neural Networks. Trudy ISP RAN/Proc. ISP RAS, vol 31, issue 2, pp. 41–52 (2019). [https://doi.org/10.15514/ISPRAS-2019-31\(2\)-4](https://doi.org/10.15514/ISPRAS-2019-31(2)-4)
11. Kotelnikov, V.A.: Theory of Potential Noise Immunity. (Teoriya potentsial'noy pomekhoustojchivosti). - Radio i svyaz' (1956). (in Russian)
12. Harkevich, A.A.: Selected Works. Vol. 3. Information Theory. Pattern Recognition. (Izbrannyye trudy. T.3. Teoriya informatsii. Opoznanie obrazov.) – M.; Nauka (1972). (in Russian)
13. Ipatov, V.: Broadband systems and code division of signals. Principles and Applications. (Shirokopolosnyye sistemy i kodovoe razdelenie signalov. Principy i prilozheniya.) - M.: Tekhnosfera (2007). (in Russian)
14. Qiao, Yu.: THE MNIST DATABASE of handwritten digits (2007). Accessed 04.08.2021
15. Cook, C., Bernfeld, M.: Radar signals. Theory and application. (Radiolokatsionnyye signaly. Teoriya i primeneniye.) - Sovetskoe Radio, Moscow (1971). (in Russian)
16. Ahmed, N., Rao, K.R.: Orthogonal transformations in digital signal processing (Ortogonal'nyye preobrazovaniya pri obrabotke cifrovyykh signalov): Per. s angl./Pod red. I.B. Fomenko. - M.: Svyaz', (1980). (in Russian)
17. Vorob'ev, V.I., Gribunin, V.G.: Theory and practice of wavelet transforms. (Teoriya i praktika vevlet-preobrazovaniya.) - S.-Peterburg: Voennyj universitet svyazi (1999). (in Russian)
18. Sikarev, A.A., Lebedev, O.N.: Microelectronic devices for forming and processing complex signals. (Mikroelektronnyye ustrojstva formirovaniya i obrabotki slozhnykh signalov.) - M.: Izd-vo <<Radio i svyaz'>> 1983. (in Russian)
19. Haar, A.: Zur theorie der orthogonalen funktionensysteme. Georg-August-Universitat, Gottingen (1909)
20. Genchaj, R., Sel'chuk, F., Uitcher, B.: Introduction to wavelets and other filtering techniques in finance and economics. (Vvedenie v vevlety i drugie metody fil'tracii v finansah i ekonomike.) - Academic Press (2001). (in Russian)

21. Alexandridisa, A.K., Zapranisb, A.D.: Wavelet neural networks: a practical guide. *Neural Networks*, vol. 42, pp. 1–27 (2013)
22. Cui, Z., Chen, W., Chen, Y.: Multi-scale convolutional neural networks for time series classification (2016). arXiv preprint [arXiv:1603.06995](https://arxiv.org/abs/1603.06995)



Factors of a Mathematical Model for Detection an Internal Attacker of the Company

I. V. Mandritsa^(✉), V. V. Antonov, and Siyanda L. Madi

Department of Organization and Technologies Defense of Information, Institute of Digital
Development, North-Caucuses Federal University, Stavropol, Russia
imandritsa@ncfu.ru

Abstract. For commercial organizations, the issue of tolerance and loyalty of employees is always relevant, since it is closely related to the issue of leakage of trade secrets from those who have access to it. Thus, the identification of an internal violator is becoming increasingly relevant in the development of information processing technologies, when in the information systems of the organization the most vulnerable place remains a person. An employee who processes information and has full access to it, becomes the target of an attacker, and can also be an intruder and become a source of leakage of confidential information. Analysis of metadata about the current psycho-emotional state of the employee with the help of digital technologies allows you to make an assumption about his current and future emotional state, readiness to honestly perform work and predict the likelihood of this employee or violate the law on trade secrets. Thus, the security department of the company, having collected metadata and analyzing them, is able to prevent the threat and get before the leak a certain probability of malicious intent for each employee, and this will increase the security of information.

Keywords: Internal violator · information security · metadata collection · probability of malicious intent

1 Introduction

Information systems face several security threats, many of which are initiated by so called trusted employees inside of an organization. For the purposes of this paper, an insider is someone that has or had access to the information system of an organization and does not comply with the security policy of the organization. We will focus on threats that are malicious, i.e., in cases where the insider deliberately causes harm to an organization. The problem of insider threats is a problem with a technical and a behavioral nature. The paper proposes a prediction model, which combines a number of different approaches. The ultimate goal of the paper is to look into some of the factors, more specifically social learning, that influence an insider's decision to act, as well as a number of indicators and precursors of malicious acts, especially those that leave a technological, detectable trail on a system, using metadata.

Improvement of information processing technologies entails a decrease in the possibility of unauthorized receipt of confidential information by an external attacker, while

an employee who has access to confidential information can become a source of leakage of valuable information, due to his malicious intent or careless handling of the information processing system. As in the case of unintentional mistakes in the treatment of the information resources of the company, which is the result of emotional fluctuations: apathy, fatigue, bad mood, and in the case of deliberate transfer of information to third parties, hostile actions regarding the welfare of the company are affected by the emotional state of a person.

That is, to varying degrees, the negative actions of an employee of the company are influenced by various emotional deviations caused by a reaction to external factors that are identical in nature. Knowing the set of these factors and the degree of their impact, it is possible to conduct a presumptive analysis of the emotional state and, as a consequence, the possible behavior of the subject, for this purpose a model of the internal violator is compiled, reflecting the spectrum of negative emotional states of a person. Metadata for the model is taken from open sources.

Metadata—information about other information, or data related to additional information about the content or object. Metadata reveals information about the features and properties that characterize any entities, allowing you to automatically search for and manage them in large information flows [1]. Metadata creation can take place both manually and automatically. From various sources of both open and “leaks” of information from closed sources (debtors’ databases, databases on offenses on the roads, etc.), flows of data about a person, his material support, possible problems with the law are formed. From social networks there is a stream of data on the worldview, political and social views of a person. You can also use internal data of firm and video surveillance, indicating the movements of a person inside the object.

Common factors that are considered when making threat prediction models are sociodemographic factors such as age, gender, education, migration background and ethnicity, religious affiliation, marital status, household, employment, and income. For more accurate predictions we propose taking psychological factors like social learning into account as well.

Social learning theory is a theory of learning process and social behavior which proposes that new behaviors can be acquired by observing and imitating others. According to social learning theory, people engage in crime because of their association with others who engage in crime. Their criminal behavior is reinforced and they learn beliefs that are favorable to crime [2]. They essentially have criminal models that they associate with. As a consequence, these individuals come to view crime as something that is desirable, or at least justifiable in certain situations. Learning criminal or deviant behavior is the same as learning to engage in conforming behavior: it is done through association with or exposure to others.

With the use API from multiple web services and social media applications we can determine whether an employee could become a threat or not. By looking at what kind of people they follow or if any of their close friends and family members have committed any criminal offences for instance can tell us if the employee is more inclined to deviant behavior. Information compiled from this factor as well as others will analyzed and put into a mathematical model which will calculate the probability of the employee being a threat. Should the probability be higher than a certain threshold which is determined

by the head of information security then the employee is locked out of the organizations system.

At this stage of work, we propose the following concept of two scientific hypotheses – who is such an employee for the purposes of controlling access to the company’s trade secrets [3]. The first hypothesis (assumption) states that our object-worker is defined as an employee with a “public face” (in society, in a collective), as a predetermined “unstable deterministic system” of a set of indicators (factors) describing his current psycho-emotional (the first flow of metadata) and physical-economic (the second flow of metadata) states, and accordingly is interpreted by us as a “stable” or “unstable” state of his motivation to “become or not to become” an internal violator of the firm [4].

The second hypothesis (assumption), predetermines which of the two independent streams of metadata proposed according to the first hypothesis, about the confirmed (or unconfirmed) facts of its developing “crisis” in the ethics of behavior (conflicts, empathy, selfishness, etc.) and in labor indicators (productivity, quality, simple, etc.), should have a different philosophy of comprehension that is converted into different levels of ranking the probabilistic state of stability and instability of the object. The fact is that the internal environment of functioning in the enterprise “requires” from our object of research the fulfillment of a set of indicators of functioning in the company, but which reflect its “some” current in development “intermediate” state of motivation and satisfaction, which for the research model is a certain “rank” of the state [5].

2 Materials and Methods

We propose to introduce two terms for the subsequent filling of the mathematical model with factors:

- the type of employee “Alien”, whose “working” usual state corresponds to the type - “resists” the functional duties imposed on him by the information system of the enterprise - due to the “duality” of the alienity of the socio-environment of the enterprise - because it is “not a family” for him and if the “internal” mechanisms of motivational criteria for the comfort of staying in this environment exceed the “threshold” of tolerance, then they (feelings and consciousness) translate his state according to the “Alien” type “Internal violator”;
- type of employee “Own” of the enterprise, whose “working” usual state corresponds to the type - “helps” and “cooperates” with the indicators imposed on him by the information system of the enterprise, according to the composition of functional duties and the quantity and quality of their performance, the employee works optimally productively, until the moment when the “internal” mechanisms of motivational criteria for the threshold of optimality of efforts and the evaluation of these efforts do not “bring down” his type “His” in the type of “Internal violator” for the reason is the underestimation of the payment of his efforts and labor. However, these two types of “Alien” and “Own” coexist in one employee at once, but in the mode of triggering an externally internal emotional state from external factors.

The following Fig. 1 shows a “dual” gestalt, or type of behavior of any employee, as a scheme for providing a mathematical model of HV by collecting metadata about

him. If the first type of information can be generated completely automatically using well-known API-request requests for each employee of the company and data of DLP systems, which automatically create metadata - daily, then the second type of KPI - indicators of economic activity - in semi-automatic (once at the end of the week). KPI indicators will be obtained by calculation subjective means - by the immediate supervisor of the employee according to the KPI methodology (Key performance indicators) [6].

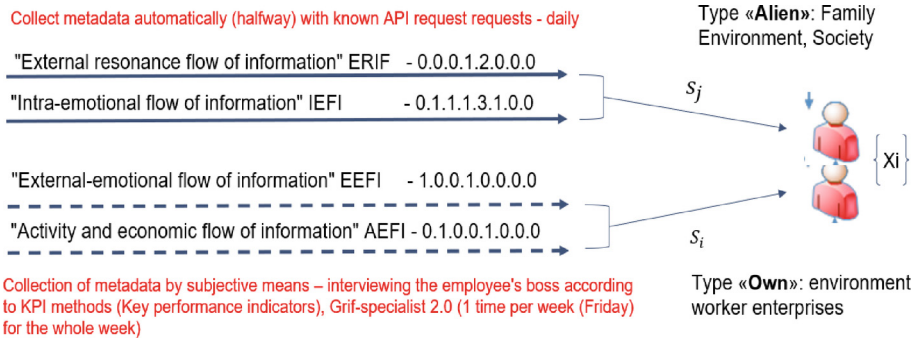


Fig. 1. Collection of metadata about the employee of the company

Using the formal logic of scientific research, it is impossible not to agree with the thesis that it is a mistake to mix emotions (psychomotor indicators) and economic (physical) indicators collected using API-request requests from internal sources of the company due to the different “nature” of these factors. Accordingly, the two streams of metadata should have different degrees (levels) of the estimated probability of its stable state, that the employee, having reached the “peak of the crisis” of his state (for various employees - he is personal), will not go into the state of “internal violator”.

Consequently, different streams of metadata about an employee must necessarily have “estimated” probabilities in different ways. One confirmed event returns a value of 0 in 1 (Fig. 1), which is important for calculating the probability, which means that there is a fact of a certain proportion (weight) of the information and psychological impact of this flow of information on the Xi-worker until the subsequent confirmation of the fact.

After all the stages of the model, at the end of the algorithm it will be necessary to calculate the probabilistic states “Own” - or “Alien” - for employee Xi from the number of Q(Xi) incidents affecting him “from the outside”, as shown earlier in Fig. 3 - decomposition of the stages of the model in the form of an algorithm for calculating the probability of the employee’s state “Alien”. $S_i S_j$

Accordingly, this figure reflects the hypothesis of calculating the total probability of the observed impact of incidents on the state of the employee - “Alien” according to the generalized Bayes formula (1) $\rho_j S_j (X_i)^{ext} X_i$

$$P\left(\frac{D_{Xi}}{K^{Sj}}\right) = P(D_{Xi}) * P\left(\frac{K^{Sj}}{D_{Xi}}\right) / P(K^{Sj}) \tag{1}$$

where: $P\left(\frac{D_{Xi}}{K^{Sj}}\right)$ – the resulting total probability of the state Sj – “Alien”;

$P(D_{Xi})$ – preliminary marginal probability (threshold hypothesis) of employee dissatisfaction X_i .

Next, you need to enter the values of the thresholds of probability that the employee is approaching or moving away from the “state” of VN. Figure 2 presents hypotheses for evaluating (ranking) the collected metadata about the employee.

Rank of cumulative probability for X_i -threats according to n-observations of its confirmation from external closed sources		
	ERIF	IEFI
Almost zero:	$f(1) \rightarrow \rho = 0,0;$	$f(1) \rightarrow \rho = 0,0$
Minor	$f(2) \rightarrow \rho = 0,1;$	$f(2) \rightarrow \rho = 0,1$
Minor +:	$f(3) \rightarrow \rho = 0,3;$	$f(3) \rightarrow \rho = 0,2$
Significant:	$f(4) \rightarrow \rho = 0,5;$	$f(4) \rightarrow \rho = 0,3$
High:	$f(5) \rightarrow \rho = 0,7.$	$f(5) \rightarrow \rho = 0,4$

Fig. 2. Ranking the probability from the number of events on the collected API-request ERIF and IEFI

3 Results and Discussion

Next, taking as an experiment an employee of the company - office manager and conditional values for the collected metadata about his states, we obtained the following results of the CT protection model. By transforming the probabilities of incidents observed from closed and open sources conditionally occurred with the employee in the external environment (type - alien) and in the internal environment of the company (type - your own), conditionally confirming the impact of the i -threat on the X_i -worker on the first data stream ERIF $\vec{\rho}(x_i^{ERIF})$ and the second data stream IEFI $\vec{\rho}(x_i^{IEFI})$, in their aggregate probability $\cup \vec{\rho}(S_j)$ - that there is a current value - “Alien”, we will compare it with the full Bayes probability according to the formula (2). Observations are conditionally carried out 1 time per day in 1 week.

$$\begin{aligned}
 \bar{p} &= \frac{1}{7} * \sum_1^7 (0, 0 + 0, 0 + 0, 0 + 0, 0 + 0, 0 + 0, 2 + 0, 2) + \frac{1}{7} \\
 &* \sum_1^7 (0, 0 + 0, 1 + 0, 0 + 0, 0 + 0, 0 + 0, 2 + 0, 0) \\
 &= 0, 057 + 0, 042 = 0, 1
 \end{aligned}
 \tag{2}$$

By comparing, we get two responses from the security trigger for this employee X_i : “Access is open” or “Access is denied” [5]. Since later in Fig. 1 for a conditional

employee of the company for example: “Office Manager” a hypothetical threshold of the dissatisfaction coefficient was set to $D = (0.6)$, and the resulting probability from two streams = 0.1, we get the next complete Bayes probability (3) and (4), which will confirm or refute the reliability of the values obtained. $\cup \bar{p}(S_j)$

$$P_A(H_1) = \frac{0,5 * 0,6}{0,5 * 0,6 + 0,5 * 0,1} = \frac{0,3}{0,35} = 0,86 \quad (3)$$

$$P_A(H_2) = \frac{0,5 * 0,1}{0,5 * 0,6 + 0,5 * 0,1} = \frac{0,05}{0,35} = 0,14 \quad (4)$$

That is, the state of “Alien” for employee Xi is reliably 86%. Therefore, the trigger of the company’s access system will be confidently defined for the employee as “Access is allowed”, which is reflected in Fig. 3.

Threshold of marginal probability $D_i =$
employee satisfaction coefficient X_i (hypothesis)

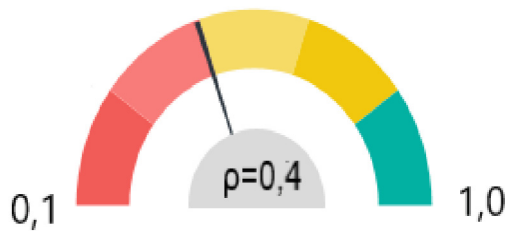


Fig. 3. Trigger of the information security system of the employee’s access to the trade secrets of the company scan

4 Conclusion

The paper considers the scale and typology of types of employee conditions as the main factors of metadata that need to be collected and taken into account at the enterprise, for all employees who have access to CT, which made it possible to propose the concept of a multifactor mathematical model for identifying an internal violator based on measuring the probability of risks of an internal threat of CT from the personnel of the enterprise, as well as an algorithm for triggering a system of access to the company’s trade secrets. The appearance of an internal violator is a multifactorial, in fact, dynamic process that has its own typology and a set of relevant threats that lead to such a phenomenon as the Internal Violator (often referred to as an insider), which, being under the influence of resonant factors, becomes a threat of leakage of CT of the enterprise. Hypotheses and conclusions of formal logic led to an understanding of the factors of protection of CT from the risk of its leakage and the likely loss of integrity or “failure” in use from an internal intruder allowed to describe the mathematical model.

References

1. Zvonarev, S.V.: Fundamentals of mathematical modeling: a textbook / S.V. Zvonarev.—Ekaterinburg: Publ. Ural. University, 2019. —112 c. – access mode
2. Mandritsa, I.V., et al.: Study of risks of business information by stages of the business process of organization in the collection: problems of information security of socio-economic systems. In: VII All-Russian Scientific and Practical Conference with International Participation. Simferopol, pp. 20–29 (2021)
3. Martyanov, E.A.: The possibility of identifying an insider by statistical methods. Syst. Means Inform. **27**(2), 41–47 (2017)
4. Medvedev, V.I., Larina, E.A.: Struggle with internal threats. Identifying an Insider - “Current Accounting”, February 2014
5. Shcheglov, A.Yu., Shcheglov, K.A.: Mathematical Models and Methods of Formal Design of Information Systems Protection Systems. Textbook, 93 p. ITMO University, St. Petersburg (2015)
6. Minaev, V.A., et al.: System-dynamic modeling of information impacts on society. Voprosy radioelektronika. (11), 35–43 (2017)



Comparative Analysis of Methods and Algorithms for Building a Digital Twin of a Smart City

Vladislav Lutsenko¹  and Mikhail Babenko^{1,2} 

¹ North Caucasus Center for Mathematical Research, North-Caucasus Federal University, Stavropol, Russia

officialvladlutsenko@gmail.com

² Sirius University of Science and Technology, Sochi, Russian Federation

Abstract. With the development of next-generation information technologies, especially big data and digital twins, the topic of building smart cities is increasingly dominating discussions about social change and economic performance. The purpose of this article is to analyze methods for building digital twins of a smart city. The paper describes the concepts underlying digital twins. Examples of the implementations of methods for building digital twins are investigated. Advantages of data mining and neural network modeling over other methods in the context of the considered characteristics are revealed. Based on the comparative analysis, it is shown that all methods can be complementary, as they are aimed to optimize processes, as well as predict and analyze problems.

Keywords: Digital twin · Smart city · Neural networks · Data mining · Big data

1 Introduction

A smart city is an innovative city that uses information and communication technology (ICT) to improve the quality of life of citizens and increase the efficiency of city services [1]. To achieve this goal, information from different urban systems is combined and analyzed to provide more efficient services. This data is generated from various sources, such as wireless sensor networks and Internet of Things (IoT) devices installed in buildings, streets, vehicles, etc. New ways of collecting and analyzing data are gradually replacing established mechanisms of city management. Unlike statistical samples, which have time to become obsolete by the time they are analyzed, big data can be processed in real time, which increases the quality and speed of decision-making. The culture of big data encompasses cyber-physical systems, cloud computing, and IoT. The current global attention to big data on the Internet is related to the development of artificial intelligence (AI) [2]. Big data in the field of urban management complements traditional types of information about the city and expands its scope of application [3]. Thus, thanks to big data it became possible to monitor behavioral patterns and analyze the urban lifestyle at the intersection of such familiar categories as population, economic development, building and infrastructure, etc.

The Ministry of Construction of Russia has developed an index of urban digitalization “IQ of cities”, which aims to evaluate the effectiveness of the urban management system and increase the competitiveness of Russian cities. The index is estimated on forty seven indicators, which are divided into ten areas. One of the indicators of the “IQ of cities” is the presence of a digital twin city.

The concept of Digital Twin (DT) refers to the development of the digital counterpart of the physical system and linking their information throughout the life cycle of the physical counterpart [4]. It is assumed that the use of digital twins will contribute to the development of a smart city and increase its sustainability [5]. However, distinguishing digital twins from other types of models is critical to understanding the applicability of using a digital twin for city-scale modeling; that is, the level of data integration, which varies by model type, between the two analogs (digital and physical) of the digital twin is critical to determining whether or not the developed model is a digital twin [6]. Thus, a clear definition of the potential benefits, as well as the problems and requirements of the development of digital twins of the city is important in the formation of smart cities.

The purpose of this article is to analyze the existing methods and technologies for building digital twins of a smart city. To do this, we need to give the concept of a digital twin city, investigate methods and algorithms for building a digital twin smart city, and then conduct a comparative analysis of them. The article has the following structure. Section 2 defines the digital twin city, describes its architecture and functional requirements. Section 3 discusses the methods and algorithms used in the construction of the digital twin smart city. Then, in Sect. 4, their comparative analysis is performed. Finally, conclusions are drawn in Sect. 5.

2 The Digital Twin of a Smart City

A digital twin is a digital (virtual) model of any object, system, process, or person, the purpose of which is to obtain a true representation of the corresponding real object [11].

Since the first mention of digital twins, the popularity of this topic has been increasing, as more and more scientists began to focus their research on digital twins [10]. The Fig. 1 shows an exponential increase in the number of publications in Scopus, as well as in ScienceDirect (in English only), which contain the term “digital twin” in the article title, abstract or as keywords from 2011 to 2020.

The main reason why DT technology is considered as mainstream in Industry 4.0 is its many benefits, including reducing errors, uncertainties, inefficiencies and costs in any system or process. Let us list some of the benefits of digital twins:

- Rapid prototyping.
- Cost-effectiveness.
- Problem prediction/system planning.
- Optimization of solutions.
- Availability.
- Safety.

A digital twin of the city is a prototype of a real city, on the basis of which one can analyze the life cycles of the object, its reaction to possible changes and external

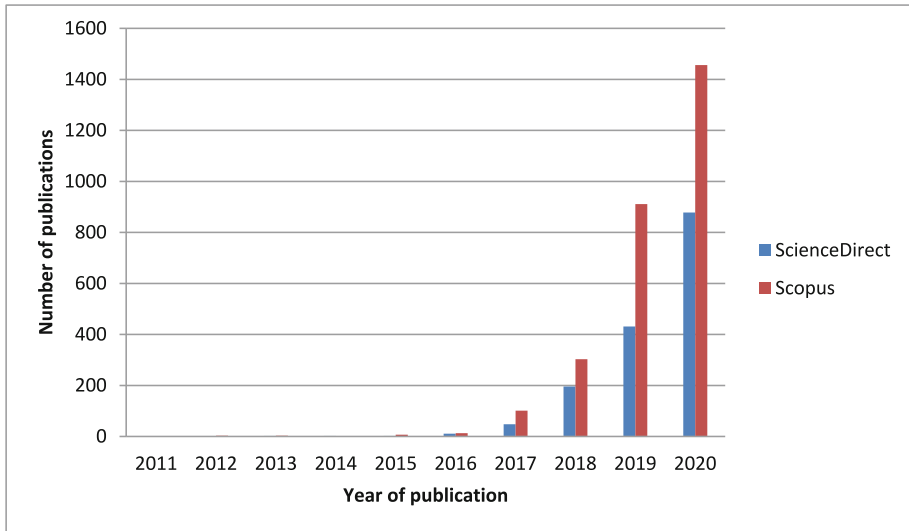


Fig. 1. Number of publications on digital twins from 2011 to 2020 in Scopus and ScienceDirect

influences. The effective operation of the digital twin city requires a continuous flow of data generated by various sources in the digital infrastructure of the smart city. Digital twins have primarily been used in the manufacturing sector, but other areas of study and business are beginning to find new potential uses. An ideal digital twin would be identical to its physical counter-part and have a complete, real-time dataset of all information on the object/system. As the object/system increases in complexity a digital twin may be identical in only relevant areas and have only the real-time data necessary to support any desired simulations. How accurate and useful a digital twin is, depends on the level of detail put into it and how comprehensive the available data is.

In seeking to explore the potential of digital doyens for the smart city, it is necessary to examine what benefits a smart city can bring to cities. A smart city can be considered through three areas: technological, human and institutional, where the main components can be considered transportation, environment, energy, health, security and education [7]. An article [8] identified four key areas necessary for the development of smart cities: planning, physical infrastructure, information and communication technology infrastructure and smart solutions. It was recognized that the approaches used in the smart city contribute to improving the environmental and economic sustainability of cities, as well as improving the provision of services to their residents [9]. It is expected that the benefits of developing smart cities will cover almost all functions and areas of the city.

Digital twins allow for the simulation of many options before taking physical action in the real world to identify the strengths and weaknesses of each plan. This is especially important in safety critical situations, where only one option can be chosen and there may be a number of competing plans to choose from. This is exemplified by the rescue operation in Thailand to save a lost soccer team that occurred in July 2018 [31]. A 3D map of the terrain, a complex cave system, was created using GIS data, water and

oxygen information inside the caves. Weather forecasts were also used in order to create an accurate digital twin that could simulate rescue operations and ensure the safety of the rescuers and the lost team. The use of a digital twin ensured that when rescuers acted, it was a best-case scenario after testing multiple options. Digital twins can have applications in a number of different domains. With the data generated by smart cities, digital twins can be used to model urban planning and policy decisions. An example of a work-in-progress digital twin of a city is Virtual Singapore, which is a three-dimensional (3D) city model and data platform [30].

Each of the city's systems can be represented through its digital twin. For example, a digital twin for smart buildings can be built and used to monitor environmental and social aspects (e.g. temperature, air quality, light). A digital twin of the transportation network can be built to simulate traffic flow, congestion, and accidents under expected population growth scenarios. In addition, a digital twin of rivers can be emulated to simulate water levels under various scenarios to effectively develop water management and flood prevention strategies. Digital twins for water, stormwater, and wastewater can be built to simulate their dependencies as well as their interactions with other systems. A digital twin of the electricity network can also be built, since electricity supply is important for the functioning of other urban infrastructures [12].

3 Methods

DT use various methods for modeling real objects and technological processes, including methods of statistical and intelligent data analysis, computational modeling methods, etc. [13]. Each of the methods applied to DT modeling, imposes special requirements to the necessary computational resources. In this section, the various methods used for the creation of digital twins are discussed.

Numerical simulation is a computer simulation method that implements a mathematical model of a physical system. Numerical modeling is necessary to study the behavior of systems whose mathematical models are too complex to obtain analytical solutions, as in most nonlinear systems.

An example of the application of numerical simulation to build a digital twin can be found in the article [18]. This article presents the results of modeling of dust distribution on the territory of Tbilisi city in winter under westerly winds.

Optimization modeling provides search and finding of the best (optimal) solutions, based on mathematical algorithms. Optimization model consists of the target function, the area of acceptable solutions and the system of constraints defining this area. The main task of optimization modeling is to find an extremum of functions under existing constraints in the form of equations and inequalities.

An example of the application of optimization modeling is the optimization of biomass logistics [14]. In this study, a mathematical model of biomass logistics and the inclusion of interdependent operations related to harvesting, collection, storage, pre-treatment and transportation are considered. Another example is construction optimization [15]. The paper solves the problem of optimizing the construction plan, taking into account the minimization of time and cost of construction at maximum safety, quality and stability. The problem is reduced to the development of multi-criteria optimization algorithms.

Simulation modeling is a research method based on the fact that the system under study is replaced by a model that simulates this system. Experiments are performed on the model and as a result information about the real system is obtained [16]. Examples are the models presented in the article [17]. These models have been used to select the most effective strategies to mitigate the effects of coronavirus infection (COVID-19). In addition, it is shown how simulation modeling helps the government to make the most informed decisions.

Intelligent data analysis (Data Mining) is used to extract implicit, previously unknown and potentially useful information needed to make strategically important decisions in various areas of human activity [19]. As part of the construction of a digital twin city model based on data mining allows to solve the following major problems:

- identification of the influence of factors on each other;
- identification of the degree of influence of factors on the indicator;
- predicting the values of the factors and the indicator.

An example of the use of intelligent data analysis can serve as an article [20]. In this study, an intelligent waste management system is developed to optimize waste collection, based on the data read from the sensors placed in the waste containers.

Neural network modeling is based on artificial neural networks (ANNs) and convolutional neural networks (CNNs). There are different definitions of ANNs, but there are notable similarities with respect to their origin and functioning. The article [21] defines ANNs as computational networks that attempt to mimic the nerve cell networks of the biological central nervous system. A similar definition is given in the article [22], in which ANNs are considered as a new generation of information processing paradigms designed to simulate some models of human brain behavior. The main advantages of ANNs are considered to be their accuracy, speed, volume, convergence, scalability, fault tolerance and performance [23]. A convolutional neural network is a special architecture of artificial neural networks aimed at image analysis.

Neural network modeling is used to solve the problem of detecting objects in an image [24]. An example of a neural network model for urban traffic assessment is the “AIETMS (Artificial Intelligence-Enabled Traffic Monitoring System)” [26].

It is worthwhile to note that most present-day traffic monitoring activity happens at the Traffic Management Centers (TMCs) through vision-based camera systems. However, most existing vision-based systems are monitored by humans which makes it difficult to accurately keep track of congestion, detect stationary vehicles whilst concurrently keeping accurate track of the vehicle count. Therefore, TMCs have been laying efforts on bringing in some levels of automation in traffic management. Automated traffic surveillance systems using artificial intelligence have the capability to not only manage traffic well but also monitor and access current situations that can reduce the number of road accidents. Similarly, an AI-enabled system can identify each vehicle and additionally track its movement pattern characteristic to identify any dangerous driving behavior, such as erratic lane changing behavior. Another important aspect of an AI-enabled traffic monitoring system is to correctly detect any stationary vehicles on the road. Oftentimes, there are stationary vehicles which are left behind and that impedes the flow of preceding vehicles and causes vehicles to stack up. This results in congestion that hampers the free

mobility of vehicles. Intelligent traffic monitoring systems are thus an integral component of systems needed to quickly detect and alleviate the effects of traffic congestion and human factors.

In the last few years, there has been extensive research on machine and deep learning-based traffic monitoring systems. Certain activities such as vehicle count, and traffic density estimation are limited by the process of engaging human operators and requires some artificial intelligence intervention. Traffic count studies for example require human operators to be out in the field during specific hours, or in the case of using video data, human operators are required to watch man hours of pre-recorded footage to get an accurate estimation of volume counts. This can be both cumbersome and time-consuming. Similarly, when it comes to seeing traffic videos from multiple CCTV cameras, it becomes extremely difficult to analyze each traffic situation in real time. Therefore, most TMCs seek out deploying automated systems that can, in fact, alleviate the workload of human operators and lead to effective traffic management system. At the same time, the associated costs are comparatively lower due to savings associated with not needing to store multiple hours of large video data. The study [26] applies several state-of-the-art deep learning algorithms based on the nature of certain required traffic operations. Traditional algorithms [27–29] often record lower accuracies and fail at capturing complex patterns in a traffic scene; hence, we tested and deployed deep learning-based models trained on thousands of annotated traffic images. Thus, the proposed system as shown in Fig. 2 can perform the following:

1. Monitoring traffic congestion
2. Traffic accidents, stationary or stranded vehicle detection
3. Vehicle detection and count
4. Managing traffic using a stand-alone Graphical User Interface (GUI)
5. Scaling traffic monitoring to multiple traffic cameras.

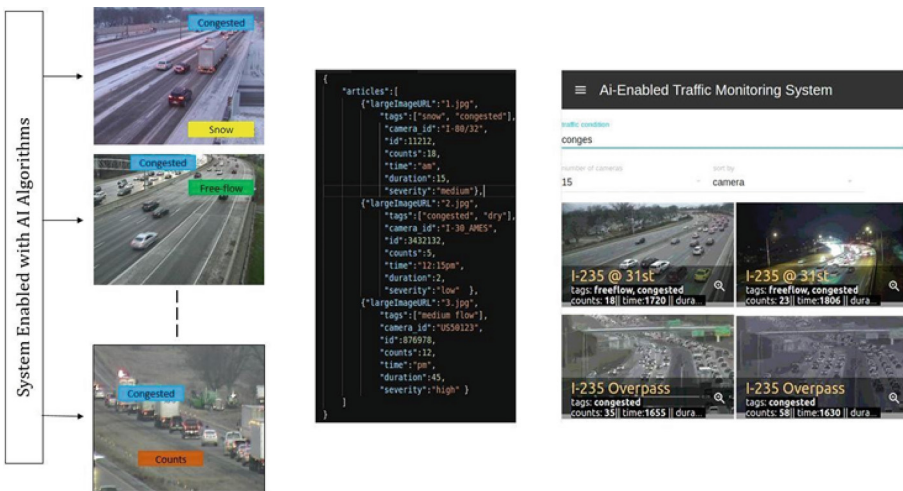


Fig. 2. Example of a neural network for traffic monitoring

In addition, neural network modeling is used in the construction of recommendation systems. For example, decision support systems for urban traffic management in order to reduce harmful emissions into the atmosphere [25].

4 Comparative Analysis

Methods of building digital twins have attracted worldwide attention and are considered to be the key to a smart city. These methods have been analyzed in terms of several characteristics, and the results of the comparison are presented in Table 1.

Data sources for data mining and neural network modelling can be big data. All methods aim to improve efficiency, customer satisfaction and management accuracy, and in their own way reduce costs by facilitating intelligent management.

All of the methods presented make use of IoT technologies. However, data mining and neural network modelling focus more on data-driven technologies, while numerical, optimisation and simulation modelling touch on cyber-physical integration technologies (e.g. virtual and augmented reality).

Intelligent data analysis and neural network modelling can make use of big data technologies. Since big data cannot be processed by conventional data processing tools in an acceptable time, big data processing has its own advanced tools, algorithms, platforms, etc.

Data for intelligent analysis and neural network simulation comes from physical objects, information systems and the Internet, which are generated by activities in the physical world. Data in simulation modeling comes not only from the physical world, but also from virtual models. All methods use the same tools to collect data from the physical world (e.g. sensors and RFID).

Regarding visualisation, data mining and neural network modelling tend to use two-dimensional and static tools such as tables, charts, graphs, etc. Because of virtual models, visualisation in simulation and numerical modelling is more visual, which is mostly three-dimensional and dynamic, such as image, video, virtual and augmented reality, etc.

Thus, in terms of data, intelligent analysis and neural network modelling are more advanced techniques. Numerical, optimisation and simulation modelling are more efficient in applications. It should be noted that all methods differ from each other in detail, but are interrelated in the general direction of their functions.

5 Conclusion

This article has focused on the definition and main benefits of digital twins for cities, as well as examples of solutions that can now be considered the first steps towards building a full-fledged digital twin. Examples of implementation methods of building digital twins were considered. A comparative analysis of existing methods for the construction of digital twins of smart cities was performed. It was revealed the superiority of data mining and neural network modeling over other methods in the context of the considered characteristics. The relationship between the digital twin and big data in the development of a smart city was shown.

The digital twin allows the government to manage real-time two-way comparisons between a physical object and a digital representation, paving the way for cyber-physical integration. Combined with accurate analysis and forecasting capabilities, a city managed by a digital twin may become more predictable for efficient and accurate management in all areas. Further development in the field of smart city digital twins depends largely on advances in neural network construction and the development of ‘advanced’ machine learning algorithms.

Funding. This work was carried out at the North Caucasus Center for Mathematical Research within agreement no. 075-02-2022-892 with the Ministry of Science and Higher Education of the Russian Federation. The reported study was funded by RFBR, Sirius University of Science and Technology, JSC Russian Railways and Educational Fund “Talent and success”, project number 20-37-51004 “Efficient intelligent data management system for edge, fog, and cloud computing with adjustable fault tolerance and security”.

References





1. Anthopoulos, L., Janssen, M., Weerakkody, V.: Smart service portfolios: do the cities follow standards?. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp. 357–362 (2016)
2. Kitchin, R., McArdle, G.: What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets. *Big Data Soc.* **3**(1), 1–10 (2016)
3. Ivanov, S., et al.: Digital twin of city: concept overview. In: 2020 Global Smart Industry Conference (GloSIC), pp. 178–186. IEEE (2020)
4. Grieves, M., Vickers, J.: Digital twin: mitigating unpredictable, undesirable emergent behavior in complex systems. In: Kahlen, F.-J., Flumerfelt, S., Alves, A. (eds.) *Transdisciplinary Perspectives on Complex Systems*, pp. 85–113. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-38756-7_4
5. Hämmäläinen, M.: Smart city development with digital twin technology. In: 33rd Bled eConference-Enabling Technology for a Sustainable Society: June 28–29, 2020, Online Conference Proceedings. University of Maribor (2020)
6. Kritzing, W., et al.: Digital twin in manufacturing: a categorical literature review and classification. In: *IFAC-PapersOnLine*, vol. 51, no. 11, pp. 1016–1022 (2018)
7. Nam, T., Pardo, T.A.: Conceptualizing smart city with dimensions of technology, people, and institutions. In: Proceedings of the 12th Annual International Digital Government Research Conference: Digital Government Innovation in Challenging Times, pp. 282–291 (2011)
8. Kumar, H., et al.: Moving towards smart cities: solutions that lead to the Smart City transformation framework. *Technol. Forecast. Soc. Change* **153**, 119281 (2020)

9. Beier, R., Fritzsche-El Shewy, J.: UN-habitat, the new urban agenda and urban refugees—a state of the art. *Z'Flucht. Zeitschrift für Flucht-und Flüchtlingsforschung*. **2**(1), pp. 128–142 (2018)
10. Singh, M., et al.: Digital twin: origin to future. *Appl. Syst. Innov.* **4**(2), 36 (2021)
11. Glaessgen, E., Stargel, D.: The digital twin paradigm for future NASA and US Air Force vehicles. In: 53rd AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics and Materials Conference 20th AIAA/ASME/AHS Adaptive Structures Conference 14th AIAA, p. 1818 (2012)
12. Haggag, M., et al.: Resilient cities critical infrastructure interdependence: a meta-research. *Sustain. Resilient Infrastruct.* 1–22 (2020)
13. Korambath, P., et al.: A smart manufacturing use case: furnace temperature balancing in steam methane reforming process via kepler workflows. *Procedia Comput. Sci.* **80**, 680–689 (2016)
14. Malladi, K.T., Sowlati, T.: Biomass logistics: a review of important features, optimization modeling and the new trends. *Renew. Sustain. Energy Rev.* **94**, 587–599 (2018)
15. Kandil, A., El-Rayes, K., El-Anwar, O.: Optimization research: enhancing the robustness of large-scale multiobjective optimization in construction. *J. Constr. Eng. Manag.* **136**(1), 17–25 (2010)
16. Maria, A.: Introduction to modeling and simulation. In: *Proceedings of the 29th Conference on Winter Simulation*, pp. 7–13 (1997)
17. Currie, C.S.M., et al.: How simulation modelling can help reduce the impact of COVID-19. *J. Simul.* **14**(2), 83–97 (2020)
18. Surmava, A.A., et al.: Numerical simulation of dust distribution in city tbilisi territory in the winter period. *J. Georgian Geophys. Soc.* **24**(1) (2021)
19. Chen, M.S., Han, J., Yu, P.S.: Data mining: an overview from a database perspective. *IEEE Trans. Knowl. Data Eng.* **8**(6), 866–883 (1996)
20. Oralhan, Z., Oralhan, B., Yiğit, Y.: Smart city application: internet of things (IoT) technologies based smart waste collection using data mining approach and ant colony optimization. *Internet Things* **14**(4), 5 (2017)
21. Graupe, D.: *Principles of Artificial Neural networks*. Advanced Series on Circuits and Systems, 6th edn (2007)
22. Karayiannis, N.B., et al.: New radial basis neural networks and their application in a large-scale handwritten digit recognition problem. In: *Recent Advances in Artificial Neural Networks: Design and Applications*, pp. 39–94 (2000)
23. He, H., Garcia, E.: A learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**(9), 1263–1284 (2009)
24. Yu, L., Chen, X., Zhou, S.: Research of image main objects detection algorithm based on deep learning. In: 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), pp. 70–75. IEEE (2018)
25. Elbir, T., et al.: Development of a GIS-based decision support system for urban air quality management in the city of Istanbul. *Atmos. Environ.* **44**(4), 441–454 (2010)
26. Mandal, V., et al.: Artificial intelligence-enabled traffic monitoring system. *Sustainability* **12**(21), 9177 (2020)
27. Land, E.H.: An alternative technique for the computation of the designator in the retinex theory of color vision. *Proc. Natl. Acad. Sci. USA* **83**, 3078–3080 (1986)
28. Rahman, Z.-U., Jobson, D.J., Woodell, G.A.: Multi-scale retinex for color image enhancement. In: *Proceedings of the 3rd IEEE International Conference on Image Processing*, Lausanne, Switzerland, 19 September 1996, pp. 1003–1006. IEEE, Piscataway (1996)

29. He, K., Sun, J., Tang, X.: Single image haze removal using dark channel prior. *IEEE Trans. pattern Anal. Mach. Intell.* **33**, 2341–2353 (2010)
30. Alam, K.M., El Saddik, A.: C2ps: a digital twin architecture reference model for the cloud-based cyber-physical systems. *IEEE Access* **5**, 2050–2062 (2017)
31. Srimee, N., Cooharajanone, N., Chandrachai, A.: Destination selection in Thailand toward the risk in the eyes of tourist: a case study of Tham Luang Cave. *PSAKU Int. J. Interdisc. Res.* **9**(1) (2020)



Modification of the Projection Method to Correct Errors in RNS

Egor Shiriaev^{1,3}  , Viktor Kuchukov^{2,3} , and Nikolay Kucherov^{1,3} 

¹ North-Caucasus Federal University, Stavropol, Russia

eshiriaev@ncfu.ru

² North Caucasus Center for Mathematical Research NCFU, Stavropol, Russia

³ Sirius University of Science and Technology, Sochi, Russian Federation

Abstract. In this paper, we study methods for correcting errors in the system of residual classes. Traditional correction codes have disadvantages, such as Reed-Solomon codes have a large redundancy, others have either the same disadvantages or have greater computational complexity. The residue number system has self-correcting properties that have less redundancy and computational complexity. Thus, error detection methods were considered. Based on the study, it was found that the most effective method for detecting errors is the nullivisation method.

Keywords: Residue Number system · Chinese Remainder Theorem · Distributed Storage Systems · Correction Codes · Syndrome Method · Nullivisation

1 Introduction

One of the main properties of any digital information system is reliability [1]. The problem of reliability is especially acute in distributed systems [2], such as cloud storage [3] and cloud computing systems [4]. In a decentralized system, data is constantly migrating over a shared network, between service providers and customers. Under such conditions, the likelihood of errors due to various reasons increases. Errors can be caused by both internal and external events.

The most dangerous are random errors that occur at a low hardware and software level. Redundancy, in this case, involves the replication of information and confirmation of its identity. However, this requires an increase in times, both the amount of data storage equipment and physical space for their placement and other overhead costs proportional to the required replication.

However, there is another way to control information errors. This method is based on the self-correcting properties of the residue number system (RNS) [5]. RNS, due to its properties, allows to reduce the computational complexity of data processing and control the correctness of information. Many scientists and scientific groups have been researching this topic recently [6–9]. However, in the methods presented in [6–9], for the most part, to detect an error, either methods are used to return a number from RNS to a weighted number system (WNS), or base extension methods. These methods have high computational complexity. However, there are methods for detecting errors without such

computationally complex operations. In this paper, we will consider an error detection method based on nullivisation a number in an RNS, as well as its application with an error correction method in an RNS.

Thus, this work is presented as follows. Section 1 will describe the RNS and its main properties and characteristics. In Sect. 2, methods for error correction will be considered. Section 2.2 will present methods for converting numbers from RNS to PSS. Section 3 will present a nulling method for determining the error in the RNS. Section 4 will present an experimental study of the methods considered in this work.

2 Materials and Methods

RNS is a non-positional number system based on the Chinese Remainder Theorem (CRT) [10]. CRT is the following: provided that the natural numbers p_1, p_2, \dots, p_n – coprime, then for any integers x_1, x_2, \dots, x_n , such that $0 \leq x_i < n$ at $i \in \{1, 2, \dots, n\}$, there is such X , which, when divided by p_i gives the remainder x_i .

The connection between CRT and RNS is established based on the proof of the theorem. Consider the following system of linear equations modulo:

$$\begin{cases} X \equiv x_1 \pmod{p_1} \\ X \equiv x_2 \pmod{p_2} \\ \dots \\ X \equiv x_n \pmod{p_n} \end{cases} \quad (1)$$

In that case, if x_1, x_2, \dots, x_n and p_1, p_2, \dots, p_n satisfy the condition of the theorem, then there exists a unique solution to system (1) such that:

$$X = \left| \sum_{i=1}^n x_i P_i P_i^{-1} \right|_P, \quad (2)$$

where $P = p_1 \cdot p_2 \cdot \dots \cdot p_n$ – basis of modules p , $P_i = \frac{P}{p_i}$, P_i^{-1} – multiplicatively inverse element [10].

So, the number X in RNS this is a set of leftovers x_1, x_2, \dots, x_n obtained by modulo dividing by a set of relatively prime numbers p_1, p_2, \dots, p_n or RNS modules, with a constraint such that $X \in [0, P]$.

Execution of non-modular operations in RNS – computationally complex operations and their execution are open questions in RNS and are being actively researched at present. [21–24]. One of the most important of these operations is the operation of converting a number from RNS to WNS. The Eq. (2) presents one of the earliest and most computationally complex ways of converting numbers from RNS to WNS, however, there are other methods that reduce the computational complexity of the operation due to the properties of RNS and CRT or using approximate arithmetic.

Redundant RNS (RRNS) [6] suggests that there is a working range of bases p_1, p_2, \dots, p_n and redundant (control) – $p_{n+1}, p_{n+2}, \dots, p_k$ at $j \in [n, k]$, and:

$$\frac{P_k}{p_j} \geq P_n, \quad (3)$$

where P_n – operating range basis, P_k – redundant range basis. The most common use of RRNS is to control errors in RNS, provided that the control remainder/base pairs are absolutely accurate.

2.1 Projection Method

It follows from the theorem presented in [11] that based on inequality (3), we can state that $X < P_n \leq \frac{P_k}{p_j}$ or $X < \frac{P_k}{p_j}$, then the assertion is also true that

$$X < p_1 p_2, \dots, p_{i-1} p_{i+1}, \dots p_{n+1}, \tag{4}$$

Hence the number X can be represented by its residuals x_1, x_2, \dots, x_n by modules p_1, p_2, \dots, p_n in one single way.

Thus, excluding i we will also get a remainder-base pair X , what does it have if condition (4) is satisfied. Then we can introduce the concept of the projection of a number X .

Number projection X_i in RNS – it’s the same number X obtained by removing the residue from the RNS x_i and the basis on which it was received. So any number represented in RNS, have n projection.

It also follows from the theorem presented and proven in [11] that

$$\begin{aligned}
 X_1 = X_2 = \dots = X_i = \dots = X_{n+1} \\
 \text{provided that} \\
 X < \frac{P}{p_i}
 \end{aligned}
 \tag{5}$$

which allows us to assert the following - the number X has no error if statement (5) is valid. However, this projection method is not effective. To check, it is necessary to calculate all the projections of the number and perform the operation of returning the number to the WNS for each. However, this way you can detect one error in the number. The development of this method was derived from the fact that the control base is always correct and cannot be distorted.

Then you can adjust the number by iteratively calculating the projections by replacing the remainder under study.

Thus, we need n iterations to find the error. The weak point of the method is the conversion of a number from RNS to WNS, which requires a lot of computing resources.

There is another type of projection method, it allows you to reduce the number of iterations by introducing an additional control base. Its essence lies in the fact that when creating a projection, not one residue is deleted, but a pair of them. However, in this case, a restriction is imposed on the method due to the parity of the length of the moduli vector.

Another way to increase the efficiency of the error correction method is to modify the error detection method. As for example, the use of other, more computationally simple methods for converting from RNS to WNS.

2.2 Converting Numbers from RNS to WNS Based on CRT

As already mentioned in Sect. 3, CRT guarantees the uniqueness of the representation of a number in the range $(0, P]$ returned to the WNS according to the formula (2). This formula is called the Garner method or the CRT method [5]. This formula can also be rewritten in the following form (6):

$$X = \left| X'_1 + X'_2 + \dots + X'_n \right|_P \quad (6)$$

where $X'_i = x_i \cdot \left| P_i^{-1} \right|_{p_i} \cdot P_i$.

Analyzing method, we can tell that its weak point is the operation of modulo division of a large number X by a large number P . There are many different methods that allow you to get rid of this drawback, such methods will be considered below.

2.3 Modified CRT Method

The modified CRT method was presented in [12], it consists in modifying the CRT (Sect. 4.1) by the Mixed Radix System (MRS) [13].

The modification consists in introducing mixed coefficients γ_i the meaning of which is to obtain the inverse multiplicative number from the basis of each module (7):

$$\begin{aligned} \gamma_1 &= \frac{(P_1 \cdot \left| P_1^{-1} \right|_{p_1})^{-1}}{p_1} \\ \gamma_i &= \frac{P_i}{p_1} \cdot \left| P_i^{-1} \right|_{p_i} \end{aligned} \quad (7)$$

Having obtained the MRS coefficients, you can restore the number based on the MRS method [14]:

$$\begin{aligned} X &= x_1 + p_1 \cdot \left| \gamma_1 \cdot x_1 + \gamma_2 \cdot x_2 \right|_{p_2} + p_1 p_2 \cdot \left[\frac{\gamma_1 \cdot x_1 + \gamma_2 \cdot x_2 + \gamma_3 \cdot x_3}{p_2} \right]_{p_2} + \\ &+ \dots + p_1 p_2 \dots p_{n-1} \left[\frac{\gamma_1 \cdot x_1 + \gamma_2 \cdot x_2 + \dots + \gamma_n \cdot x_n}{p_2 p_3 \dots p_n} \right]_{p_n} \end{aligned} \quad (8)$$

Thus, the advantage of this method can be considered the absence of dividing a large number X by a large number P . The disadvantages can be considered the storage of additional constants in memory, as well as many auxiliary multiplications and additions.

2.4 Core Function Method

The Akushsky core function [15] is a monotonic function of the form:

$$C(X) = \sum_{i=1}^n w_i \cdot \left[\frac{X}{p_i} \right] = \sum_{i=1}^n \frac{w_i}{p_i} \cdot X - \sum_{i=1}^n \frac{w_i}{p_i} \cdot x_i$$

Then you can convert a number from RNS to WNS based on the expression (9):

$$X \equiv \frac{P \cdot C(X) + \sum_{i=1}^n w_i \cdot P_i \cdot x_i}{C(P)} \quad (9)$$

where $C(P) = \sum_{i=1}^n \frac{w_i}{p_i} \cdot P_i$, $w_i = \left| \left| P_i^{-1} \right|_{p_i} \cdot P_n \right|_{p_i}$.

However, this method is inefficient. It requires a lot of divisions and multiplications by a large number P . Consider its modification [16].

Let us introduce the notion of an orthogonal basis [17]. An orthogonal basis in an RNS is a representation of the basis of the number of RNS in a vector representation $B = \{1, 0, \dots, 0\}, \{0, 1, \dots, 0\} \dots \{0, 0, \dots, 1\}$. Then to find B_i it is necessary to obtain a representation of the basis for each base $-P_i$ and the reciprocal of the base, i.e. its multiplicative inverse $- \left| P_i^{-1} \right|_{p_i}$. Then the orthogonal basis of a number can be found as:

$$B_i = P_i \cdot \left| P_i^{-1} \right|_{p_i} \tag{10}$$

Then the core function B_i

$$C(B_i) \equiv \frac{(B_i \cdot P_n)}{P} - \frac{w_i}{p_i}$$

Based $C(B_i)$ can be obtained and $C(P)$

$$C(P) \equiv \left| \sum_{i=1}^n x_i \cdot C(B_i) \right|_{P_n}$$

And then the translation is performed according to the following formula

$$X \equiv \left| \frac{P}{P_n} \cdot \left(C(P) + \sum_{i=1}^n \frac{w_i}{p_i} \cdot x_i \right) \right|_P \tag{11}$$

This method has similarities with the Garner method and its inherent disadvantages, such as division by a large number P . However, there is a modification of this method.

2.5 Core Diagonal Function Method

In this case, the authors of [18] modified the core function method based on its monotonicity and diagonal properties of functions, setting its weights w_i equal to one.

Then the formulas will look like this:

$$D(X) = (x_1 \cdot k_1 + x_1 \cdot k_1 \dots x_n \cdot k_n)_P$$

where $k_i = P_i \cdot \left| P_i^{-1} \right| - \frac{1}{p_i}$

Or when viewed from the point of view of the value of the number itself from the function $D(X)$

$$X \equiv \frac{(\sum_{i=1}^n x_i \cdot P_i + P \cdot D(X))}{P}$$

However, this formula is not efficient, since it requires many actions with a large number P . Then we transform this method into a core function.

Diagonal Core Function B_i

$$C_d(B_i) \equiv \frac{(B_i \cdot P_n)}{P} - \frac{1}{p_i}$$

Diagonal basis function $C_d(P)$

$$C_d(P) \equiv \left| \sum_{i=1}^n x_i \cdot C_d(B_i) \right|_{P_n}$$

And then the translation is performed according to the following formula

$$X \equiv \left| \frac{P}{P_n} \cdot \left(C_d(P) + \sum_{i=1}^n \frac{x_i}{p_i} \right) \right|_P \quad (12)$$

Despite the modification of the method, the core function method is still not efficient enough. However, there is a method based on approximate arithmetic.

2.6 Core Diagonal Function Method

An approximate method for converting numbers from RNS to WNS was obtained in [19]. This method was based on the properties of fractional values for the CRT approximation - this was used in RNS to develop methods for determining the sign of a number, as well as the division operation [20]. Thus, we can get an approximate (rounded) value of X based on the following formula

$$\widehat{X} = \left| \sum_{i=1}^n k_i \cdot x_i \right|_1 \quad (13)$$

where $k_i = \frac{|P_i^{-1}|_{p_i}}{p_i}$.

This method can be modified to a more accurate calculation of the value of the desired X . In [19], such a modification was carried out. It follows from it that the number can be restored by introducing an approximate coefficient $N = \log_2 P \cdot \mu$, where $\mu = \sum_{i=1}^n p_i$. Then, based on the properties of the binary RNS, we can introduce 2^N for approximate calculation X . Approximate factor k_i takes the following form:

$$k'_i = \left| \frac{|P_i^{-1}|_{p_i}}{p_i} \cdot 2^N \right| \quad (14)$$

Then

$$\left[\frac{X}{P} \right] = \left| \sum_{i=1}^n k'_i \cdot x_i \right|_{2^N} \quad (15)$$

And the number itself in the WNS will look like

$$X = \left\lfloor \frac{\left\lfloor \frac{X}{P} \right\rfloor \cdot P}{2^N} \right\rfloor \tag{16}$$

This method is the most efficient due to the low computational complexity of the operations, however, it requires a lot of precomputations.

Thus, based on the translation methods in this section, it is possible to return a number to the WNS to check the number for an error. However, such methods cannot be called effective in terms of error detection. Since it is necessary to transfer the number from RNS to WNS at each iteration of the projection method. Let’s consider an alternative method.

2.7 Nullivisation Method for Determining Error in RNS

Akushsky I. Ya. and Yuditsky D. Yu. in the work [11], in chapter 5, the method of nullivisation numbers in RNS was given.

The meaning of the nullivisation method is to check the correctness of the number. Since only a correctly represented number in the RNS can turn into zero. Nullivisation is performed based on constant precalculation M_i^j :

$$M_i^j = \{t_1^1, t_2^1, \dots, t_n^m\} \tag{17}$$

where $t_i^j \in 1, 2, \dots, p_i - 1, i \in \{0, n\}, j \in \{0, p_i - 1\}$. Moreover, each member of the vector M_i^j less i equals zero. Then the set of constants can take the following form:

$$\begin{aligned} M_1^j &= \{t_1^j, t_2^j, \dots, t_{n+k}^j\} \\ M_2^j &= \{0, t_2^j, \dots, t_{n+k}^j\} \\ &\dots \\ M_n^j &= \{0, 0, \dots, 0, t_n^j, \dots, t_{n+k}^j\} \end{aligned}$$

where k – the number of redundant modules. The number of nullivisation constants is determined by the expression $\sum_{i=1}^n p_i - 1$.

Then nullivisation occurs on the basis of calculations X_{S_i} values:

$$X_{S_i} = x - M_i^j = \{0, S_2, S_3, \dots, S_{n+1}\} \tag{18}$$

Moreover, it is known from [11] that it is necessary $n - 1$ such operations to ensure nullivisation of the number. In this case, the value j defined as equal x_i in X_{S_i} .

Thus, this method of identifying an error in the RNS is more efficient than converting the number from the RNS to the WNS. However, in this case, we can state this based on mathematical formulas and pseudocode methods. In order to get an actual evaluation of the methods, it is necessary to conduct performance simulations.

3 Results

The research will be conservative. Two sets of reciprocal primes are being prepared. Thus, this study will provide the most accurate picture of the software implementation of the methods, as well as assess the possible load.

Consider the obtained performance measurement results:

For greater clarity, we will build a graph based on the data in the table, where the abscissa axis is the size of the module in bits, and the ordinate axis is time in microseconds.

Then we have the following illustration (Fig. 1).

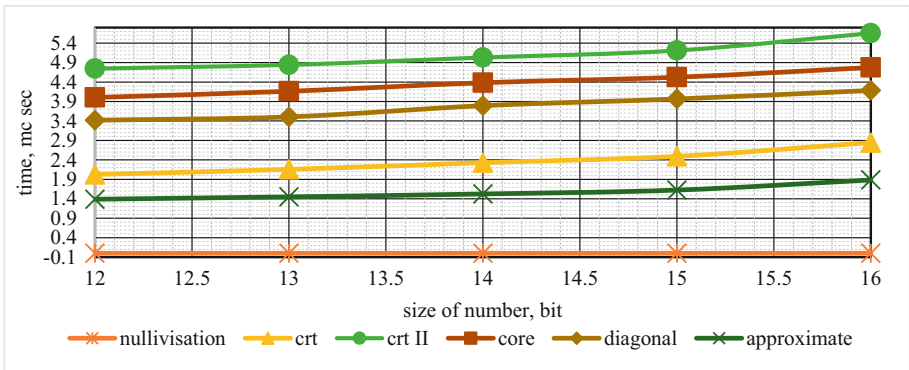


Fig. 1. Graph of the performance of the methods depending on the size of the bases.

Considering the illustration (Fig. 1), you can notice that the nullivisation method is the most productive, since its method has only $n - 1$ iteration. And the method itself uses only assignment, subtraction and modulo division operations. Since this uses numbers only in the range from 1 before r bit where r – base size. Then all operations are computationally simple. It should also be noted that the performance drop of other methods is quite low and is within one microsecond. Next, consider the study of the second set of base vectors.

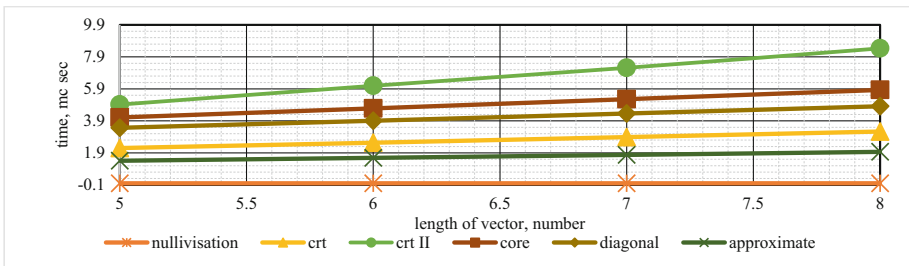


Fig. 2. Graph of the performance of methods depending on the length of the RNS base vector

Just as in the case of the previous experiment, let us consider and illustrate the result of the experiment. The abscissa axis is the number of bases in the RNS basis vector. Y-axis - time in microseconds (Fig. 2).

In this case, the most productive method is still the same - the nullivisation method. All methods have a linear drop in performance, since with an increase in the number of bases, the number of operations performed also increases. The observed linear growth allows us to say that the methods have stable operation, which is not disturbed with increased load (as in the case of Fig. 1). Thus, the nullivisation method has a clear advantage. However, as was said in Sect. 6, the constants are a matrix $n \times m$, где $m = \sum p - 1$. And with a base size of 16 bits, the number of nullivisation constants can reach a million. Therefore, for an objective study, it is necessary to conduct studies on the used computer memory by software implementations of these methods.

This study takes into account the constants that are calculated at the stage of precalculations, i.e. permanently stored in the computer memory during data processing. The sets of modules are the same as in the previous stage of research. The value of memory is given in Kbytes.

Let's build a graph and consider the resulting illustration. The abscissa axis is the module size in bits, the ordinate axis is the total amount of memory used by the method constants in Kbytes.

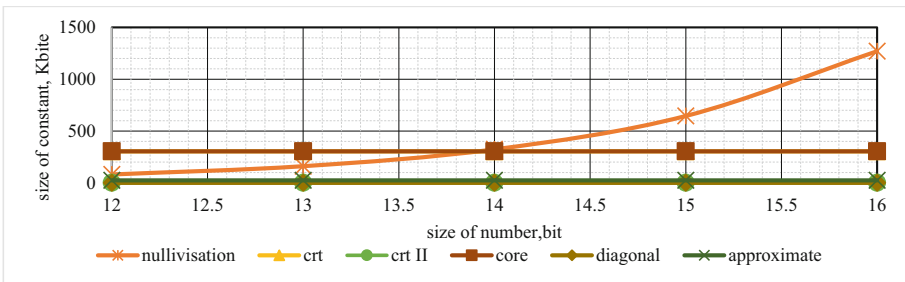


Fig. 3. Plot of memory used by method constants versus base size.

As can be seen in Fig. 3, the memory used by the methods does not practically increase in comparison with the nullivisation method. Since, despite a certain increase in the values of quantities - in the RAM of a computer, the place they occupy practically does not increase, based on the features of the development environment and the presentation of data in memory, when, as in the case of the nullivisation method, it grows exponentially. This situation comes from the numerical value of the bases, since the number of constants directly depends on the bases. From the point of view of consumed memory, the nullivisation method is the least productive when the size of the bases is more than 14 bits.

Consider studies with a different set of modules. Just like in the previous experiment - visualize the data as a graph (Fig. 4).

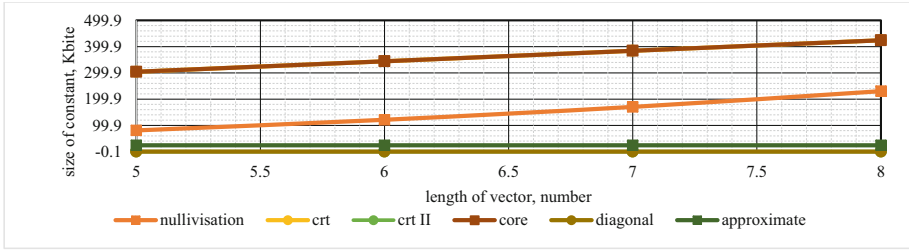


Fig. 4. Plot of memory used by method constants depending on the size of the bases depending on the length of the RNS base vector

In this study, the worst methods are the kernel function method and the modified CRT method. Nullivisation in this case is between the most capacious methods, along with the approximate method, and the costliest.

In this case, we will conduct an additional study, namely, we will compare the methods of nullivisation and approximation.

4 Conclusion

Thus, in this work, a study was made of the performance of SOC correction codes, namely, the projection method from the side of error detection. A study of the literature on this topic was carried out. In the work, the RNS itself was considered from the point of view of its properties. Correction codes in WNS and RNS, as well as methods for converting a number from RNS to WNS. The nullivisation method was designated as the target method, which was considered in a separate section. Further, an experimental study of software implementations was carried out. In the course of which it was found that the nullivisation method has high performance in terms of program execution time, which is its advantage, the disadvantages can be identified as the need for additional memory of the computing system.

Based on the study, we can conclude that the use of the projection method for error correction together with nullivisation can increase the performance of the computing system in terms of checking information for correctness. The additional redundancy to the RRNS is also not significant compared to other error correction methods in the WNS.

This method can be used in various computing systems, such as distributed computing systems for cloud storage and cloud computing.

Acknowledgments. This work was carried out at the North Caucasus Center for Mathematical Research within agreement no. 075-02-2022-892 with the Ministry of Science and Higher Education of the Russian Federation. The study was financially supported by the Russian Foundation for Basic Research within the framework of the scientific project No. 20-37-51004 “Effective intelligent data management system in edge, fog and cloud computing with adjustable fault tolerance and security” and Russian Federation President Grant MK-1203.2022.1.6, and SP-3186.2022.5.

References

1. Gromov, Yu.Yu., Diedrich, I.V., Ivanova, O.G., Paladiev, V.V., Yakovlev, A.V.: Reliability of information systems (2015)
2. Babeshko, V.N., Medvedeva, V.A., Kishchenko, I.I.: Heterogeneous distributed systems in foggy network infrastructures. In: Innovation in Construction Through the Eyes of Young Professionals, pp. 39–40 (2014)
3. Logvinova, Yu.V.: Cloud data storage. *Online Electron. J.* **43** (2016)
4. Golovan, A.M., Klashanov, F.K., Petrova, S.N.: Cloud computing. *Vestnik MGSU* **6**, 411–417 (2011)
5. Garner, H.L.: The residue number system. Papers presented at the Western Joint Computer Conference, 3–5 March 1959, pp. 146–153, March 1959
6. Yang, L.L., Hanzo, L.: Redundant residue number system based error correction codes. In: IEEE 54th Vehicular Technology Conference. VTC Fall 2001. Proceedings (Cat. No. 01CH37211), vol. 3, pp. 1472–1476. IEEE, October 2001
7. Sachenko, A., Zhengbing, H., Yatskiv, V.: Increasing the data transmission robustness in WSN using the modified error correction codes on residue number system. *Elektronika ir elektrotechnika* **21**(1), 76–81 (2015)
8. Shiryaev, E., Bezuglova, E., Babenko, M., Tchernykh, A., Pulido-Gaytan, B., Cortés-Mendoza, J.M.: Performance impact of error correction codes in RNS with returning methods and base extension. In: 2021 International Conference Engineering and Telecommunication (En&T), pp. 1–5. IEEE, November 2021
9. Yatskiv, V., Tsavolyk, T., Zhengbing, H.: Multiple error detection and correction based on modular arithmetic correcting codes. In: 2015 IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), vol. 2, pp. 850–854. IEEE, September 2015
10. Pei, D., Salomaa, A., Ding, C.: Chinese Remainder Theorem: Applications in Computing, Coding, Cryptography. World Scientific, Singapore (1996)
11. Parthasarathy, S.: Multiplicative inverse in mod (m). *Algologic Tech. Rep.* **1**, 1–3 (2012)
12. Akushsky, I.Ya., Yuditsky, D.I.: Machine arithmetic in residual classes. *Sov. radio* (1968)
13. Bi, S., Gross, W.J.: The mixed-radix Chinese remainder theorem and its applications to residue comparison. *IEEE Trans. Comput.* **57**(12), 1624–1632 (2008)
14. Yassine, H.M., Moore, W.R.: Improved mixed-radix conversion for residue number system architectures. *IEE Proc. G (Circuits Devices Syst.)* **138**(1), 120–124 (1991)
15. Akkal, M., Siy, P.: A new mixed radix conversion algorithm MRC-II. *J. Syst. Architect.* **53**(9), 577–586 (2007)
16. Akushskii, I.J., Burcev, V.M., Pak, I.T.: A new positional characteristic of non-positional codes and its application. In: Coding Theory and the Optimization of Complex Systems. SSR, Alm-Ata’Nauka’Kazah (1977)
17. Krishnan, R., Ehrenberg, J., Ray, G.: A core function based residue to binary decoder for RNS filter architecture. In: Proceedings of the 32nd Midwest Symposium on Circuits and Systems, pp. 837–840. IEEE, August 1989
18. Ananda Mohan, P.V.: RNS to binary conversion using diagonal function and Pirlo and Impedovo monotonic function. *Circuits Syst. Sig. Process.* **35**(3), 1063–1076 (2016)
19. Chervyakov, N.I., Molahosseini, A.S., Lyakhov, P.A., Babenko, M.G., Deryabin, M.A.: Residue-to-binary conversion for general moduli sets based on approximate Chinese remainder theorem. *Int. J. Comput. Math.* **94**(9), 1833–1849 (2017)
20. Hung, C.Y., Parhami, B.: An approximate sign detection method for residue numbers and its application to RNS division. *Comput. Math. Appl.* **27**(4), 23–35 (1994)

21. Miranda-López, V., Tchernykh, A., Cortés-Mendoza, J.M., Babenko, M., Radchenko, G., Neschachnow, S., Du, Z.: Experimental analysis of secret sharing schemes for cloud storage based on RNS. In: Mocskos, E., Neschachnow, S. (eds.) CARLA 2017. CCIS, vol. 796, pp. 370–383. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73353-1_26
22. Babenko, M., et al.: RNS number comparator based on a modified diagonal function. *Electronics* **9**(11), 1784 (2020)
23. Tchernykh, A., et al.: An efficient method for comparing numbers and determining the sign of a number in RNS for even ranges. *Computation* **10**(2), 17 (2022)
24. Shiryaev, E., Golimblevskaia, E., Babenko, M., Tchernykh, A., Pulido-Gaytan, B.: Improvement of the approximate method for the comparison operation in the RNS. In: 2020 International Conference Engineering and Telecommunication (En&T), pp. 1–6. IEEE



An Overview of Modern Fully Homomorphic Encryption Schemes

Ekaterina Bezuglova^{2,3}(✉)  and Nikolay Kucherov^{1,3} 

¹ North-Caucasus Federal University, Stavropol, Russia

² North Caucasus Center for Mathematical Research NCFU, Stavropol, Russia
bezuglovakaterina@mail.ru

³ Sirius University of Science and Technology, Sochi, Russian Federation

Abstract. In this paper, research is carried out to ensure the security of cloud storages. The work is devoted to homomorphic encryption. Even though the very concept of homomorphic encryption appeared in the late 70s, a fully homomorphic encryption scheme was obtained only in 2009. Since that time, a variety of fully homomorphic encryption schemes have been developed. In this paper, we analyze the most popular encryption schemes, as well as their software implementations. As a result of the study, it was found that the most promising scheme is CKKS. This scheme can perform homomorphic arithmetic on rational numbers, and already has several software implementations distributed under a free license.

Keywords: Homomorphic Encryption · Cloud Storage · Cloud Computing · Lattice Encryption · Ring Error Learning · Gentry scheme · DGHV · BFV · CKKS

1 Introduction

Now, progress in the IT industry has led to the fact that most of the information created and consumed by a person is in digital format. In such a situation, services for remote storage and processing of information, the so-called cloud storage, are gaining more and more popularity. Cloud storages are distributed data storage systems (DDSS) [1]. In addition to storing data, clouds also process them. In addition, due to the high demand for computing power, cloud service [2] providers in recent years have actively begun to expand services to provide their cloud networks not only for data storage, but also for computing. That is, the user does not need to have expensive equipment with large computing power, instead, he can rent part of the power from a cloud service provider.

Such a problem creates the task of building a security system that eliminates all or at least part of the vulnerabilities in cloud structures.

The solution to this problem can be considered fully homomorphic encryption (FHE) [3]. HE is an evolution of homomorphic encryption. Which, in fact, has existed since the middle of the 20th century.

A homomorphic cipher is an encryption scheme that has the sign of homomorphism. In the case of encryption, this is the ability to perform either homomorphic

addition or homomorphic multiplication. That is, the encrypted data can either be added or multiplied.

Until 2009, it was believed that a cipher could be homomorphic either by addition or by multiplication, but this year Gentry presented the HE schemes in his dissertation [4], marking the beginning of active development in this area.

Now, there are many different FHE schemes and their implementations. In this paper, we will consider the most popular and modern FHE schemes; because of the review, a comparative characteristic of the schemes and their implementations is presented.

The work will be presented as follows: Sect. 2 will present the basic concepts of HE and FHE, Sect. 3 will present modern schemes of FHE, Sect. 4 will present an overview with their comparative characteristics, and in conclusion, conclusions obtained from a comparative characteristic will be presented.

2 Materials and Methods

2.1 First Homomorphic Ciphers and Semi Homomorphic Encryption

In general, the HE can be shown quite simply in the form of a formula. Let's take the information m and present it in two parts m_1 and m_2 , and the encryption key k then the encryption function can be displayed as $Enc(k, m_i)$, while decryption $Dec(k, m_i)$. The ciphertext can be represented as $c = Enc(k, m_i)$. Then the formula explaining the principle HE will look like this:

$$Dec(Enc(k, m_1) * Enc(m_2)) = Dec(m_1 * m_2) = Dec(c) = m_1 + m_2 = m \quad (1)$$

where $*$ - homomorphic operation of addition/multiplication. Initially, the term "homomorphic encryption" was introduced in the work of Ronald Rivest, Leonard Adleman and Michael Dertouzos, who are the authors of the RSA algorithm, in 1978 [5] (the RSA algorithm itself was developed in 1977, but the work was published in 1978).

In general, homomorphic encryption can be described in terms of 4 main operations:

1. Encryption key generation - in homomorphic encryption, there are two keys, as in any asymmetric encryption - the public key pk and the secret key sk . Based on the concept of asymmetric encryption, one key is used to encrypt a message (public), and the other to decrypt (secret). This separation of keys makes it possible to increase the stability of the system, since knowing only the public key, the user cannot decrypt the system, this also increases the efficiency of the system without lowering its security.
2. Encryption - encryption is carried out as follows: first, the plain text is converted, then the plain text is converted into closed text, i.e., $CT = E_{sk}(PT)$.
3. Calculations - performed on the ciphertext using the key pk .
4. Decryption is done with a key sk

Let's look at some of the most famous SHE schemes. Since the main topic of the work is still FHE, the rest of the HE schemes will be given in the table at the end of the section.

The first PHE system is RSA. Let's consider this system. RSA is an encryption system developed by Rivest et al. sometime after the discovery of public key cryptography by

Diffie and Hellman. RSA is considered the first public key cryptosystem to be developed. The RSA homomorphic property of Rivest et al. presented immediately after the main publication [5]. The term privacy homomorphism was indeed introduced by Rivest et al. in the seminal work, which confirms the fact that RSA is the first PHE scheme. RSA uses the factorization problems of the product of two large primes to provide high cryptographic strength. Consider the basic operations in RSA:

- *KeyGen*: start with 2 large primes p and q , after which it is necessary to calculate $n = pq$ and $\phi = (p - 1)(q - 1)$. Next is selected e so $\gcd(e, \phi)$ and d , as a multiplicative inverse to e (i. e. $ed \equiv 1 \pmod{\phi}$). Eventually (e, n) represented as a pair of public keys, and (d, n) then a pair of secret keys.
- *Encryption*: carried out in several stages. The first stage converts the message into the so-called plain text, which is within $0 \leq m < n$. The second stage performs the encryption as follows:

$$c = E(m) = m^e \pmod{n}, \quad \forall m \in M \quad (2)$$

where c – ciphertext.

- *Decryption*: message m is recovered from c based on secret key (d, n) based on the following formula:

$$m = D(c) = c^d \pmod{n} \quad (3)$$

- *Homomorphic multiplication*: take open texts $m_1, m_2 \in M$, let's encrypt them. Then homomorphic multiplication can be represented by the following expression

$$\begin{aligned} E(m_1) \cdot E(m_2) &= (m_1^e \pmod{n}) \cdot (m_2^e \pmod{n}) = \\ &= (m_1 \cdot m_2)^e \pmod{n} = E(m_1 \cdot m_2) \end{aligned} \quad (4)$$

The homomorphism of RSA shows that $E(m_1 \cdot m_2)$ can be directly calculated using $E(m_1)$ and $E(m_2)$, without deciphering them. Thus, RSA has a homomorphism over the operation of multiplication, but the scheme does not have a homomorphic addition.

Next, consider the Goldwasser-Micali (GM) scheme [6]. Goldwasser and Michali proposed their probabilistic public key encryption scheme, which is called GM. In the scheme under consideration, cryptography is conditioned by the solution of the quadratic residual problem [7]. This problem can be considered as follows: $a \pmod{n}$ is called a quadratic residue provided that $x^2 \equiv a \pmod{n}$. This checks whether the number is q quadratic modulo n . Then the operations in the GM cryptosystem can be described as follows:

- *KeyGen*: similarly, RSA builds keys based on the ratio $n = pq$, where p и q – various large prime numbers, after which you need to choose x which is the quadratic value of nonresidues modulo n $\left(\frac{x}{n}\right) = -1$. Then (x, n) is the public key, and (p, q) secret key.
- *Encryption*: message m must be represented as a string of bits. After for each message bit m_i creates a quadratic non-residue value y_i such that $\gcd(y_i, n) = 1$. Then it is necessary to encrypt each bit for c_i based on the following expression:

$$c_i = E(m_i) = y_i^2 x^{m_i} \pmod{n}, \quad \forall m_i \in \{0, 1\} \quad (5)$$

where $m = m_0 m_1 \dots m_r$, $c = c_0 c_1 \dots c_r$, a r – block size in message space. x is in Z_n^* and is chosen randomly from this space for each encryption operation. It is worth noting that Z_n^* – multiplicative subgroup of integers modulo n . The subgroup includes all numbers that are less than n and coprime with n .

- *Decryption*: because $x \in Z_n^*$ ($1 < x \leq n - 1$), then $x \bmod n$ – this is a quadratic residue only for $m_i = 0$. Therefore, to decrypt the ciphertext c_i , need to check c_i is it a quadratic residue modulo n , if c_i is a quadratic residue, then m_i returns 0, otherwise m_i returns 1.
- *Addition homomorphism*: performed for each bit $m_i \in \{0, 1\}$:

$$\begin{aligned} E(m_1) + E(m_2) &= (y_1^2 x^{m_1} \pmod{n}) + (y_2^2 x^{m_2} \pmod{n}) = \\ &= (y_1^2 + y_2^2) x^{m_1 + m_2} \pmod{n} = E(m_1 + m_2) \end{aligned} \quad (6)$$

So, GM encryption amount $E(m_1 \oplus m_2)$ can be directly computed from separately encrypted bits $E(m_1)$ and $E(m_2)$. Given that the message and the ciphertext belong $\{0, 1\}$, the operation is like XOR. Therefore, GM is homomorphic with respect to binary addition.

Next, consider the El-Gamal scheme [8]. Based on a key exchange algorithm based on problems with the discrete Diffie-Hellman logarithm [9], Taher Elgaman created his own encryption scheme. This scheme is a hybrid encryption system. The secret key is calculated for a symmetric encryption system. Consider the basic operations in El-Gamal:

- *KeyGen*: generator based g a cyclic group is generated G order n . At G it is possible to generate all elements of a group using the cardinalities of one of its own elements. Further based on the element $y \in Z^*n$ calculated $h = y$. Based on this, the public key (G, n, g, h) and secret key x .
- *Encryption*: message m is encrypted based on g and x . x must be chosen randomly from a set $\{1, 2, \dots, n - 1\}$. Then the result of the encryption operation is a ciphertext pair $c = (c_1, c_2)$, which can be represented by the relation:

$$c = E(m) = (g^x, mh^x) = (g^x, mh^{xy}) = (c_1, c_2) \quad (7)$$

- *Decryption*: the ciphertext is taken from and computed $s = c_1^y$, where y – the secret key. Decryption is defined by the following expression:

$$c_2 \cdot s^{-1} = mg^{xy} \cdot g^{-xy} = m \quad (8)$$

- *Multiplicative homomorphism*: is defined by the following expression:

$$\begin{aligned} E(m_1) \cdot E(m_2) &= (g^{x_1}, m_1 h^{x_1}) \cdot (g^{x_2}, m_2 h^{x_2}) \\ &= (g^{x_1 + x_2}, m_1 \cdot m_2, h^{x_1 + x_2}) = E(m_1 \cdot m_2) \end{aligned} \quad (9)$$

Considering the ElGamal cryptosystem, it can be argued that the scheme is multiplicatively homomorphic. However, it does not support ciphertext homomorphic addition operations.

2.2 Fully Homomorphic Encryption

Consider the history of the appearance of FHE on the scheme presented by Gentry in his dissertation [4]. The Gentry scheme encrypts the message by injecting noise into it. The Gentry scheme began its history with SWHE based on ideal cryptographic lattices [11].

Let us denote that the SWHE scheme [11] allows homomorphic evaluation of the ciphertext only for a limited number of operations. This means that the decryption function can recover the original message up to a certain threshold. The envy threshold of the ciphertext noise level that must be reduced to convert the noisy ciphertext back to plaintext. To increase the Gentry threshold, revolutionary scheme methods were used: squashing and bootstrapping, to obtain the ciphertext. These methods allow you to perform several homomorphic operations on it in a larger number. Methods can be repeated an unlimited number of times. In other words, it is possible to evaluate an unlimited number of ciphertext operations that make the scheme completely homomorphic. Consider the operations in this scheme:

- *KeyGen*: a ring is used to generate the key R and basis B_I ideal I . Then $IdealGen(R, B_I)$ generates a couple (B_J^{sk}, B_J^{pk}) , where $IdealGen()$ – is an algorithm that derives the simple public secret key bases of an ideal lattice with basis B_I , where $I + J = R$. Then $Samp()$ needed when generating keys for selecting a cost of an ideal, in which each coset is computed by shifting the ideal by a certain length. Then the public key can be obtained at the output, which is the following function $(R, B_I, B_J^{pk}, Samp())$, however, the secret key is only B_J^{sk} from original function.
- *Encryption*: to encrypt a message, you must select a vector \vec{r} and \vec{g} randomly. Using the basis B_{pk} , which is chosen as the most unsatisfactory base from the ideal lattice L , then message $\vec{m} \in \{0, 1\}^n$ encrypted based on the following message:

$$c \vec{\rightarrow} = E(\vec{m} + \vec{r} \cdot B_I + \vec{g} \cdot B_J^{pk}) \quad (10)$$

In the expression (16) $\vec{m} + \vec{r} \cdot B_I$ called the noise parameter.

- *Decryption*: to decipher \vec{c} , you need to take the secret key, and execute the following expression:

$$\vec{m} = \vec{c} - B_J^{sk} \cdot \left[(B_J^{sk})^{-1} \cdot \vec{c} \right] \text{ mod } B_I \quad (11)$$

where $\lfloor \cdot \rfloor$ – nearest integer function that returns the nearest integers for the coefficients of the vector.

Next, consider the homomorphism of the operations of addition and multiplication.

- *Homomorphism over addition/multiplication*: holds for plaintext vectors $m_1, m_2 \in \{0, 1\}^n$ additive homomorphisms are easily verified as follows:

$$\begin{aligned} \vec{c}_1 * \vec{c}_2 &= E(\vec{m}_1) * E(\vec{m}_2) = \vec{m}_1 * \vec{m}_2 + \\ &+ (\vec{r}_1 * \vec{r}_2) \cdot B_I + (\vec{g}_1 * \vec{g}_2) \cdot B_J^{pk} \end{aligned} \quad (12)$$

It follows that $\vec{c}_1 * \vec{c}_2$ preserve the format and are in the ciphertext space. Consider decrypting the sum of ciphertexts. For this, it is calculated $(\vec{c}_1 * \vec{c}_2) \bmod B_J^{pk} = \vec{m}_1 + \vec{m}_2 + (\vec{r}_1 * \vec{r}_2) \cdot B_I$ for ciphertexts. Moreover, the noise level should be less $\frac{B_J^{pk}}{2}$. Under this condition, the decryption is performed correctly, and the result is equal to $m_1 * m_2$, which to take modulo B_I noise.

Many other schemes have been developed based on this scheme since 2009. Most FHE schemes perform operations on the field of integers, but there is at least one scheme that performs operations on non-integers. The next section will discuss the most interesting schemes for integers and the scheme for non-integers.

3 Results

3.1 DGHV

Consider the DGHV scheme, which is an extension of Gentry's scheme [11]. Its feature is the use of polynomial keys, and this scheme is also designed with symmetric encryption. The scheme parameters are polynomials in the security parameter λ . η —is the key size p in bits. Key p generated in such a way that p is big η -bit prime number. A random large prime number is chosen for encryption. q and a small number $r \ll p$. Hence the message $m \in [0, 1]$ encrypted in ciphertext c by the following expression:

$$c = Enc(m) = m + 2r + pq, \quad (13)$$

Then decryption occurs according to the expression

$$m = Dec(c) = (c \bmod p) \bmod 2, \quad (14)$$

Now consider the homomorphism of addition and multiplication operations in this scheme, analyze the addition operation based on the following expression:

$$\begin{aligned} Dec(c_1 + c_2) &= Dec(Enc(m_1) + Enc(m_2)) = \\ &= \left((m_1 + m_2 + 2(r_1 + r_2) + p(q_1 + q_2)) \bmod p \right) \bmod 2 = \\ &= (m_1 + m_2 + 2(r_1 + r_2)) \bmod 2 = m_1 + m_2 \end{aligned} \quad (15)$$

Then the homomorphism of the operation of multiplication is defined as

$$\begin{aligned} Dec(c_1 \cdot c_2) &= Dec(Enc(m_1) \cdot Enc(m_2)) = \\ &= \left((m_1 m_2 + 2(m_1 r_2 + m_2 r_1 + 2r_1 r_2) + q_1 q_2 p) \bmod q \right) \bmod 2 \\ &= m_1 \cdot m_2. \end{aligned} \quad (16)$$

Moreover, the condition must be observed, such that the noise $2(m_1 r_2 + m_2 r_1 + 2r_1 r_2)$ must be less than p , otherwise, the decryption of the result will be performed incorrectly.

3.2 BFV

Based on the Brakerski Fan and Vercauteren scheme in 2012 [12], they developed a new encryption scheme based on learning with ring errors. The authors apply relinearization in the same way as in the BGV scheme [13], but their version is more efficient. They also simplified the bootstrapping process by switching modules.

Consider key generation in this scheme. Since for B - limited distribution χ around the ring R , used for learning purposes with errors in the ring, the secret key sk , it's possible choose how $s \leftarrow \chi$, and the public key is calculated as:

$$pk = ([-(a \cdot s + e)]_q, a) \quad (17)$$

where $e \leftarrow \chi$ и $a \leftarrow R_q$, where R_q – set of polynomials from \mathbb{Z}_q , where \mathbb{Z}_q – set of integers in the interval $(-\frac{q}{2}, \frac{q}{2}]$.

We establish that the message consists of integers integer $t > 1$ and belongs to the space R_t . Let's establish that $p_0 = pk[0]$, $p_1 = pk[1]$ and sample $u, e_1, e_2 \leftarrow \chi$, then to encrypt the message $m \in R_t$ you need to make the following calculations:

$$c = ([p_0 \cdot u + e_1 + \Delta \cdot m]_q, [p_1 \cdot u + e_2]_q) \quad (18)$$

where $\Delta = [\frac{q}{t}]$.

Then decryption can be performed by the following formula:

$$m = \left[\left[\frac{t \cdot |c(s)|_q}{q} \right] \right]_t = \left[\left[\frac{t \cdot [c_0 + c_1 \cdot s]_q}{q} \right] \right] \quad (19)$$

In [12], it was found that for successful decryption of a message, the condition must be met that the error must be less than or equal to the value $2 \cdot \delta_R \cdot B^2 + B$. B sequences they prove the lemma and show that, since $m \in R_t$, decryption works. Consider the homomorphism of operations in this encryption scheme based on addition and multiplication operations, establishing that the encrypted message is a polynomial s :

$$[c(s)]_q = [c_0 + c_1 \cdot s]_q = \Delta \cdot m + v \quad (20)$$

Then the homomorphism of the operation of addition can be defined as:

$$|c_1(s) + c_2(s)|_q = \Delta \cdot [m_1 + m_2]_t \quad (21)$$

Multiply two ciphertexts $ct_1(s)$ and $ct_2(s)$ it is possible by the following expression:

$$ct_1(s) \cdot ct_2(s) = c_0 + c_1 \cdot s + c_2 \cdot s^2. \quad (22)$$

Often, in different HE schemes, there is a problem that ciphertexts have different lengths, or the length becomes incorrect. To solve this problem in circuits, such as in BFV, a relinearization algorithm is used. This operation requires a special relinearization

key rlk , which is obtained by sampling $a \leftarrow R_{p,q}$ and $e \leftarrow \chi'(\chi' = X)$ and formula calculation (23).

$$rlk = \left(\left[-(a \cdot s + e) + p \cdot s^2 \right]_{p,q}, a \right) \tag{23}$$

To consider the next HE class, it is necessary to designate such a number system as the residue number system, which is used in some HE schemes, including the BFV described above [13].

3.3 HEAAN/CKKS

The problem of non-integers in homomorphic encryption was solved using approximate methods, i.e. Homomorphic Encryption for Arithmetic of Approximate Numbers (HEAAN) [14] scheme HE is better known by the names of its authors Cheon, Kim, Kim and Song (CKKS).

When developing CKKS, the authors decided to modify its RNS to speed up arithmetic operations. Thus, CKKS is a version of the homomorphic encryption scheme for approximate number arithmetic, originally proposed by Cheon, Kim, Kim, and Song, that provides approximate arithmetic over the field of complex numbers.

This scheme can be used for arithmetic over $C^{N/2}$. Plaintext space and ciphertext space share the same area

$$Z_Q[X]/(X^N + 1)$$

where N – power of two.

Batch encoding this scheme $C^{\frac{N}{2}} \leftrightarrow Z_Q[X]/(X^N + 1)$ maps an array of complex numbers to a polynomial with a property: $decode(encode(m_1) \otimes encode(m_2)) \approx m_1 \odot m_2$, where \otimes is a component multiplication, and \odot is a non-cyclic convolution.

The CKKS scheme supports standard recommended parameters chosen to provide 128-bit security for a uniform ternary private key. $s \in_u \{-1, 0, 1\}^N$, according to the group of homomorphic encryption standards.

CKKS encodes a complex number field using Lagrange polynomials [15]. The CKKS scheme has great potential for development, since its arithmetic is performed on the field of rational numbers with fixed precision. Let us consider homomorphic arithmetic operations in CKKS in more detail.

This scheme consists of five algorithms (*KeyGen*, *Enc*, *Dec*, *Add*, *Mult*) with constants *Bclean* and *Bmult*() for noise estimation. Then one can describe the scheme HE over the polynomial ring $R = Z[X]/(\Phi M(X))$.

- *KeyGen*(1^λ). Generation of secret value sk , public information pk for encryption and evaluation key evk .
- *Enc* _{p} $k(m)$. For a given polynomial $m \in \mathcal{R}$ encrypted text output $c \in \mathcal{R}_{II_C}^{\parallel}$. Encryption c for m will satisfy the condition $\langle c, sk \rangle = m + e(\backslash b \text{ mod } q_L)$ for some small e . Constant *Bclean* the encryption capability limiting point, i.e. the error polynomial of the new ciphertext satisfies $|e|_{\infty}^{can} \leq Bclean$ with high probability.

- $Dec_{sk}(c)$. For ciphertext c at the level of the output polynomial $m \leftarrow \langle c, sk \rangle \pmod{q_l}$ for the secret key sk .

Unlike most existing schemes, CKKS does not have a separate plaintext space from an inserted error. Output $m' = m + e$ decryption algorithm is slightly different from the original message m , but it can be considered an approximate value for approximate calculations when $|m|_{\infty}^{can}$ quite small compared to $|m|_{\infty}^{can}$. The intuition of approximate encryption has been partly used before, such as the switch key for homomorphic multiplication in [14–16].

Algorithms for homomorphic operations must satisfy the following properties.

$Add(c_1, c_2)$. For given ciphers m_1 and m_2 output encryption $m_1 + m_2$. The output ciphertext error is limited to the sum of two errors in the input ciphertexts.

$Mult_{evk}(c_1, c_2)$. For a pair of ciphertexts (c_1, c_2) output the ciphertext $c \in \mathcal{R}_{q_L}^k$ satisfying the condition $\langle c_{mult}, sk \rangle = \langle c_1, sk \rangle \cdot \langle c_2, sk \rangle + e_{mult} \pmod{q_l}$ for some polynomial $e_{mult} \in \mathcal{R}$ with $|e_{mult}|_{\infty}^{can} B_{mult}(l)$.

Thus, in this section, we have considered four FHE schemes, and then we will carry out a comparative analysis of these schemes.

4 Discussion

In this paper, the FHE, Gentry, DGHV, BFV, and CKKS schemes were considered. The first three schemes apply encryption over the field of integers, while CKKS over the field of fixed-point rational numbers. Gentry based his scheme on cryptographic lattices, which other schemes use. The DGHV scheme has some modifications over the Gentry scheme, such as polynomial keys and key switching, but it is also obsolete. There were several circuits between DGHV and BFV that introduced certain innovations, but the BFV circuit can be called the most advanced among integer circuits. Schemas BFV uses various mechanisms to improve and speed up the schema. For example, training with errors in the ring, new techniques in relinearization, bootstrapping. There have also been attempts to speed up the scheme using residue number systems (RNS).

The CKKS scheme, on the other hand, is the result of combining all the previous achievements of researchers. The authors were able to implement in the RNS system over the field of complex numbers using approximate methods, which makes it possible to apply the scheme over rational numbers with a fixed point. This makes it possible to apply this scheme in those applications that use non-integer numbers (for example, neural networks [24]). However, the scheme has some drawbacks, for example, a smaller number of multiplications before switching keys, more computationally complex rescaling and relinearization operations, as well as the amount of noise, and in this work [25], a comparison was made of the speed of the scheme, which showed that the BFV scheme is faster.

Separately, it is worth noting the implementation of schemes, at the moment there are many implementations of homomorphic encryption open source and freely distributed (Table 1).

Table 1. Summary table of FHE libraries

Library	Numbers with fixed precision	Integer numbers		Language	OS	
	CKKS	BFV	BGV		Windows	Linux
HElib [17]	Supports		Supports	C++		Supports
Microsoft SEAL [18]	Supports	Supports		C++	Supports	Supports
PALISADE [19]	Supports	Supports	Supports	C++	Supports	Supports
HEAAN [20]	Supports			C++		Supports
Lattigo [21]	Supports	Supports		Go	Supports	Supports
Pyfhel [22]	Supports	Supports		Python	Supports	Supports
SEAL-python [23]	Supports	Supports		Python	Supports	Supports

Analyzing this table, the Gentry and DGHV schemes were not implemented in free software. The most popular scheme in the implementation is CKKS, each library considered implements this scheme. Most of the schemes are implemented in C++ and the Linux operating system. Thus, a large area is open for researchers to study FHE implementations, as well as their modifications.

5 Conclusion

Thus, in this work, an analytical review of FHE schemes was carried out. HE was considered as a whole, which began its history in 1978, when the very concept of HE was introduced by Rivest et al. in the seminal work on RSA. Many SHE schemes have been developed during this time. During the study, it was found that since 2009, when the first FHE scheme was developed, many researchers became interested in developing their own schemes, most of which were based on Gentry's scheme, or continued his ideas.

However, one scheme generated more interest, namely the CKKS scheme. This scheme performs operations on the field of complex numbers, which allows it to process rational numbers with fixed precision. In addition, the circuit has the ability to process data in the form of polynomials and perform polynomial operations.

The analysis showed that the CKKS scheme has many effective mathematical tools, and researchers are also showing great interest in it. In terms of FHE software implementations, most of the free license implementations have a CKKS implementation.

Thus, CKKS is the most interesting for further research.

Acknowledgments. This work was carried out at the North Caucasus Center for Mathematical Research within agreement no. 075-2-2022-892 with the Ministry of Science and Higher Education of the Russian Federation. The study was financially supported by the Russian Foundation for Basic

Research within the framework of the scientific project No. 20-37-51004 “Effective intelligent data management system in edge, fog and cloud computing with adjustable fault tolerance and security” and Russian Federation President Grant MK-1203.2022.1.6.



References

1. Moysiadis, V., Sarigiannidis, P., Moscholios, I.: Towards distributed data management in fog computing. *Wirel. Commun. Mob. Comput.* (2018)
2. Tadapaneni, N.R.: Different Types of Cloud Service Models (2017)
3. Armknecht, F., et al.: A guide to fully homomorphic encryption. *Cryptology ePrint Archive* (2015)
4. Gentry, C.: A fully homomorphic encryption scheme. Stanford University (2009)
5. Rivest, R.L., Adleman, L., Dertouzos, M.L.: On data banks and privacy homomorphisms. *Found. Secure Comput.* **4**(11), 169–180 (1978)
6. Goldwasser, S., Micali, S.: Probabilistic encryption & how to play mental poker keeping secret all partial information. In: *Proceedings of the 14th Annual ACM Symposium on Theory of Computing*, pp. 365–377. ACM
7. Ogiwara, M.: On paddability of the quadratic residuosity problem. *IEICE Trans.* (1976–1990), **73**(2), 208–211 (1990)
8. ElGamal, T.: A public key cryptosystem and a signature scheme based on discrete logarithms. In: Blakley, G.R., Chaum, D. (eds.) *CRYPTO 1984*. LNCS, vol. 196, pp. 10–18. Springer, Heidelberg (1985). https://doi.org/10.1007/3-540-39568-7_2
9. Hariss, K., Chamoun, M., Samhat, A.E.: On DGHV and BGV fully homomorphic encryption schemes. In: *2017 1St Cyber Security in Networking Conference (CSNet)*, pp. 1–9. IEEE, October 2017
10. Fan, J., Vercauteren, F.: Somewhat practical fully homomorphic encryption. *Cryptology ePrint Archive* (2012)
11. Halevi, S., Polyakov, Y., Shoup, V.: An improved RNS variant of the BFV homomorphic encryption scheme. In: Matsui, M. (ed.) *CT-RSA 2019*. LNCS, vol. 11405, pp. 83–105. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-12612-4_5
12. Cheon, J.H., Kim, A., Kim, M., Song, Y.: Homomorphic encryption for arithmetic of approximate numbers. In: Takagi, T., Peyrin, T. (eds.) *ASIACRYPT 2017*. LNCS, vol. 10624, pp. 409–437. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-70694-8_15
13. Chan, W.C.C., Chyan, C.J., Srivastava, H.M.: The Lagrange polynomials in several variables. *Integral Transform. Spec. Funct.* **12**(2), 139–148 (2001)
14. Brakerski, Z.: Fully homomorphic encryption without modulus switching from classical GapSVP. In: Safavi-Naini, R., Canetti, R. (eds.) *CRYPTO 2012*. LNCS, vol. 7417, pp. 868–886. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-32009-5_50
15. Brakerski, Z., Gentry, C., Vaikuntanathan, V.: (Leveled) fully homomorphic encryption without bootstrapping. *ACM Trans. Comput. Theory (TOCT)* **6**(3), 1–36 (2014)
16. Cheon, J.H., Stehlé, D.: Fully homomorphic encryption over the integers revisited. In: Oswald, E., Fischlin, M. (eds.) *EUROCRYPT 2015*. LNCS, vol. 9056, pp. 513–536. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-662-46800-5_20
17. Helib. — Текст : электронный // GitHub : [сайт]. <https://github.com/homenc/HElib>
18. Microsoft SEAL. — Текст : электронный // GitHub : [сайт]. <https://github.com/microsoft/SEAL>
19. PALISADE HOMOMORPHIC ENCRYPTION SOFTWARE LIBRARY.—Текст : электронный. Palisade-Crypto : [сайт]. <https://palisade-crypto.org>

20. HEAAN. — Текст : электронный // GitHub : [сайт]. <https://github.com/snucrypto/HEAAN>
21. Lattigo. — Текст : электронный // GitHub : [сайт]. <https://github.com/tuneinsight/lattigo>
22. Pyfhel. — Текст : электронный // GitHub : [сайт]. <https://github.com/ibarrond/Pyfhel>
23. SEAL-python. — Текст : электронный // GitHub : [сайт]. <https://github.com/Huelse/SEAL-Python>
24. Orlandi, C., Piva, A., Barni, M.: Oblivious neural network computing via homomorphic encryption. *EURASIP J. Inf. Secur.* **2007**, 1–11 (2007)
25. Golimblevskaia, E., Shiriaev, E., Kucherov, N.: Survey software implementations of homomorphic encryption methods. In: Radionov, A.A., Gasiyarov, V.R. (eds.) *RusAutoCon 2020. LNEE*, vol. 729, pp. 601–613. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-71119-1_59



Model of Error Correction Device in RNS-FRNN

Egor Shiriaev^{1,3}  and Viktor Kuchukov^{2,3} 

¹ North-Caucasus Federal University, Stavropol, Russia
eshiriaev@ncfu.ru

² North Caucasus Center for Mathematical Research NCFU, Stavropol, Russia

³ Sirius University of Science and Technology, Sochi, Russian Federation

Abstract. The spread of cloud technologies has made it possible to increase the efficiency and flexibility of computer networks, as well as information storage. However, the decentralized structure and the large geographical distance of data processing centers imposes an additional burden on the infrastructure of such technologies. Both in big data streams and in the processing of information on nodes, an error often occurs, which reduces the reliability of the structure. In this work, we have developed a model of an error correction device in the residue number system using finite ring neural networks. The result of this work is to increase the efficiency of error handling, which in turn increases the reliability of the system.

Keywords: Cloud technologies · Residue number system · Finite Ring Neural Network · Syndrome method · Base expansion

1 Introduction

Cloud technologies (CT), due to such properties as flexibility, transparency, efficiency, are one of the most promising areas in the field of IT technologies [1]. CT uses various technologies for working with information, such as parallel and distributed, as well as grid computing [2]. Considering that CTs are often used as a service for providing access to the cloud structure, they are usually divided into three categories:

1. Software as a Service (SaaS).
2. Platform as a service (PaaS).
3. Infrastructure as a Service (IaaS).

In the case of SaaS, the user is provided with some software that is located remotely from him. It cannot control the cloud systems (infrastructure, operating system, etc.), but has full control over the application. Software can also be a virtual machine with an operating system (OS) installed on it [3].

PaaS is essentially a platform for deploying applications with support for cloud solutions [4].

IaaS - in this case, the entire cloud infrastructure is provided, the user can use it for software development, data storage, information processing, etc. In this case, the user

has full control over the dedicated cloud segment. This service is the most expensive; it is used by large firms and corporations to deploy their infrastructure in additional structures that they do not have [5].

Thus, we can say that now CT is not only popular as a new technology but is firmly established as part of the global network, production, and IT in general. The most common characteristic that describes other cloud services can be called its cost. The cost is calculated from the cost of its creation and support of basic characteristics, such as security, reliability, etc. The purpose of this work is to increase the reliability of cloud services. Reliability in CT - determines the ability to save information in the correct form during its storage and transmission. Both the hardware component of the cloud solution and the software component are responsible for this. Very often in computer technologies there are collisions, interference, extraneous noise, surges, voltage drops, etc., which can lead to information distortion. Distortions of information lead to the need for distortion control and error correction, which in turn encourages cloud service providers to apply control methods.

There are many methods for error control. The most traditional approach is to create backups, at least three. However, this imposes significant costs and significantly increases the physical size of the storage. However, this allows you to correct the error after the fact of its detection, which is not always effective, and if the backup copy is lost for some reason, then the information can be lost forever. To combat this, so-called corrective codes are used. Correction codes [6] make it possible to control the integrity of information in real time. It is also worth noting that such codes require additional redundancy, which, like backup, imposes multiple costs.

There are many works, some of which will be discussed below, in which the residue number system (RNS) is used as a method of error control and correction [7]. RNS is a non-positional number system that is based on the modular properties of numbers. RNS is a consequence of the Chinese Remainder Theorem (CRT) [8]. This is also related to such a mathematical term as the ring of residues. In RNS any number X can be represented as a set of numbers x_1, x_2, \dots, x_n obtained by dividing by a set of coprime modules p_1, p_2, \dots, p_n , where n - natural number of modules, i.e. $x_i = X \bmod p_i$, where $i \in \{1, 2, \dots, n\}$. CRT guarantees the correct display of the number in RNS when $X < P$, where $P = \prod_{i=1}^n p_i$. RNS has many useful properties. Since the RNS residuals are independent, it is possible to perform calculations on them in parallel and quickly. The operations in RNS shown below:

$$C = A * B = (a_1 * b_1) \bmod p_1, (a_2 * b_2) \bmod p_2, \dots, (a_n * b_n) \bmod p_n \quad (1)$$

where $* \in \{+, \times\}$.

In addition, RNS has self-correcting properties, which will be discussed below.

Let us consider several works devoted to the problem of error correction. If we consider the overall picture, then most scientific groups, when working with error correction in terms of information transmission, even though transmission networks on the Internet have their own methods of protecting against such collisions, such works are related to our subject, since the cloud infrastructure is quite complex and multifaceted.

For example, in his work, D. P. Hart [9] conducts research related to the reduction of subpixel errors. The author's method is to eliminate false vectors using particle

image velocimetry. This is achieved by creating correlation tables of neighboring regions and element-by-element comparison of the results. Thanks to these manipulations, it is possible to eliminate electronic and optical noise, which is a good result for images. However, this work is aimed at correcting distorted image pixels, when information can be distorted more insignificantly, for example, a few bits, which in some situations can be critical.

Error correction codes are also used when reading various information, such as DNA. DNA must be read with high accuracy. Jan Schröder et al. [10] propose a new method for error correction when reading short reads. The error correction algorithm is based on the use of a common suffix for use in basic data structures. The authors claim that the accuracy of error correction for real data is more than 88%, which is superior to previously published approaches.

RNS also has a rich history of its corrective properties. For example, in their work, S. Pontarelli et al. [11] propose the concept of an FIR filter built on RNS for the purpose of error handling. Error correction occurs due to the syndrome method. In general, this paper shows an interesting implementation of the FIR filter based on the error syndrome method in RNS. The syndrome method in RNS is quite popular along with other methods such as the projection method, but the syndrome method has a higher efficiency.

In this paper, we will demonstrate the scheme of the RNS error syndrome method using finite ring neural networks, which increases the efficiency of the method. This device will improve the reliability of cloud structures.

The work consists of the following: Sect. 2 investigates the detection and correction of errors in RNS; Sect. 3 presents the development of a model for the RNS-FRNN error correction device; Sect. 4 provides a discussion of the applicability of this device and further research; Sect. 5 presents the conclusions obtained in the course of this work.

2 Materials and Methods

2.1 RNS and Error Detection

In the previous section, we considered RNS. In this section, we will show the relationship between the generalized weighted number system (WNS) and RNS X based on the following system of comparisons:

$$\begin{aligned} X &\equiv x_1 \text{ mod } p_1 \\ X &\equiv x_2 \text{ mod } p_2 \\ &\dots \\ X &\equiv x_n \text{ mod } p_n \end{aligned}$$

The system shows that any remainder is identical to the number modulo assigned to it. You can also restore the number in WNS based on the following formula:

$$X = \left| \sum_{i=1}^n x_i B_i \right|_P, \quad (2)$$

where $B_i = \frac{P}{p_i} P_i^{-1}$ – orthogonal basis of the module RNS, P_i^{-1} – multiplicative inverse of a number. This method is called the CRT translation method or the Garner method.

Then if the number $X < P$ it will successfully restore to WNS. This property also allows you to control errors in the RNS, since often the distortion of one of the remainders of the set leads to an overflow of the range, i.e. $X > P$. However, this method of determining the error cannot be called effective. The operation of converting a number from RNS to WNS is complex and requires a large amount of computing resources, even when using other translation methods.

In this paper, we will consider another method of error detection called nullification. The method was developed by I. Ya. Akushsky and D. Yu. Yuditsky in [12]. This method is simple and requires only the compilation of a table based on a set of modules. Since the set of modules is static for all data, it does not need to be recalculated. The table is calculated using the following expression:

$$M_i^j = \{t_1^1, t_2^1, \dots, t_n^1, t_1^2, t_2^2, \dots, t_n^m\}, \tag{3}$$

where $t_i^j \in 1, 2, \dots, p_i - 1, i \in \{0, n\}, j \in \{0, p_i - 1\}$.

The process of nullification itself consists in successively reducing the number to zero. If the number X turns to 0 in $n - 1$ operations, it is considered that the number is correct:

$$X_{S_i} = x - M_i^j = \{0, S_2, S_3, \dots, S_{n+1}\} \tag{4}$$

This method of error detection avoids the operation of converting a number from RNS to WNS.

2.2 RNS and Error Correction

If we consider traditional error correction codes, they all have high redundancy, which is necessary to control the main code [6]. This is the main disadvantage of such codes. Correction codes in RNS also require redundancy, but much less than traditional ones. To correct a single error, two control bases are sufficient. The control base is usually taken more than the working one. There is a strict requirement for control residues - they must be absolutely accurate. Such absolute accuracy allows you to correct errors by various methods.

Consider the syndrome method [13].

Let's say in the number $X = \{x_1, x_2, \dots, x_n, x_{n+1}, x_{n+2}\}$ error detected. Then, using the base system extension methods [14], we obtain new, distorted, control residuals x'_{n+1} and x'_{n+2} . Having accurate and distorted control residuals, we get the error syndrome according to the following formula:

$$\phi_i = x_{n+i} - x'_{n+i} \text{ mod } p_{n+1} \tag{5}$$

The syndromes are mapped to an error table. The error table is compiled by enumeration of all numbers from 0 to $P - 1$ on a set of modules. For more efficient tabulation, control residues (which are syndromes) are compiled in pairs, since usually 2 sets of

Table 1. Error table

$\phi^{(1)}, \phi^{(2)}$	Error	$\phi^{(1)}, \phi^{(2)}$	Error	$\phi^{(1)}, \phi^{(2)}$	Error
(0, 0)	(0, 0, 0)	(3, 10)(1, 2)	(0, 1, 0)	(6, 9)(4, 1)	(0, 2, 0)
(1, 1)(6, 4)	(1, 1, 1)	(4, 0)(2, 3)	(1, 2, 1)	(0, 10)(5, 2)	(1, 0, 1)
(2, 2)(0, 5)	(0, 2, 2)	(5, 1)(3, 4)	(0, 0, 2)	(1, 0)(6, 3)	(0, 1, 2)
(3, 3)(1, 6)	(1, 0, 3)	(6, 2)(4, 5)	(1, 1, 3)	(2, 1)(0, 4)	(1, 2, 3)
(4, 4)(2, 7)	(0, 1, 4)	(0, 3)(5, 6)	(0, 2, 4)	(3, 2)(1, 5)	(0, 0, 4)
(5, 5)(3, 8)	(1, 2, 0)	(1, 4)(6, 7)	(1, 0, 0)	(4, 3)(2, 6)	(1, 1, 0)
(6, 6)(4, 9)	(0, 0, 1)	(2, 5)(0, 8)	(0, 1, 1)	(5, 4)(3, 7)	(0, 2, 1)
(0, 7)(5, 10)	(1, 1, 2)	(3, 6)(1, 9)	(1, 2, 2)	(6, 5)(4, 8)	(1, 0, 2)
(1, 8)(6, 0)	(0, 2, 3)	(4, 7)(2, 10)	(0, 0, 3)	(0, 6)(5, 9)	(0, 1, 3)
(2, 9)(0, 1)	(1, 0, 4)	(5, 8)(3, 0)	(1, 1, 4)	(1, 7)(6, 10)	(1, 2, 4)

control often correspond to a working base. For example, a table of the first 20 numbers for a set of working bases (2,3,5) and control (7,11) has the following form:

Further, after receiving the syndrome and comparing it with the error, it is necessary to add it to the number X without control bases, i.e.

$$X^{correct} = X + X' = (x_1, x_2, \dots, x_n, 0, 0) = (x'_1, x'_2, \dots, x'_n, 0, 0) \tag{6}$$

Thus, in RNS it is possible to quickly fix the error [14].

2.3 Finite Ring Neural Network

The finite ring neural network [15] (FRNN) combines the properties of a biological neuron and RNS [16]. In fact, if we imagine that the number of synapses is equal to the number of bases in RNS, then the representation of RNS in neural networks is natural. The self-correcting properties of RNS, or rather the fault tolerance of RNS, are also common with neural networks, due to the multiplicity of inputs and outputs of individual neurons. FRNN got its name from the fact that RNS has arithmetic with finite computing structures, rings, which are used in the implementation of modular operations.

The FRNN architecture is a three-level hierarchy: parameter mapping; bit calculation display; displaying the operation of a finite ring. Variable arithmetic is modular addition and multiplication, as well as combinations of these operations (obtaining the remainder from division is an exception and is usually performed on the input layer).

Thus, FRNN is a typical artificial neural network built on RNS with the possibility of repetition (closing the neuron to itself) until the desired result is obtained. Such a model is quite efficient and allows implementing almost any operation in RNS on FRNN with subsequent construction of a neurocomputer.

In our work, with the help of FRNN, we implement 3 operations: error checking, extension of the module base, and error correction.

3 Results

The developed model has 4 connected blocks. Input block. Error Detection Block. Error correction block and output block (see, Fig. 1).

FRNN is implemented on three blocks - input, control, and error correction. Let us look at each of these blocks. The input block is a virtual block necessary for the general representation of the model, in fact, in a real device, the input numbers are already received in the RNS, as in the entire system. The input receives information in the usual positional representation and the WNS-RNS translation is performed (see, Fig. 2).

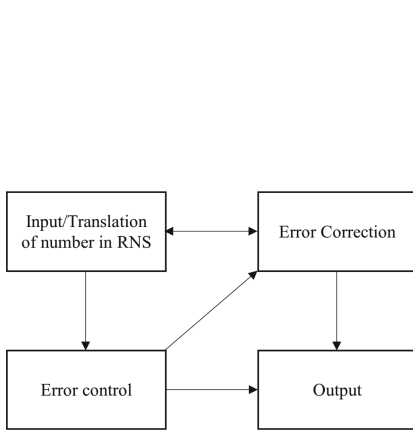


Fig. 1. Model

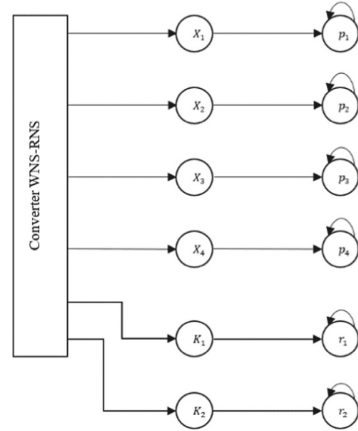


Fig. 2. WNS-RNS Number Translation Unit

After that, the obtained values are transmitted further to the error control and correction unit. The error control unit (see, Fig. 3 (a)) performs the zeroing. The output is signaled, “0” - no error “1” - error present. If there is no error, then the information is transmitted further, otherwise the error correction block starts its work (see, Fig. 3 (b)). Error correction works according to the method presented in Sect. 2.2. The base extension method was presented in [17]. The corrected number is sent to the output.

Thus, the general model has the following form (see, Fig. 4). Where highlighted - 1) error control block; 2) error correction block.

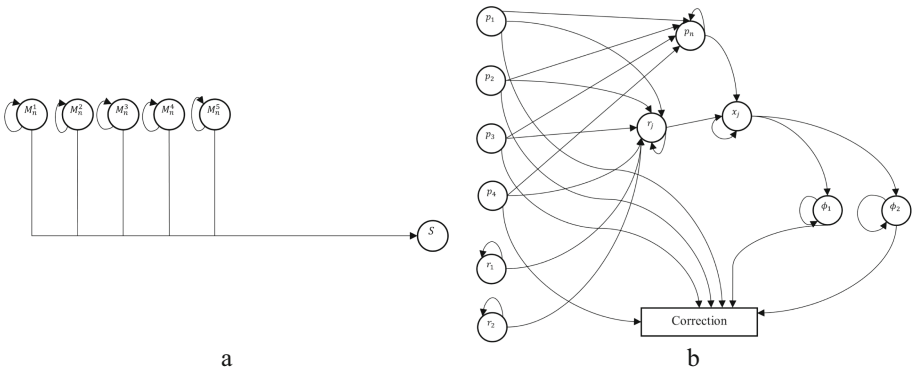


Fig. 3. Blocks of control and error correction

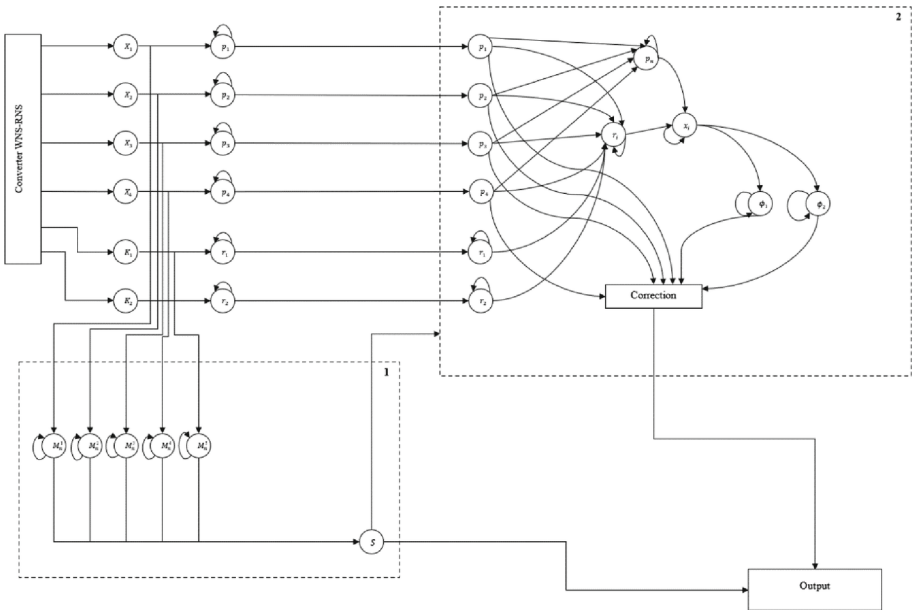


Fig. 4. Error correction device model in RNS-FRNN

4 Discussion

The developed device model allows you to effectively control and correct errors in the information presented in the RNS. In this paper, we use the result presented earlier in [17] and demonstrate a method for expanding the base of the number of RNS using FRNN. It increased the efficiency of the operation. In addition, it is also worth noting that, in general, working with FRNN is possible thanks to such works as [18–20], which reflected the construction of neural networks with RNS. This model differs from typical works related to the syndrome method by the method of error detection, nulling, and by the

base extension method, since the computationally complex RNS-WNS-RNS translation is usually used.

5 Conclusion

In this work, we have developed a model of the RNS-FRNN error correction device. An analysis of existing work on correcting errors in the field of IT was carried out. It has been established that the problem of error correction is present in various fields of science where IT is applied, such as DNA reading. RNS is also used for error correction due to its self-correcting properties, which is why RNS is gaining popularity in this matter. The developed model implies the use of such techniques in RNS as zeroing, number base expansion in RNS and the syndrome method, these techniques are implemented in FRNN. The developed model will effectively process the information presented in the RNS to find errors, the model can be integrated into cloud technologies that use RNS.

Acknowledgments. This work was carried out at the North Caucasus Center for Mathematical Research within agreement no. 075-02-2022-892 with the Ministry of Science and Higher Education of the Russian Federation. The reported study was funded by RFBR, Sirius University of Science and Technology, JSC Russian Railways and Educational Fund “Talent and success”, project number 20-37-51004 “Efficient intelligent data management system for edge, fog, and cloud computing with adjustable fault tolerance and security”, and Russian Federation President Grant SP-3186.2022.5.

References

1. Chervyakov, N., Babenko, M., Deryabin, M., Garianina, A.: Development of information security’s theoretical aspects in cloud technology with the use of threshold structures. In: 2014 International Conference on Engineering and Telecommunication, pp. 38–42. IEEE, November 2014
2. Tropp, J.A.: Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inf. Theory* **50**(10), 2231–2242 (2004)
3. Cusumano, M.: Cloud computing and SaaS as new computing platforms. *Commun. ACM* **53**(4), 27–29 (2010)
4. Pahl, C.: Containerization and the PaaS cloud. *IEEE Cloud Comput.* **2**(3), 24–31 (2015)
5. Hay, B., Nance, K., Bishop, M.: Storm clouds rising: security challenges for IaaS cloud computing. In: 2011 44th Hawaii International Conference on System Sciences, pp. 1–7. IEEE, January 2011
6. Sanna, M., Izquierdo, E.: A survey of linear network coding and network error correction code constructions and algorithms. *Int. J. Digit. Multimedia Broadcast.* (2011)
7. Garner, H.L.: The residue number system. Papers presented at the the Western Joint Computer Conference 3–5 March 1959, pp. 146–153, March 1959
8. Pei, D., Salomaa, A., Ding, C.: Chinese Remainder Theorem: Applications in Computing, Coding, Cryptography. World Scientific, Singapore (1996)
9. Hart, D.P.: PIV error correction. *Experiments Fluids* **29**(1), 13–22 (2000)
10. Schröder, J., Schröder, H., Puglisi, S.J., Sinha, R., Schmidt, B.: SHREC: a short-read error correction method. *Bioinformatics* **25**(17), 2157–2163 (2009)

11. Pontarelli, S., Cardarilli, G.C., Re, M., Salsano, A.: A novel error detection and correction technique for RNS based FIR filters. In: 2008 IEEE International Symposium on Defect and Fault Tolerance of VLSI Systems, pp. 436–444. IEEE, October 2008
12. Akushsky, I.Ya., Yuditsky, D.I.: Machine arithmetic in residual classes. *Owls, Radio*
13. Shiryaev, E., et al.: Performance impact of error correction codes in RNS with returning methods and base extension. In: 2021 International Conference Engineering and Telecommunication (En&T), pp. 1–5. IEEE (2021)
14. Shenoy, A.P., Kumaresan, R.: Fast base extension using a redundant modulus in RNS. *IEEE Trans. Comput.* **38**(2), 292–297 (1989)
15. Wang, S.C.: Artificial Neural Network. In: *Interdisciplinary Computing in Java Programming*, pp. 81–100. Springer, Boston (2003)
16. Chervyakov, N.I., Galkina, V.A., Strekalov, Yu.A., Lavrinenko, S.V.: Neural network of the finite ring (2006)
17. Babenko, M.G., et al.: Neural network method for base extension in residue number system. In: *ICCS-DE*, pp. 9–22 (2020)
18. Chervyakov, N.I., Spelnikov, A.B., Mezentshev, O.S.: Neural network of a finite feed-forward ring for operations in elliptic curves. *Neurocomp. Dev. Appl.* (1–2), 28–34 (2008)
19. Timoshenko, L.I.: Implementation of modular operations in the ring of polynomials using neural networks. *Int. J. Appl. Basic Res.* (1-1), 22–24 (2015)
20. Dolgachev, A.A., Irkhin, V.P., Andreev, R.N., Melnik, V.A.: Neural network of the finite ring (2019)



Review of Modern Technologies of Computer Vision

Ekaterina Bezuglova^{2,3}(✉) , Andrey Gladkov^{1,3} , and Georgy Valuev² 

¹ North-Caucasus Federal University, Stavropol, Russia

² North Caucasus Center for Mathematical Research NCFU, Stavropol, Russia
bezuglovakaterina@mail.ru

³ Sirius University of Science and Technology, Sochi, Russian Federation

Abstract. Today, the use of artificial intelligence technologies is becoming more and more popular. Scientific and technological progress contributes to increasing the power of hardware, as well as obtaining effective methods for implementing methods such as machine learning, neural networks, and deep learning. This created the possibility of creating effective methods for recognizing images and video data, which is what computer vision is. At the time of 2022, a huge number of methods, technologies, and techniques for using computer vision were received, in this paper a study was conducted on the use of computer vision in 2022. Results were obtained on the decrease in the popularity of computer vision in the scientific community, its introduction into industry, medicine, zoology and human social life, the most popular method of computer vision is the ResNet neural network model.

Keywords: Computer Vision · Artificial intelligence · Convolutional Neural Networks · ResNet · OpenCV · YOLO

1 Introduction

Computer vision (CV) is a field of Artificial Intelligence (AI), including machine learning (ML) [1]. CV refers to many different systems, such as automated process control, video surveillance, information organization, object or environment modeling, human-machine interaction, augmented reality, etc. [2]. Every year this list is updated with new systems. In addition, the scope of CV application is also expanding, such as one of the most popular areas - medicine (detection of oncology). Even though, in fact, the history of the application of computer vision begins in the 1970s, CV is considered a new area of research. Various factors contribute to its development. The main one is the scientific and technological progress that has made it possible to create high-performance processors and store huge amounts of information, which is an important factor for processing data such as images. Also, development factors are advances in related areas - AI, ML, and neural networks - new methods to increase the speed and accuracy of data processing allow the development of effective CV methods.

The purpose of this work is to conduct a study of the current state of CV, the effectiveness and applicability of its methods. We will explore the application of CV in various areas that are not directly related to IT, which will allow us to consider CV methods in more detail.

CV is a technology that obtains information from images or video data. Video data can be obtained in various ways: video sequence, image frames, or 3D object models. Also associated with CV are artificial intelligence methods such as pattern recognition and learning methods. In other words, CV interacts closely with image processing and machine vision, and uses signal processing. Thus, CV is a broad area of scientific interests related to various areas of IT sciences, as well as others, such as biology, medicine, etc.

The work consists of the following: in Sect. 2, research is carried out on scientific papers that are devoted to CV technology at the time of mid-2022; Sect. 3 provides a classification of the models considered in the study; Sect. 4 provides a reasoning on the received studies, as well as the obtained statistics on the relevance of CV technology, as well as interest in certain CV solutions; Sect. 5 presents the conclusions obtained in the course of this work.

2 Materials and Methods

This section will present the works that were selected for analysis. A feature of this study is that the works were selected from various areas of scientific interest and for various applied problems.

Obviously, the main application of CV is image processing. CV methods can help where processing of large amounts of images and video data is needed. For example, Yujie Lei et al. in their work [3] conduct a study on the use of CV to detect the primate species Slow Loris. The need for effective methods of detection and recognition is since this species is endangered, and the timely detection of primates will increase the protection of this species from extinction. In this work, a dataset of more than 50 TB in size is used, which contains video data on the activity of the studied primates from April 2017 to June 2018. This dataset is used for training and validating CV. It contains both images of single individuals and groups. In addition, this paper uses the YOLO framework developed by Redmon et al. to solve the problem of deep detection speed, which gave a new detection method [4]. In this work, the YOLOv5 version [5] is used. In general, the framework is a multilayer convolutional neural network [6] for image recognition.

Consider the initial stages of YOLOv5 work:

1. Filling data - before the convolution operation, the boundaries of the original matrix are filled with indents. Some values are padded at the matrix boundary to increase the size of the matrix, usually 0 is chosen.
2. The step of the convolution is as follows, moving the kernel from the upper left corner diagonally, one step, one row-column. Thus, the number of rows and columns in each slide is called a pitch. To avoid loss of information during the reconciliation, data padding is applied, while the size of the work step is set in such a way that the compression of part of the information has the effect of reducing the information at the input.

- Establishing a relationship between the number of channels and filters: The number of output layers of a channel is only related to the number of channels in the current filter.

YOLOv5 applies 2D image convolution. That is, the output image is obtained by the formula:

$$\sigma = \left\lfloor \frac{n + 2p - k}{s} \right\rfloor + 1$$

where k – original image size ($k \times k$), n – output image size, s is the step size, p – size of the convolution kernel. The general block diagram of YOLOv5 is used in the work (Fig. 1):

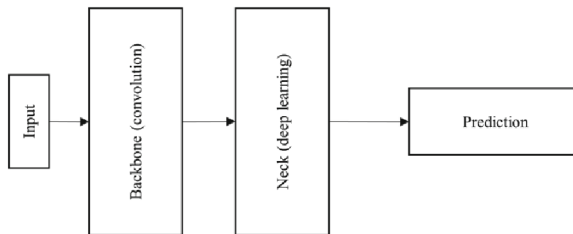


Fig. 1. General scheme of YOLOv5

According to the authors, their model built based on the YOLOv5 framework has high accuracy (about 95%) and a small model size (16.4 MB). That allows you to take this development for image processing and recognition of dynamic objects (such as primates). Consider other CV models.

During the CoronaVirus Disease 2019 (COVID-19) pandemic [7], a lot of research and development was carried out around the world using new IT CVs, and Fabricio Crespo et al. in their work [8] conducted a study of CV as a model for identifying medical mask on the face. This study plays an important role because Proper wearing of masks has become a problem in most countries affected by the COVID-19 pandemic. The main problem, according to the authors, is the incorrect wearing of masks, which is a difficult task to detect. In their paper, the authors propose a solution by exploring multiple datasets and CV methods. The authors use various datasets for their model, such as: Winder face [9], Face Mask Label Dataset (FMLD) [10], and MAFA [11]. Winder face differs from the others in that it does not feature masked faces.

In this paper, the authors build a CV model by detecting and classifying faces. For detection, RetinaFace [12] is used - a deep learning model for CV, which is a single-shot multi-level face localization in images, i.e., this is a face detection technique using an end-to-end detector. RetinaFace can perform several face detections tasks: face detection; two-dimensional alignment of the face; building a three-dimensional model of the face. RetinaFace is based on the ResNet architecture [13] using a Fully Pyramidal Network (FPN) to obtain a detailed image function. Classification occurs at the expense of ResNetSt. ResNetSt is a modification of ResNet by introducing a feature map with

multi-threaded access to improve the quality of recognition. In addition, ResNetSt has distributed computing units divided into function groups. This modification has better transfer learning results [14]. The authors use Stochastic Gradient Descent (SGD) as an optimizer [15]. The results of the study showed that the use of ResNetSt in conjunction with SGD makes it possible to recognize a mask on a face with an accuracy of more than 96%.

Consider a library for working with CV. Google engineers have developed a new computer vision library called SCENIC [16]. SCENIC is an open-source all-in-one library that allows you to process various inputs in a pipelined way and distribute training according to the methods of interest to the user, for example, such as ViT [17], DETR [18], CLIP [19], MLP mixer. [20], T5 [21], and BERT [22].

According to the authors, SCENIC is already being used by Google, in such projects as ViViT [23], OmniNet [24] and others, studies on the scaling behavior [25] and efficiency [26] of various models.

The library also offers project management, i.e. certain model settings can be distributed over several projects or control several models at the same time. The library is implemented within the Python programming language. If we consider the content of the library, then it offers various ML methods for use in CV, in addition, some of the solutions used in this library are used for processing text information, which can also be applied in CV for text recognition on images or video data.

In the work of Jue Su et al. [27], studies are carried out on the study of the physico-mechanical properties of flexible biomass particles using CV. The authors are modernizing the Particle Size Distribution (PSD) method [28] by applying CV to increase the efficiency and speed of material processing, since the traditional method is too time consuming. For their research, the authors use: a special stand for attaching the material and a camera. Further, the resulting image is processed using the CV module included in the Matlab software product [29]. In general, the method proposed by the authors makes it possible to establish with a certain accuracy the particle sizes and their main characteristics based on images using CV.

Stéfano Frizzo Stefenon et al. [30] classify insulators using a neural network based on CV. CV is used in this work to detect contamination on insulators. The model itself is based on solutions such as Keras and TensorFlow for Python. CV in this work is used from the OpenCV library [31]. Thus, the authors developed a deep learning CNN for CV application. Based on the obtained CV data, the neural network classifies the images of insulators according to the degree of contamination. According to the authors, their model has an accuracy of about 97.50%

We also consider the work of David A. Wood et al. [32], which is devoted to deep learning for labeling head MRI datasets for computer vision applications. In their work, the authors raise the issue of preparing datasets for CV deep learning, since manual preparation takes a lot of time. The authors substantiate the work carried out by the successful results in compiling datasets of computed tomography of the head using the natural language processing method, which made it possible to obtain data with satisfactory accuracy. The authors use a modification of BERT, which is specially designed for medical research - BioBERT [33]. The authors also modified it by adding a custom

module for processing medical reports. Thus, the authors built a neural network that classifies neuro-medical MRI reports of the head and puts binary labels (normal/abnormal) using CV on the MRI result.

3 Results

During the study, various methods for applying CV to certain libraries and models were considered. Thus, it is possible to classify methods (see, Table 1).

Table 1. Classification of methods

No	Library	Open-source	Language	Methods	CPU	GPU	Narrow focus
1	YOLO	+	Python	RCNN, fast RCNN and faster RCNN, Deep SORT	+	+	Object recognition
2	RetinaFace	+	Python	Deepface[], arcfac[],	+	+	Face recognition
3	ResNet	+	Python	CNN	+	+	Object recognition
4	OpenCV	+	C++	Deep NN (DNN),	+	+	Library for CV
5	Mathlab CV	-	C/C++	SURF, Viola–Jones, HOG, KLT	+	+	CV
6	BERT	+	Python	Transformer (machine learning model)	+	-	Natural language processing, learning CV models
7	BioBERT	+	Python	Transformer (machine learning model)	+	-	biomedical BERT
8	ViT	+	Python	Token-to-Token ViT [], CaiT []	+	-	CV
9	OmniNet	+	Python	VQA, HMDB, Captioning, PENN	+	+	Multi-modal multi-task learning

(continued)

Table 1. (continued)

No	Library	Open-source	Language	Methods	CPU	GPU	Narrow focus
10	SCRENIC	+	Python	OmniNet, ViT, CLIP, DETR, BERT, etc	+	+	Library for CV
11	CLIP	+	Python	CNN	+	+	NN training for CV (video and text)
12	DETR	+	Python	CNN	+	+	CV

Thus, by analyzing the table, we can include the following. Most of the studied methods, models, libraries use the python programming language. This is due to the presence of a rich toolkit, the language it has for working with AI results. Also, most open-source projects other than CV is a Matlab module. Exceptional models have different features in the CV. Network training, data sets, neural networks for CV work, as well as ready-made libraries. Such a rich open-source toolkit for 2022 allows you to both conduct research using CV and introduce new technologies, methodologies and prototypes based on CV.

Discovered cases of discovery of the most popular publications found in ResNet. ResNet is a neural network used for infection. The network coverage gained the greatest popularity due to simple and high consumption, which entails many extensions (up to 100). It can also be associated with oncological diseases, for example, with the development of malignant neoplasms of the skin.

4 Discussion

Thus, in this work, a study was conducted of scientific publications on the topic of CV published in 2022. Those papers were selected that presented practical research and results and were also published in leading scientific journals. If we consider the direction of publications in 2022, then many publications were associated with the introduction of CV in production, agriculture, etc. (see examples [34–36]).

This situation allows us to say that CV is not a hard-to-reach theoretical innovation in the sciences of artificial intelligence, but a proven technology that is used in everyday life, and interest in it does not subside. Consider a chart that displays the number of publications by the keyword “computer vision methods”. Statistics obtained using google scholar is shown in Fig. 2.

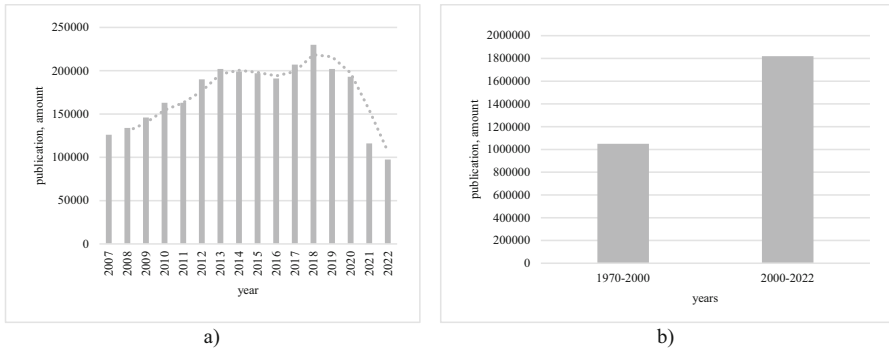


Fig. 2. Statistics on the number of publications on the topic “computer vision methods”: a) over the past 15 years b) in the 20th and 21st centuries

Analyzing the illustration (Fig. 2), as already mentioned in Sect. 1, CV began its development in the 70s of the last centuries, thus analyzing the number of publications for 30 years of the end of the 20th century and 22 years of the 21st century, interest in CV increased significantly, namely 1.733 times. The statistics were also obtained for the last 15 years. Analyzing it, you can see that the peak of interest in CV falls on 2018, this is explained by the fact that in the mid-2010s, the popularity of AI technologies arose, both in the scientific community and among society as a whole, and active development of equipment for working with it began., so it became popular to conduct research on the GPU instead of the CPU (in the works presented in the study, the vast majority of simulations were carried out on the GPU), but since 2018, interest has declined and continues to fall. This may be since the main directions for use were determined for CV, and the main methods of application were found. During the study, it was found that, depending on the application, certain CV models are used, such as OpenCV, ResNet, etc., investigator, now most research is not aimed at developing fundamental approaches, but at creating applied solutions, which fair for proven technologies that are being introduced into human activities.

We will also consider such statistics on the methods and technologies considered in this paper (Fig. 3).

Analyzing the obtained statistics, we can draw the following conclusions. Over the past 15 years, the largest number of publications have solutions based on Matlab CV. This is since Matlab was one of the first to introduce the CV module into its mathematical apparatus, given the ease of handling and distribution (most large research centers have licenses for this software) Matlab in this area. However, if you look at the illustration Fig. 3 b), you can see that the popularity of Matlab has declined, while open-source projects, on the contrary, have begun to grow. This is due to the availability of solutions and the development of the use of Python as an AI tool, as well as technological progress (it has become possible to process data on the GPU), so the ResNet mentioned in Sect. 3 for 2022 is the most popular due to its ease of use and implementation. ResNet is usually chosen by research groups whose main activity is not related to the IT field because of the ease of setting up and implementing CV models. However, representatives of narrow

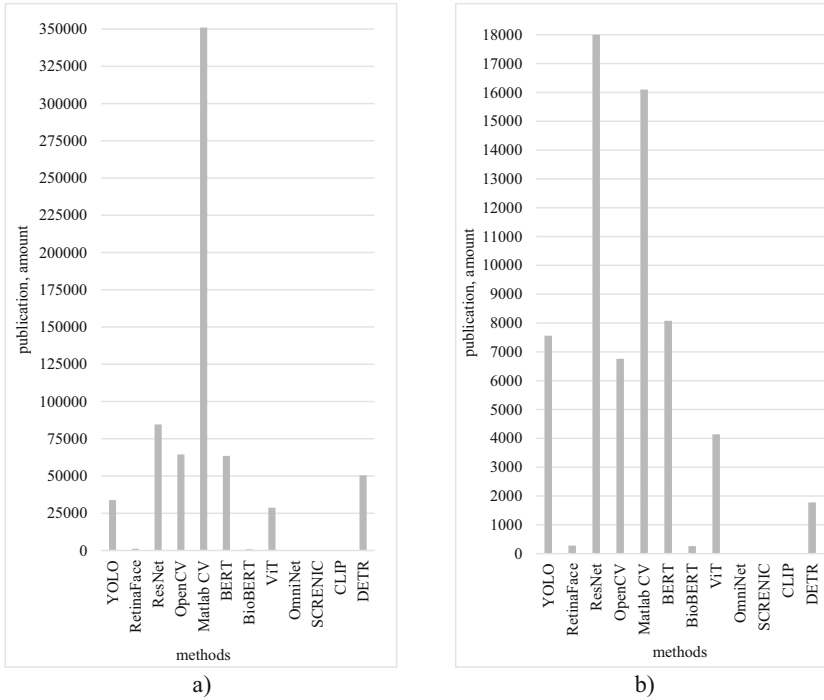


Fig. 3. Statistics on the number of publications on the topic “computer vision methods”: a) for the last 15 years b) for 2022

technical fields, such as electrical engineering, bioengineering, etc. refer to traditional Matlab.

In this paper, a study was conducted of the place occupied by CV in the scientific community in 2022. Despite the decline in popularity of this technology, it has proven itself enough for use in various fields. However, it is worth asking a question. What further development of CV is possible?

Recently, more and more publications have appeared on the topic of using CV for cryptography. CV as an application for cryptography is a poorly developed area, however, there are already several publications on this topic, for example, work [37], in which CV is used for so-called visual cryptographic schemes. This paper proposes the use of CV to build a secret sharing scheme based on some image. Thus, the application of CV in cryptography is the development of shorthand using AI. This direction is promising due to the constant development of security methods, so in the future we will conduct a deeper study of this direction.

5 Conclusion

In this work, a study was made of the state of CV technology at the time of 2022, as well as its methods and related technologies. It was found that this technology is currently relevant and in demand, despite the decline in popularity. A study of publications on

various topics and areas was conducted: medicine, zoology, electric power industry, etc. During the study, it was found that CV is being successfully implemented in various scientific fields, for example, for monitoring animals, controlling the wearing of medical masks, analyzing the results of medical research, as well as monitoring the state of electrical devices.

The results of the analysis of the used libraries of CV methods were also obtained. Such libraries allow you to effectively build various models and projects related to CV, introduce their management, configuration, and synchronization.

Thus, based on the study, we can say that CV technology for 2022 is being successfully introduced into industry, as well as into human life and activity. In 2022, the most popular CV method is ResNet, but its effectiveness is the subject of debate, which requires more research.

Acknowledgments. This work was carried out at the North Caucasus Center for Mathematical Research within agreement no. 075-02-2022-892 with the Ministry of Science and Higher Education of the Russian Federation. The reported study was funded by RFBR, Sirius University of Science and Technology, JSC Russian Railways and Educational Fund “Talent and success”, project number 20-37-51004 “Efficient intelligent data management system for edge, fog, and cloud computing with adjustable fault tolerance and security”.

References

1. Mitchell, T., Buchanan, B., DeJong, G., Dietterich, T., Rosenbloom, P., Waibel, A.: Machine learning. *Annu. Rev. Comput. Sci.* **4**(1), 417–433 (1990)
2. Shapiro, L.G., Stockman, G.C.: *Computer Vision*, vol. 3. Prentice Hall, Upper Saddle River (2001)
3. Lei, Y., et al.: Development of a slow loris computer vision detection model. *Animals* **12**(12), 1553 (2022)
4. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788 (2016)
5. Yao, J., Qi, J., Zhang, J., Shao, H., Yang, J., Li, X.: A real-time detection algorithm for Kiwifruit defects based on YOLOv5. *Electronics* **10**(14), 1711 (2021)
6. O’Shea, K., Nash, R.: An introduction to convolutional neural networks. arXiv preprint [arXiv:1511.08458](https://arxiv.org/abs/1511.08458) (2015)
7. Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W.C., Wang, C.B., Bernardini, S.: The COVID-19 pandemic. *Crit. Rev. Clin. Lab. Sci.* **57**(6), 365–388 (2020)
8. Crespo, F., Crespo, A., Sierra-Martínez, L.M., Peluffo-Ordóñez, D.H., Morocho-Cayamcela, M.E.: A computer vision model to identify the incorrect use of face masks for COVID-19 awareness. *Appl. Sci.* **12**(14), 6924 (2022)
9. Yang, S., Luo, P., Loy, C.C., Tang, X.: Wider face: a face detection benchmark. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5525–5533 (2016)
10. Batagelj, B., Peer, P., Štruc, V., Dobrišek, S.: How to correctly detect face-masks for covid-19 from visual information? *Appl. Sci.* **11**(5), 2070 (2021)
11. Ge, S., Li, J., Ye, Q., Luo, Z.: Detecting masked faces in the wild with LLE-CNNs. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2682–2690 (2017)





12. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5203–5212 (2020)
13. Targ, S., Almeida, D., Lyman, K.: Resnet in resnet: generalizing residual architectures. arXiv preprint [arXiv:1603.08029](https://arxiv.org/abs/1603.08029) (2016)
14. Zhang, H., et al.: Resnet: split-attention networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2736–2746 (2022)
15. Amari, S.I.: Backpropagation and stochastic gradient descent method. *Neurocomputing* **5**(4–5), 185–196 (1993)
16. Dehghani, M., Gritsenko, A., Arnab, A., Minderer, M., Tay, Y.: Scenic: a JAX library for computer vision research and beyond. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21393–21398 (2022)
17. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929) (2020)
18. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12346, pp. 213–229. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58452-8_13
19. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR, July 2021
20. Tolstikhin, I.O., et al.: MLP-mixer: an all-MLP architecture for vision. *Adv. Neural. Inf. Process. Syst.* **34**, 24261–24272 (2021)
21. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **21**(140), 1–67 (2020)
22. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv preprint [arXiv:1810.04805](https://arxiv.org/abs/1810.04805) (2018)
23. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: Vivit: a video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 6836–6846 (2021)
24. Tay, Y., et al.: Omninete: omnidirectional representations from transformers. In: International Conference on Machine Learning, pp. 10193–10202. PMLR, July 2021
25. Dehghani, M., et al.: The benchmark lottery. arXiv preprint [arXiv:2107.07002](https://arxiv.org/abs/2107.07002) (2021)
26. Dehghani, M., Arnab, A., Beyler, L., Vaswani, A., Tay, Y.: The efficiency misnomer. arXiv preprint [arXiv:2110.12894](https://arxiv.org/abs/2110.12894) (2021)
27. Su, J., Zhou, C., Chen, H., Xia, N., Shi, Z.: The physical and mechanical properties for flexible biomass particles using computer vision. *Fuel* **315**, 123278 (2022)
28. Kroetsch, D., Wang, C.: Particle size distribution. *Soil Sampling Methods Anal.* **2**, 713–725 (2008)
29. Toolbox, S.M.: Matlab. Mathworks Inc. (1993)
30. Stefenon, S.F., et al.: Classification of insulators using neural network based on computer vision. *IET Gener. Transm. Distrib.* **16**(6), 1096–1107 (2022)
31. Bradski, G., Kaehler, A.: OpenCV. Dr. Dobb's J. Softw. Tools **3**, 120 (2000)
32. Wood, D.A., et al.: Deep learning to automate the labelling of head MRI datasets for computer vision applications. *Eur. Radiol.* **32**(1), 725–736 (2021). <https://doi.org/10.1007/s00330-021-08132-0>
33. Lee, J., et al.: BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**(4), 1234–1240 (2020)
34. Patricio, D.I., Rieder, R.: Computer vision and artificial intelligence in precision agriculture for grain crops: a systematic review. *Comput. Electron. Agric.* **153**, 69–81 (2018)

35. Gupta, A., Pandey, A., Kesarwani, H., Sharma, S., Saxena, A.: Automated determination of interfacial tension and contact angle using computer vision for oil field applications. *J. Petrol. Explor. Prod. Technol.* **12**, 1–9 (2021). <https://doi.org/10.1007/s13202-021-01398-6>
36. Navarro Soto, J., Satorres Martínez, S., Martínez Gila, D., Gómez Ortega, J., Gámez García, J.: Fast and reliable determination of virgin olive oil quality by fruit inspection using computer vision. *Sensors* **18**(11), 3826 (2018)
37. Sherine, A., Peter, G., Stonier, A.A., Praghash, K., Ganji, V: CMY color spaced-based visual cryptography scheme for secret sharing of data. *Wirel. Commun. Mob. Comput.* (2022)

Data Analysis and Modular Computing



Discrete Neural Network of Bidirectional Associative Memory

Aleksey V. Shaposhnikov¹ , Andrey S. Ionisyan¹ , and Anzor R. Orazhev²  

¹ Department of Mathematical Modeling, North-Caucasus Federal University, Stavropol, Russia

² North-Caucasus Center for Mathematical Research, North-Caucasus Federal University, Stavropol, Russia

aorazhev@ncfu.ru

Abstract. The paper considers bidirectional associative memory, which is one of the known neural network paradigms. To simplify the implementation of the calculation of this paradigm, a discrete mathematical model of its functioning is proposed. Reducing the complexity is achieved by switching to integer calculations because Integer multiplication is several times simpler than real multiplication. The known neural network of bidirectional associative memory neural network was compared with the proposed one. The simulation was carried out in the VHDL language. For comparative evaluation, Spartan3E, Spartan6 and XC9500 chips were used. In the experimental part, it was shown that the hardware costs for the implementation of the neural network of bidirectional associative memory have decreased by more than 3 times compared to the known one. The proposed discrete model of BAM functioning does not narrow the scope of its application in comparison with the known model and can be used to build memory devices and restore distorted or noisy information.

Keywords: neural networks · bidirectional associative memory · discrete mathematical model · math modeling · artificial intelligence

1 Introduction

Bidirectional associative memory (BAM) was proposed by Kosko [1]. A distinctive feature is that a bidirectional neural network with associative memory generalizes a single-layer auto-associative Hebb correlator to a two-layer heteroassociative scheme with pattern matching [2]. Research is currently underway in this area. In [3], the issue of state estimation for a class of neural networks of bidirectional associative memory is considered. The BAM model is considered with mixed delays, which includes a constant delay in terms of leakage, a time-varying discrete delay, and a constant distributed delay. The paper [4] considers a class of simplified BAMs with multiple delays. By analyzing the associated characteristic transcendental equation, their linear stability is investigated and the Hopf bifurcation is demonstrated. Applying the Nyquist criterion, the length of the delay is estimated, which keeps the zero equilibrium stable. The article [5] considers the global asymptotic stability of equilibrium for neural networks of continuous BAM of neutral type using the Lyapunov method. Article [6] proposes a discrete

model of the Hamming neural network, making it possible to simplify the implementation of computations significantly. In [7], the synchronization problem for fuzzy (BAM) neural networks with different time delays is formulated and investigated. Sufficient conditions are given that guarantee the global asymptotic stability of a dynamical error system using the Lyapunov–Krasovskii functional method and linear matrix inequality. In [8], the problem of the global dissipativity of high-order Hopfield bidirectional associative memory neural networks with time-varying coefficients and distributed delays are discussed.

The BAM neural network considered in the work is heteroassociative. The network generates an output vector, which is associated with the input, when a random vector is received at the input of the network. Both input and output vectors are integers, and data processing by the neural network is carried out using real data [9, 10]. Moreover, the process of data processing by the network is iterative. Therefore, the implementation of the network under consideration is a complex task that requires significant resources. By transforming a continuous model of data neuroprocessing into a discrete one, it is possible to reduce the complexity of the implementation of the network under consideration, since the main neural network operations, multiplication and addition, are simpler over discrete data compared to real ones [11]. The aim of the work is to simplify the implementation of the BAM neural network by developing a discrete mathematical model of its functioning.

This work is structured as follows. In Sect. 2, a mathematical model of BAM functioning is presented. Section 3 presents an experimental simulation of a discrete neural network of bidirectional associative memory.

2 Discrete Mathematical Model of a Bidirectional Associative Memory

In this section, a heteroassociative neural network BAM will be considered, where the network forms an output vector associated with the input when an arbitrary vector is received at the input of the network. Figure 1 shows the basic BAM configuration. The input vector A is processed by the weight matrix W of the network, resulting in a neuron output vector B . The vector B is then processed by the transposed weight matrix W' of the network, which produces new output signals representing the new input vector A . The process is repeated until the network reaches a stable state in which vector A and vector B do not change [12]. Note that neurons in layers 1 and 2 function as in other paradigms, calculating the sum of the weighted inputs and calculating the value of the activation function F from it.

This process can be expressed as follows:

$$b_i = F\left(\sum_j a_j w_{ij}\right) \quad (1)$$

or in vector form:

$$B = F(AW) \quad (2)$$

where B is the vector of output signals of neurons in layer 2, A is the vector of output signals of neurons in layer 1, W is the matrix of weights of connections between layers 1 and 2, F is the activation function.

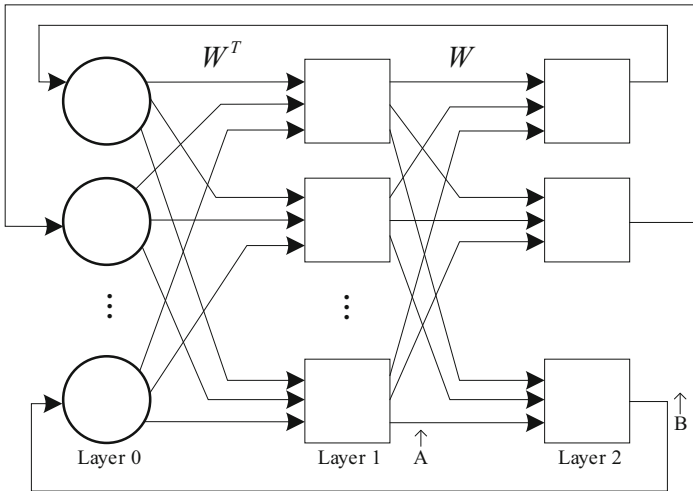


Fig. 1. Structure of the bidirectional associative memory

Likewise, $A = F(BW^T)$ where W^T is the transposition of the matrix W . The zero layer does not perform calculations and has no memory; it is only a means of distributing the layer 2 outputs to the elements of the W^T matrix.

Long-term memory (or associations) is implemented in weight arrays W and W^T . Each image consists of two vectors: vector A , which is the output of layer 1, and vector B , the associated image, which is the output of layer 2. To restore the associated image, vector A or part of it is briefly set at the outputs of layer 1. Then vector A is removed, and the network is brought to a stable state, producing the associated vector B at the output of layer 2. Next, the vector B acts through the transposed matrix W^T , reproducing the impact of the original input vector A at the output of layer 1. Each such cycle causes the output vectors of layers 1 and 2 to be refined until the point of stability in the network will not be reached. This point can be defined as resonant, as the vector is passed back and forth between the layers of the network, always processing the current output signals, but not changing them anymore. The state of neurons is a short-term memory since it can change rapidly when another input vector appears. The values of the coefficients of the weight matrix form a long-term memory and can only change over a longer time.

The network operates in the direction of minimizing the Lyapunov energy function [13]. Therefore, each cycle modifies the system towards an energy minimum, the location of which is determined by the values of the weights. This process can be visually represented in the form of a directed movement of the ball along a rubber band stretched out over the table, and each remembered image corresponds to a point “pressed” in the direction of the table surface. Figure 2 illustrates this analogy, it shows one stored image.

This process generates a minimum of gravitational energy at each point corresponding to the stored image, with a corresponding curvature of the attraction field towards this point. A freely moving ball enters the field of attraction and as a result will move towards the energy minimum, where it will stop.

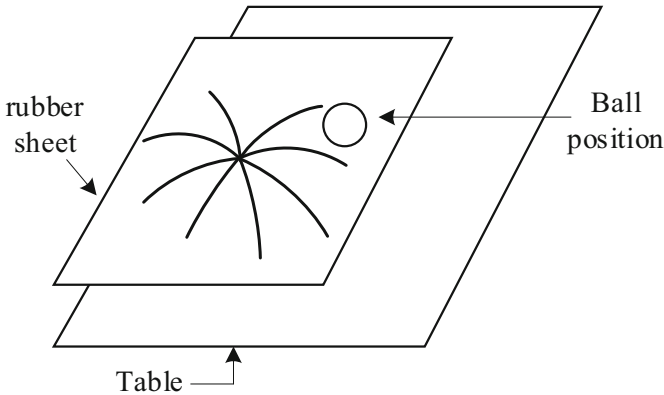


Fig. 2. Functioning of the bidirectional associative memory neural network

BAM can generalize. For example, if an incomplete or partially distorted vector is fed as A , the network tends to produce a stored vector B , which in turn seeks to correct errors in A . This may take several passes, but the network converges to reproduce the nearest stored image.

Feedback systems can tend to oscillate; this means that they can go from state to state without ever reaching stability. It is proved that all BAMs is unconditionally stable for any values of the network weights. This important property arises from the transposition relationship between the two weight matrices and means that any set of associations can be used without the risk of instability.

Typically, the network is trained to recognize multiple patterns. Training is performed using a training set consisting of pairs of vectors A and B . The learning process is implemented in the form of calculations; this means that the weight matrix is calculated as the sum of the products of all vector pairs in the training set. We write in symbolic form

$$W = \sum_i A_i^T B_i \quad (3)$$

Assume that all stored images are binary vectors. This constraint becomes less stringent if we remember that the entire contents of the university's library can be encoded into one very long binary vector. It is shown that higher productivity is achieved when using bipolar vectors. In this case, a vector component greater than 0 becomes +1, and a component less than or equal to 0 becomes -1. This is provided by the sign activation function (Fig. 3).

The network is trained by tuning the input signal, using the model of learning neurons according to the Winner Take All (WTA) principle [14]. That is, in one iteration, the state of only one neuron changes, which is selected based on two conditions:

1. The output value s_g of this neuron is greater in absolute value than the values of the other neurons of the network

$$|s_g| > |s_i|, \text{ for } i = 0 \dots n - 1, i \neq g \quad (4)$$

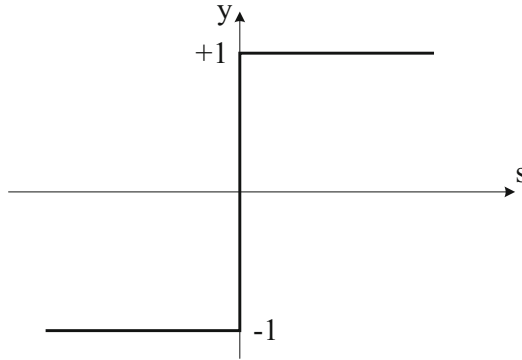


Fig. 3. Sign activation function

where g is the number of neurons that changes the state.

2. The state and output value of this neuron has different signs

$$s_g \cdot a_g < 0 \quad (5)$$

The functioning of the proposed network, in contrast to the previously discussed one, is as follows:

1. An unknown signal is applied to the network inputs.
2. The new state of neurons B of layer 2 is calculated (2).
3. In the loop for all neurons of layer B , a neuron is selected based on conditions (4,5).
4. The state of the selected k -th neuron will change.

$$b_k = -b_k \quad (6)$$

5. The new state of neurons A of layer 1 (3) is calculated.
6. In a cycle for all neurons, a neuron is selected based on conditions (4,5).
7. The state of the selected g -th neuron will change

$$a_g = -a_g \quad (7)$$

8. If the state of the network has not changed, then the end, otherwise go to step 2.

3 Experimental Modeling of a Discrete Neural Network of Bidirectional Associative Memory

Consider an example of the functioning of a discrete BAM neural network. Let's assume that it is required to train the network to memorize three pairs of binary vectors, and the vectors A_i have the same dimension as the vectors B_i . It should be noted that this is not a necessary condition for the algorithm to work; Associations can also be formed between vectors of different dimensions (Table 1).

Table 1. Initial data of functioning of discrete neural network BAM

Source vector	Associated vector
$A_1 = (0, 0, 0, 0), A'_1 = (-1, -1, -1, -1)$	$B_1 = (0, 0), B'_1 = (-1, -1)$
$A_2 = (0, 1, 0, 1), A'_2 = (-1, 1, -1, 1)$	$B_2 = (0, 1), B'_2 = (-1, 1)$
$A_3 = (1, 0, 1, 0), A'_3 = (1, -1, 1, -1)$	$B_3 = (1, 0), B'_3 = (1, -1)$
$A_4 = (1, 1, 1, 1), A'_4 = (1, 1, 1, 1)$	$B_4 = (1, 1), B'_4 = (1, 1)$

Using formula (3), we find the matrix of weight coefficients W :

$$W = \sum_{i=1}^4 A_i'^T B'_i = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 1 & -1 \\ -1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & -1 \\ -1 & 1 \\ 1 & -1 \\ -1 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 4 & 0 \\ 0 & 4 \\ 4 & 0 \\ 0 & 4 \end{pmatrix}; \tag{8}$$

$$W^T = \begin{pmatrix} 4 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \end{pmatrix}. \tag{9}$$

Iteration 1.

Let the network be simulated by the input vector $A^{(0)} = (1, 1, 1, 0)$ and $B^{(0)} = (1, 1)$ or $A^{(0)'} = (1, 1, 1, -1), B^{(0)'} = (1, 1)$, where the superscript (0) denotes iteration zero. In further consideration, we will omit '.

Compute

$$B^{(1)} == A^{(0)} \cdot W = \begin{pmatrix} 1 \\ 1 \\ 1 \\ -1 \end{pmatrix} \cdot \begin{pmatrix} 4 & 0 \\ 0 & 4 \\ 4 & 0 \\ 0 & 4 \end{pmatrix} = (8 \ 0) \tag{10}$$

the state of layer B neurons does not change.

$$A^{(1)} == B^{(1)} \cdot W^T = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 4 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \end{pmatrix} = (4 \ 4 \ 4 \ 4) \tag{11}$$

Since $A^{(0)} = (1, 1, 1, -1)$, then the 3rd bit satisfies the condition (4, 5), then taking into account (6) $A^{(1)} = (1, 1, 1, 1)$.

Iteration 2.

$$B^{(2)} == A^{(1)} \cdot W = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 4 & 0 \\ 0 & 4 \\ 4 & 0 \\ 0 & 4 \end{pmatrix} = (4 \ 4), \tag{12}$$

$$A^{(2)} == B^{(1)} \cdot W^T = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot \begin{pmatrix} 4 & 0 & 4 & 0 \\ 0 & 4 & 0 & 4 \end{pmatrix} = (4 \ 4 \ 4 \ 4). \tag{13}$$

Therefore, $A^{(2)} = A^{(1)} = (1, 1, 1, 1)$.

The network condition has recurred, which is a sign that the network has stopped functioning. Thus, the state of the network has stabilized in the fourth pair of vectors.

A comparative assessment of the consumed resources of the known and discrete BAM neural network, consisting of 8 neurons (4 neurons each in the first and second layers), was carried out based on modeling the firmware for FPGA microcircuits in the VHDL language. Were used for this Spartan3E, Spartan6 and XC9500 chips. The simulation results are summarized in Table 2.

Table 2. Hardware resources (number of logical elements) spent on designing the known and discrete BAM neural network

FPGA chip	The known neural network of BAM [1]	Discrete neural network of BAM
Spartan3E	5721	1549
Spartan6	5547	1495
XC9500	14777	3889

The simulation performed shows that the complexity of implementing neural computations for BAM using a discrete network has decreased by more than 3 times [15].

4 Conclusion

To simplify the implementation of neural computing of one of the known BAM paradigms, a discrete model of its functioning is proposed. The simulation of the proposed network in the VHDL language made it possible to conduct a comparative assessment of the known BAM neural network with the proposed one. The data obtained indicate that the hardware costs for the implementation of the BAM discrete neural network have decreased by more than 3 times compared to the known one. The proposed method is applicable for building memory devices and recovering distorted or noisy information.

Acknowledgments. The authors would like to thank the North Caucasus Federal University for supporting the contest of projects competition of scientific groups and individual scientists of the North Caucasus Federal University. The work is supported by the North-Caucasus Center for Mathematical Research under agreement № 075-02-2021-1749 with the Ministry of Science and Higher Education of the Russian Federation and by Russian Foundation for Basic Research project 1907-00130.

References

1. Kosko, B.: Bidirectional associative memories. *IEEE Trans. Syst. Man Cybern.* **18**(1), 49–60 (1988)
2. Cao, J.D., Wang, L.: Exponential stability and periodic oscillatory solution in BAM networks with delays. *IEEE Trans. Neural Netw.* **13**(2), 457–463 (2002)
3. Sakthivel, R., et al.: Design of state estimator for bidirectional associative memory neural networks with leakage delays. *Inf. Sci.* **296**, 263–274 (2015)
4. Xu, C.: Local and global Hopf bifurcation analysis on simplified bidirectional associative memory neural networks with multiple delays. *Math. Comput. Simul.* **149**, 69–90 (2018)
5. Park, J.H., et al.: A new stability criterion for bidirectional associative memory neural networks of neutral-type. *Appl. Math. Comput.* **199**(2), 716–722 (2008)
6. Shaposhnikov, A., Orazaev, A., Eremenko, E., Malakhov, D.: Hamming neural network in discrete form. In: Tchernykh, A., Alikhanov, A., Babenko, M., Samoylenko, I. (eds.) *MANCS 2021. LNCS*, vol. 424, pp. 11–17. Springer, Cham (2022). https://doi.org/10.1007/978-3-030-97020-8_2
7. Ratnavelu, K., Manikandan, M., Balasubramaniam, P.: Synchronization of fuzzy bidirectional associative memory neural networks with various time delays. *Appl. Math. Comput.* **270**, 582–605 (2015)
8. Aouiti, C., Sakthivel, R., Touati, F.: Global dissipativity of high-order hopfield bidirectional associative memory neural networks with mixed delays. *Neural Comput. Appl.* **32**(14), 10183–10197 (2020)
9. Zhao, H.: Global stability of bidirectional associative memory neural networks with distributed delays. *Phys. Lett. A* **297**(3–4), 182–190 (2002)
10. Cao, J., Liang, J., Lam, J.: Exponential stability of high-order bidirectional associative memory neural networks with time delays. *Physica D* **199**(3–4), 425–436 (2004)
11. Kosko, B.: Adaptive bidirectional associative memories. *Appl. Opt.* **26**(23), 4947–4960 (1987)
12. Aouiti, C., Sakthivel, R., Touati, F.: Global dissipativity of fuzzy bidirectional associative memory neural networks with proportional delays. *Iranian J. Fuzzy Syst.* **18**(2), 65–80 (2021)
13. Humphries, U., et al.: Global stability analysis of fractional-order quaternion-valued bidirectional associative memory neural networks. *Mathematics* **8**(5), 801 (2020)
14. Makhzani, A., Frey, B.J.: Winner-take-all autoencoders. *Adv. Neural Information Processing Systems*, vol. 28 (2015)
15. Ionisyan, A.S.: Investigation of FPGA utilization of continues and discrete bidirectional associative memory neural networks. https://github.com/anserion/bidirmem_VHDL. Accessed 16 June 2022



Modulo $2^k + 1$ Truncated Multiply-Accumulate Unit

Maxim Bergerman¹ , Pavel Lyakhov² , and Albina Abdulsalyamova¹ 

¹ North-Caucasus Center for Mathematical Research, Stavropol, Russia
maxx07051997@inbox.ru

² North-Caucasus Federal University, Stavropol, Russia

Abstract. Digital signal processing requires the calculation of large data volume. To increase the speed of data processing, a Residue Number System is used. This number system provides performing calculations in parallel, reducing time costs. In practice, moduli of the Residue Number System of a special form (2^k , $2^k - 1$, $2^k + 1$) are most often used. The article proposes a method for calculating modulo $2^k + 1$ using the Diminished-one coding technique and developed the Inverted End-Around Carry Truncated Multiply-Accumulate unit (IEAC-TMAC). This approach increases the number of moduli, affecting a decrease in the capacity of the moduli and the delay. Hardware modeling shows that, compared with the existing varieties of TMAC blocks, the proposed block demonstrates worse results in terms of hardware costs by 27–231%, depending on the block being compared. However, using two blocks of 2 times less bit width of the form ($2^k - 1$, $2^k + 1$) provides reducing the occupied area of a device in comparison with the modulo ($2^{2k} - 1$) by 24–48% times and decreasing the execution speed by 1.20–1.24 times. A promising direction for further research will be the development of digital signal processing devices with moduli of a special type (2^k , $2^k - 1$, $2^k + 1$).

Keywords: Residue number system · diminished-one · hardware implementation · digital signal processing · field-programmable gate array

1 Introduction

Digital filtering solves most practical problems in various fields of science: denoising [1], equalization [2], interpolation [3], and many others. The amount of information passing through digital filters is constantly growing, and therefore it is necessary to increase the performance of digital signal processing (DSP) devices.

The low performance of DSP devices is an urgent problem today. Scientists are conducting research to find an increase in the performance of digital devices [4, 5]. One way to solve this problem is to use modular arithmetic, such as the Residue Number System (RNS). This number system allows you to perform calculations in parallel, which allows you to increase the speed of calculation by dividing into moduli, reducing the bit depth on each channel of calculation [6]. However, the disadvantages of this number system are the complexity of performing division operations, comparing numbers,

determining the sign, etc. To solve these problems, it is necessary to perform the reverse conversion to the positional number system (PNS), which increases the hardware and time costs for the implementation of the device. To reduce the influence of this drawback, RNS moduli of special type are used: 2^k , $2^k - 1$, $2^k + 1$. This article considers a multiplication-accumulate method for a modulo $2^k + 1$ using high-performance adders.

The rest of the paper is organized as follows. Section 2 will reflect the methods that are used for the other two RNS moduli (2^k and $2^k - 1$) and the proposed method for the module ($2^k + 1$). This Section is divided into 4 subsections: Background on Residue Number System; Coding in Diminished-one for modulo $2^k + 1$ representation; State of the art Truncated Multiply-Accumulate units; Proposed modulo $2^k + 1$ TMAC architecture. Section 3 provides hardware modeling of the proposed architecture and comparison with existing ones; the simulation results. Section 4 concludes the results obtained.

2 Preliminaries

2.1 Background on Residue Number System

The Residue Number System is a non-positional value system. This means that calculations in this system do not depend on the position of the digit of the number, as, for example, in the Roman number system. Numbers in RNS are represented by a set of remainders after dividing by the bases of the system, called the moduli of the system. Calculations in modular arithmetic occur for each modulus in parallel, that is, independently of each other. For example, calculations in RNS are performed according to formula (1):

$$X * Y = (x_1 * y_1, x_2 * y_2, \dots, x_n * y_n), \quad (1)$$

where $*$ denotes arithmetic operations: addition (+), subtraction (-), or multiplication (\cdot). Each RNS modulus is pairwise coprime with other moduli, i.e., condition is met: Greatest Common Divisor (GCD) $(m_i, m_j) = 1$, for all $i \neq j$.

To perform the reverse conversion from RNS to PNS, methods such as Chinese Remainder Theorem (CRT) [7], Chinese Remainder Theorem with Fractional Values (CRTf) [8], Mixed-Radix Conversion (MRC) [9] are used.

2.2 Coding in Diminished-One for Modulo $2^k + 1$ Representation

The complexity of a modulo $2^k + 1$ arithmetic unit is determined by the representation chosen for the input operands. There are several representations, such as representations in Diminished-one (D-1) [10] and signed-LSB representations [11]. In this work, the D-1 coding method will be used.

Let x' be the number represented in D-1. To convert a number to D-1, you need to perform the conversion presented under formula (2):

$$x' = |x - 1|_{2^k + 1} \quad (2)$$

for $x \in [1, 2^k]$. Hence it follows that $x' \in [0, 2^k - 1]$. Since modulo $2^k + 1$ occupies $(k + 1)$ -bits, occupies k -bits by discarding zero in the computations and adding a zero-definition bit. The full representation of the number in D-1 is represented by formula (3)

$$X' = \underbrace{x'_k}_{1 \text{ bit}} \underbrace{x'}_{k \text{-bit}}, \quad (3)$$

where $x' = (x'_{k-1}x'_{k-2} \dots x'_1x'_0)$, and x'_k is the bit to determine the sign of the number. If the number is $x \in [1, 2^k]$, then this bit is equal to '0'. For $x = 0$, the sign bit will equal '1'. An example of representation in D-1 in binary number system at is shown in Table 1.

Table 1. Correspondence between decimal, binary and Diminished-1 representations

Decimal	Binary	Diminished-one
0	0000	1
1	0001	2
2	0010	3
3	0011	4
4	0100	5
5	0101	6
6	0110	7
7	0111	8
8	1000	0

The results of addition and multiplication in D-1 are represented by formulas (4) and (5):

$$|x + y|_{2^{k+1}} = |x' + y' + c_{out}|_{2^{k+1}} = A' \quad (4)$$

$$|x \cdot y|_{2^{k+1}} = (x' \cdot y' + x' + y')_{2^{k+1}} = P' \quad (5)$$

where A' and P' are presented in D-1 [12].

2.3 State of the Art Truncated Multiply-Accumulate Units

For calculations on a field-programmable gate array (FPGA) with an RNS modulo 2^k such high-speed adder architectures as Carry-save adder (CSA) [13] and Kogge-Stone adder (KSA) [14] are used. For modulo $2^k - 1$ End-Around Carry CSA (EAC-CSA) and End-Around Carry KSA (EAC-KSA) are used [13].

To implement digital filters by modulo 2^k and $2^k - 1$, modifications of the Multiply-Accumulate units (MAC) blocks are used, which are called Truncated Multiply-Accumulate units (TMAC) (Fig. 1) and End-Around Carry Truncated Multiply-Accumulate units (EAC-TMAC) (Fig. 2) [15]. These blocks consist of a partial product generator (PPG) and a CSA tree. For each modulus, the special kind of PPG will be different. For example, for moduli 2^k and $2^k - 1$, the PPGs are shown in Figs. 3 and 4, respectively. 4 numbers are fed to the input of the block: 2 multipliers are multiplied with each other, which are fed into the PPG, the remaining 2 numbers are summands and the adder tree is fed immediately. At the output of the block, 2 numbers are displayed, which are either fed into the next block of the corresponding TMAC, or summed up using the corresponding KSA adder.

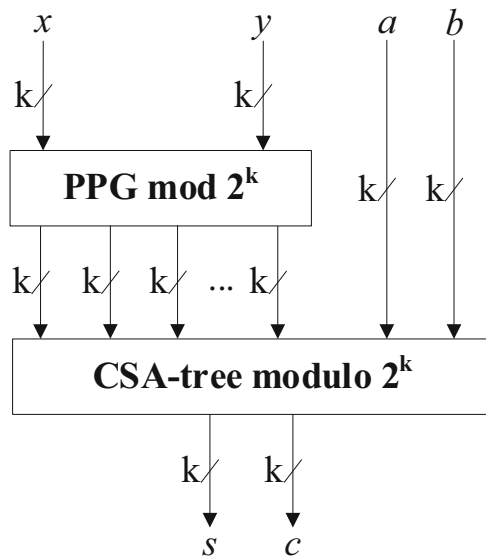


Fig. 1. TMAC block diagram

2.4 Proposed Modulo $2^k + 1$ TMAC Architecture

In this work, an Inverted End-Around Carry Truncated Multiply-Accumulate unit (IEAC-TMAC) was developed for modulo $2^k + 1$ in D-1. This block performs the product and addition according to the formula (6):

$$|x \cdot y + a + b|_{2^{k+1}} = \begin{cases} |x' \cdot y' + x' + y' + a' + b'|_{2^{k+1}}, & \text{if } x'_k \vee y'_k = 0 \\ |a' + b'|_{2^{k+1}}, & \text{if } x'_k \vee y'_k = 1 \end{cases} \quad (6)$$

The product $x' \cdot y'$ is calculated in the generator of partial products modulo $2^k + 1$ (Fig. 5). High-order carries in PPG that exceed values of k are carried to the least

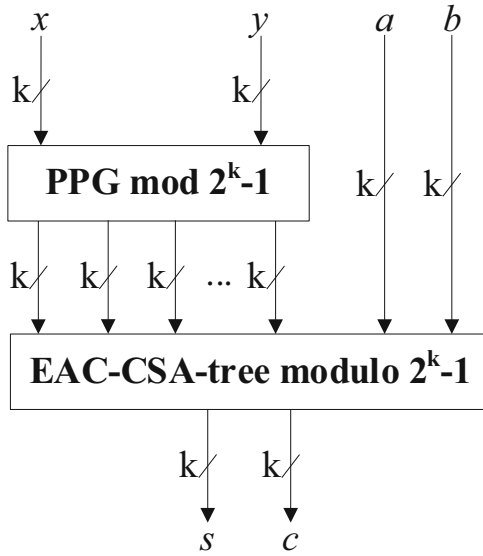


Fig. 2. EAC-TMAC block diagram

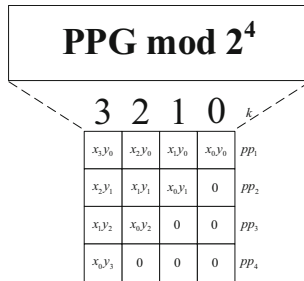


Fig. 3. Schematic of the partial product generator for modulo 2^k with $k = 4$

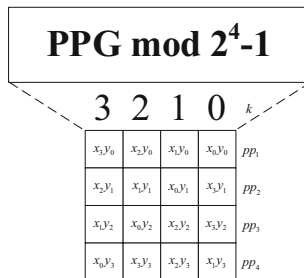


Fig. 4. Schematic of the partial product generator for modulo $2^k - 1$ with $k = 4$

significant bits and their values are inverted. To avoid an inversion error, the corrective value *cor* is supplied, which has the value

$$cor = \underbrace{00\dots 00}_{k-bit}. \tag{7}$$

Since the number “0” values are fed into the block, *a* and *b* values are fed into the block equal to:

$$a' = \underbrace{00\dots 00}_{k-bit} \text{ and } b' = \underbrace{11\dots 11}_{k-bit}, \tag{8}$$

the sum of which gives the value $|a' + b'|_{2^{k+1}} = 0$.

The received values from the PPG are fed into the IEAC-CSA adder tree (Fig. 6). Unlike other TMAC block diagrams, here you need to take into account the sign of two numbers: *x'* and *y'*. If the sign of at least one of the two numbers is equal to ‘0’ (Number “0” in Diminished-one), then the product will be calculated incorrectly. In this regard, a multiplexer is used to check the sign. After checking the meaning of the characters, the correct result is displayed, which is subsequently fed either to another IEAC-TMAC unit or to an IEAC-KSA. The proposed architecture of the IEAC-TMAC block is shown in Fig. 7.

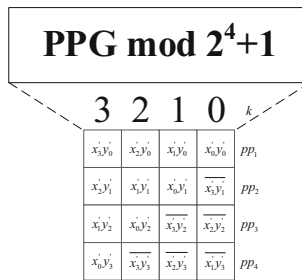


Fig. 5. Schematic of the partial product generator for modulo 2^k + 1 with k = 4

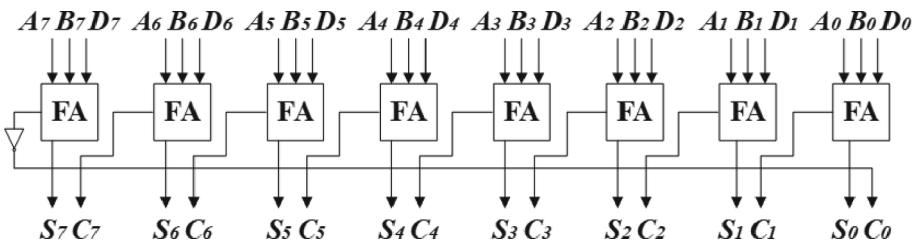


Fig. 6. IEAC-CSA adder circuit at k = 8

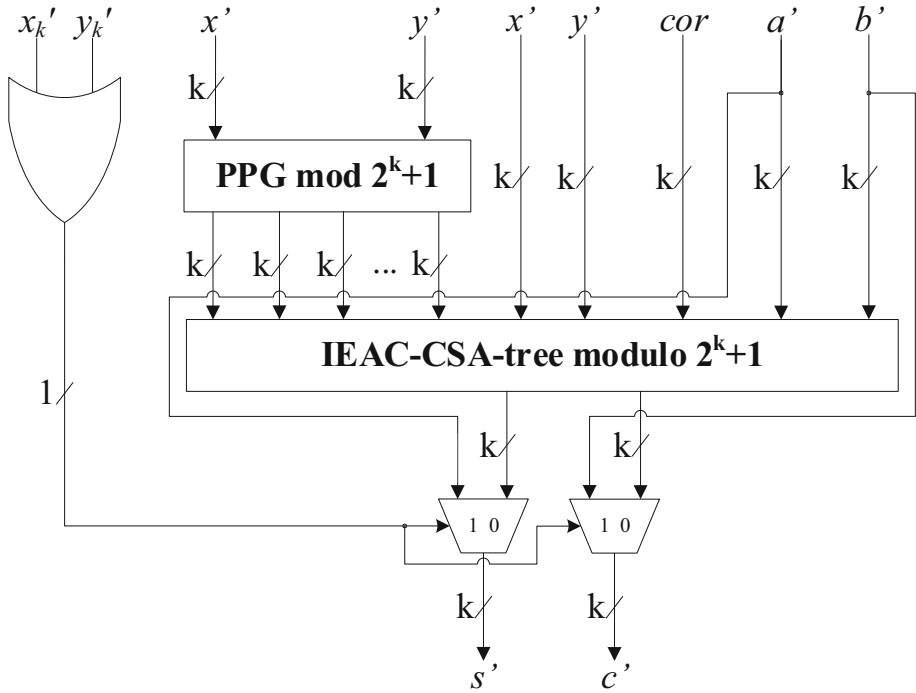


Fig. 7. Proposed IEAC-TMAC architecture diagram

3 Hardware Implementation and Results

Hardware modeling was carried out using the VHDL hardware language in the Xilinx Vivado 2018.2 environment. In this simulation, the bit depths were taken $k = 4, 8, 16$ for each TMAC block and will take into account the time and hardware costs, as well as the use of energy resources. The IEAC-TMAC block consists of the forward conversion to Diminished-one, the proposed IEAC-TMAC architecture, and the reverse conversion from Diminished-one. Hardware simulation parameters are shown in Table 2. Simulation results are presented in Table 3.

Table 2. Hardware simulation parameters on Xilinx Vivado 2018.2

Target board	xc7vx485tffg1157-1
bufg	12
Fanout_limit	10000
Max_bram	0

(continued)

Table 2. (continued)

Max_uram	0
Max_dsp	0
Max_bram_cascade_height	0
Max_uram_cascade_height	0
Synthesis strategy	Vivado Synthesis Defaults (2018)
Implementation strategy	Vivado Implementation Defaults (2018)

Table 3. Hardware results

k, bits		TMAC	EAC-TMAC	IEAC-TMAC
4	Area, LUT	13	24	43
	Delay, ns	6.954	7.740	8.293
	Power, W	4.394	6.948	4.857
8	Area, LUT	55	88	120
	Delay, ns	8.946	10.284	9.473
	Power, W	12.160	17.069	11.185
16	Area, LUT	222	400	507
	Delay, ns	12.239	12.431	13.464
	Power, W	30.374	45.427	32.203

After analyzing the obtained results, it can be concluded that the proposed IEAC-TMAC block showed the worst results in terms of hardware and time costs compared to known blocks, which was quite obvious due to the large amount of addition and the addition of multiplexers. However, in terms of power consumption, IEAC-TMAC outperforms EAC-TMAC in the corresponding bits, but is worse than the TMAC block.

The results also show that the use of two moduli of smaller capacity instead of one larger capacity, where $2^k - 1 = (2^{\frac{k}{2}} - 1)(2^{\frac{k}{2}} + 1)$, can reduce hardware costs. For example, for one modulus at $k = 8$, 88 LUT Slices are needed, and for two moduli $(2^{\frac{k}{2}} - 1, 2^{\frac{k}{2}} + 1)$ at $\frac{k}{2} = 4$ only $24 + 43 = 67$ LUTs. And due to parallel computing, the latency of one modulus will be greater compared to two moduli, which is defined as the maximum delay of two moduli. This is also observed at $k = 16$, where instead of one modulus with the cost of 400 LUT Slices, two moduli with hardware costs of $88 + 120 = 208$ LUTs can be used, which reduces the cost of hardware resources by almost 2 times. In comparison with the delay, the modulus $2^{16} + 1$ showed a worse result than two moduli $(2^8 - 1, 2^8 + 1)$.

4 Conclusion

The article proposed an IEAC-TMAC block for calculating numbers in RNS in D-1 modulo $2^k + 1$. The hardware simulation results showed that, compared to known blocks, the hardware costs of the IEAC-TMAC block showed results worse by 27–79% compared to EAC-TMAC blocks and worse by 118–231% compared to the TMAC block. In terms of running time, the IEAC-TMAC block is calculated longer by 10–19% compared to the TMAC block. Compared to the EAC-TMAC block, in terms of delay for $k = 4$ and $k = 16$ the proposed block showed worse results by 7 and 8%, respectively, and for the IEAC-TMAC block, it was better by 8%. In terms of power consumption, the proposed IEAC-TMAC block showed worse results compared to the TMAC block at $k = 4$ and $k = 16$ by 11% and 6%, respectively, and at $k = 8$ the IEAC-TMAC block was better by 8%. Compared to the EAC-TMAC block, in terms of energy consumption, the proposed block showed better results by 29–34%.

The development of the IEAC-TMAC block makes it possible to increase the number of RNS moduli used, thereby reducing the capacity for each modulus, which will increase the speed of calculations and reduce hardware costs. The use of two moduli of smaller capacity instead of one larger capacity, where $2^k - 1 = (2^{\frac{k}{2}} - 1)(2^{\frac{k}{2}} + 1)$, can reduce hardware costs by 24–48% and increase the speed of computing by 1.20–1.24 times.

A promising direction for further research is the development of digital signal processing devices using sets of RNS moduli of a special type $\{2^k, 2^k - 1, 2^k + 1\}$ using high-performance architectures of TMAC blocks.

Acknowledgments. The research in Section 2 was supported by North-Caucasus Center for Mathematical Research under agreement № 075-02-2023-938 with the Ministry of Science and Higher Education of the Russian Federation. The rest of this paper was supported by Russian Science Foundation, project 21-71-00017.



References

1. Ahmad, K., Khan, J., Iqbal, M.S.U.D.: A comparative study of different denoising techniques in digital image processing. In: 8th International Conference on Modeling Simulation and Applied Optimization (ICMSAO), pp. 1–6 (2019)
2. Kiran, S., et al.: Modeling of ADC-based serial link receivers with embedded and digital equalization. *IEEE Trans. Compon. Packag. Manuf. Technol.* **9**(3), 536–548 (2019)
3. Porshnev, S.V., Kusaykin, D.V., Klevakin, M.A.: On accuracy of periodic discrete finite-length signal reconstruction by means of a Whittaker-Kotelnikov-Shannon interpolation formula. In: 2018 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), pp. 165–168 (2018)
4. Zendegani, R., Kamal, M., Bahadori, M., Afzali-Kusha, A., Pedram, M.: RoBA multiplier: a rounding-based approximate multiplier for high-speed yet energy-efficient digital signal processing. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **25**(2), 1–9 (2016)
5. Sun, X., Guo, Y., Liu, Z., Kimura, S.: A radix-4 partial product generation-based approximate multiplier for high-speed and low-power digital signal processing. In: 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS), pp. 777–780 (2018)

6. Omondi, A., Premkumar, B.: *Residue Number Systems: Theory and Implementation*, p. 296. Imperial College Press (2007)
7. Mohan, P.V.A.: *Residue Number Systems: Theory and Applications*. Birkhauser, Basel, Switzerland (2016)
8. Chervyakov, N.I., Molahosseini, A.S., Lyakhov, P.A., Babenko, M.G., Deryabin, M.A.: Residue-to-binary conversion for general moduli sets based on approximate Chinese remainder theorem. *Int. J. Comput. Math.* **94**(9), 1833–1849 (2017)
9. Mohan, P.V.A., Phalguna, P.S.: Evaluation of mixed-radix digit computation techniques for the three moduli RNS $\{2n-1, 2n, 2n+1-1\}$. *IEEE Trans. Circuits Syst. II Express Briefs* **68**(4), 1418–1422 (2021)
10. Leibowitz, L.M.: A simplified binary arithmetic for the fermat number transform. *IEEE Trans. Acoust. Speech Sig. Process.* **ASSP-24**(5), 356–359 (1976)
11. Jaberipur, G., Parhami, B.: Unified approach to the design of modulo-adders based on signed-LSB representation of residues. In: *Proceedings of 19th IEEE Symposium on Computer Arithmetic*, pp. 57–64 (2009)
12. Živaljević, D., Stamenković, N., Stojanović, V.: Digital filter implementation based on the RNS with diminished-1 encoded channel. In: *35th International Conference on Telecommunications and Signal Processing (TSP)*, pp. 662–666 (2012)
13. Parhami, B.: *Computer Arithmetic: Algorithms and Hardware Designs*, p. 492. Oxford University Press, Oxford (2010)
14. Kogge, P.M., Stone, H.S.: A parallel algorithm for the efficient solution of a general class of recurrence equations. *IEEE Trans. Comput.* **C-22**(8), 786–793 (1973)
15. Lyakhov, P., Valueva, M., Valuev, G., Nagornov, N.: High-performance digital filtering on truncated multiply-accumulate units in the residue number system. *IEEE Access* **8**, 209181–209190 (2020)



Neural Network Skin Cancer Recognition with a Modified Cross-Entropy Loss Function

Ulyana Alekseevna Lyakhova  

Department of Mathematical Modeling, North-Caucasus Federal University, 355017 Stavropol,
Russia
uljahovs@mail.ru

Abstract. Skin cancer is currently one of the most common types of human cancer. Due to similar morphological manifestations, the diagnosis of malignant neoplasms is difficult even for experienced dermatologists. Artificial intelligence technologies can equal and even surpass the capabilities of an oncologist in terms of the accuracy of visual diagnostics. The available databases of dermoscopic images and statistical data are highly unbalanced about “benign” cases. When training neural network algorithms on unbalanced bases, there is a problem of reducing the accuracy and performance of models due to the prevailing “benign” cases in the samples. One of the possible ways to solve the problem of unbalanced learning is to modify the loss function by introducing different weight coefficients for the recognition classes. The article proposes a neural network system for the recognition of malignant pigmented skin neoplasms, trained using a modified cross-entropy loss function. The accuracy of recognition of malignant neoplasms of the skin in the proposed system was 88.12%. The use of the proposed system by dermatologists-oncologists as an auxiliary diagnostic method will expand the possibilities of early detection of skin cancer and minimize the influence of the human factor.

Keywords: Convolutional Neural Networks · Imbalanced Classification · Cost-sensitive Learning · Skin Lesion Analysis · Melanoma · Cancer

1 Introduction

Skin cancer is one of the most common types of cancer in the human body. Although there are well-characterized risk factors such as ultraviolet radiation (UV) or genetic predisposition that cause skin cancer, many remain unidentified. In this regard, there is a problem with the timely diagnosis of malignant neoplasms [1]. The five-year survival rate for melanoma diagnosed at an early stage is 95%, while the survival rate for advanced stages is approximately 15% [2]. Dermoscopy is a non-invasive method for visual diagnosis of pigmented skin lesions used in dermatology [3]. Dermoscopic images have great potential in the early diagnosis of malignant pigmented neoplasms, but their interpretation is subjective even for qualified dermatologists [4]. The average accuracy of recognition of pigmented neoplasms in practicing dermatologists is approximately 65–75% [5]. The low accuracy of recognition of oncological forms of pigmented

lesions in visual diagnosis is associated with similar early manifestations of benign and malignant neoplasms [6]. Therefore, there is currently great interest in the development of automated computer diagnostic systems that can help dermatologists in the clinical evaluation of various cases of pigmented skin lesions.

Today, Convolutional Neural Networks (CNN) are the most common tool for image analysis and classification. CNN has achieved highly accurate results on a variety of classification problems, but despite the broad outlook, there are some challenges. They are due to the large size of the networks, reaching millions of parameters, the lack of reliable balanced training data sets, and problems with overfitting. To train modern neural network architectures that include millions of parameters, large databases are needed with data evenly distributed over various categories.

Research has shown that medical databases are generally unbalanced towards classes with healthy patients, which significantly outnumber the class with cases of sick patients [7]. At the same time, machine learning algorithms that are used for binary classification problems assume a uniform distribution of classes. When training on databases with unbalanced data, the samples are dominated by cases from the majority class, which reduces the accuracy and performance of the model [8].

Machine learning algorithms assume that false negative and false positive errors in classification are equal [9]. However, in medical data recognition problems, this assumption can be dangerous [10]. False-negative classification of a pigmented neoplasm may lead to a later diagnosis of skin cancer and increase the chance of death.

One possible way to deal with data imbalance is to use resampling techniques [11]. These methods are aimed at manually balancing data by using an insufficient amount of data from a dominant class or an excessive amount of data from a minority class. Both undersampling and oversampling methods change the data distribution of different classes [12]. Significant disadvantages of these methods are the possibility of skipping diagnostically significant data during training, as well as an increase in computational costs. Another way to solve the problem of data imbalance is data augmentation using traditional affine transformations [13]. But, despite the numerous advantages of this method, in some cases, simple classical operations are not enough to significantly improve the accuracy of a neural network or overcome the problem of overfitting [14].

There is a cost-sensitive learning method that takes into account the cost associated with the false classification of samples [15]. This method is based on the use of cost matrices that contain class weights. Instead of using artificial balancing of data distribution during training, matrices contain cost factors associated with incorrect recognition of different classes [16]. Because the cost of classification errors is taken into account in machine learning of neural network algorithms [9], cost-based learning methods are the most optimal for datasets with asymmetric distribution [17].

This paper presents a neural network system for recognizing malignant pigmented lesions, trained using the cost-sensitive learning method. This method avoids problems when learning from unbalanced dermoscopic images. The rest of the work is structured as follows. Section 2 is divided into several subsections. In sub-Sect. 2.1. a description of the modification of the cross-entropy loss function, which used in training the neural network recognition system for dermoscopic images, is presented. In sub-Sect. 2.2. The definition of a neural network system for the classification of pigmented skin lesions

based on dermoscopic images is proposed. Section 3 presents a practical simulation of the proposed neural network classification system for pigmented neoplasms, which was trained using the cost-sensitive learning method. Section 4 discusses the results obtained and their comparison with known works in the field of neural network classification of dermoscopic skin images. In conclusion, the results of the work are summed up and conclusions are drawn.

2 Materials and Methods

Machine learning algorithms aim to minimize errors in the learning process. False-positive and false-negative classification errors are assigned the same value, taking into account their equal importance. As a result, the neural network system tends to correctly classify and favor the more frequently occurring classes. In cost-sensitive learning, a misclassification is subject to a penalty called cost. Cost-sensitive learning aims to minimize the cost of data misclassification by a neural network system. Instead of optimizing accuracy, the neural network tries to minimize the total cost of misclassification [18].

The paper proposes a neural network system for the recognition of malignant pigmented skin neoplasms trained using a modified cross-entropy loss function. The loss function is modified by introducing various weight coefficients for recognition classes into the learning process. The modified cross-entropy function will avoid such a problem with training a neural network system on unbalanced data when the classification results can be biased towards a larger class. The neural network recognition system for malignant pigmented neoplasms consists of various pre-trained CNN architectures that process dermoscopic images. The output signal of the proposed neural network system for recognizing pigmented skin lesions is the percentage for 2 diagnostic categories.

2.1 Modification of the Cross-Entropy Loss Function Using Weighting Coefficients

Standard machine learning methods use a loss function L that assigns a value of $L = 0$ to a correctly recognized dermoscopic image and a value of $L = 1$ to an incorrectly recognized dermoscopic image. The procedure for training a neural network model is aimed at minimizing the total cost or at minimizing the number of erroneous predictions. Since the training loss function L uses the same misclassification cost for all considered recognition categories, the neural network system is susceptible to asymmetric data distribution. This raises a problem when the loss function L is minimized if the neural network system focuses on the class with the most data and ignores the minority class. The larger the class imbalance coefficient, the more pronounced the problem becomes.

Currently, the bulk of the available dermatological data is from healthy patients and benign pigmented lesions. When trained on such dermatological bases, the standard loss function L is successfully minimized when the neural network system predicts all inputs as “benign”. In this regard, there is a need to develop special approaches to neural network learning, taking into account costs. Such approaches involve unequal misclassification costs between classes, which can be defined as a cost matrix or weighting factors.

The loss function is used to determine how well the neural network model fits the data distribution. Using cross-entropy, one can estimate the error between two probability distributions. In the case of binary classification, when the number of k classes is 2, the cross-entropy loss $L_{(cr)}$ can be calculated as follows:

$$L_{(cr)} = \frac{1}{W} \sum_{d=1}^W [-(p_d \log(\widehat{p}_d) + (1 - p_d) \log(1 - \widehat{p}_d))], \quad (1)$$

where W is the total number of dermoscopic images in the database; p_d is the actual value of the target class; \widehat{p}_d is the probability of predicting the target class.

In the case of binary classification of dermoscopic images, the class of k can take the value equal to $k = 1$ if the pigmented neoplasm belongs to the category “benign” and the value $k = 2$ if the pigmented neoplasm belongs to the category “malignant”. The cost of the costs can be considered as a penalty factor introduced during the training procedure of the neural network system, aimed at increasing the importance of the class of “malignant” pigmented neoplasms. By stricter punishment for misclassification in a given category, the neural network classifier focuses on the dermoscopic images coming from this distribution. To prevent a bias in the classification of the neural network towards more common classes, it is necessary to modify the cross-entropy function using class weights. Class weights are inversely proportional to class frequency and are calculated as follows:

$$m_k = \frac{W}{S \sum_{d=1}^S q_{dk}}, \quad (2)$$

where S is the number of classes; q_{nk} is an indication that the d image belongs to the k class.

After adding the class weight, the modified cross-entropy function $L'_{(cr)}$ can be represented as follows:

$$L'_{(cr)} = \frac{1}{W} \sum_{d=1}^W [-(m_1(p_d \log(\widehat{p}_d)) + m_2((1 - p_d) \log(1 - \widehat{p}_d)))], \quad (3)$$

where m_1 is class 1 weight; m_2 is class 2 weight.

Thus, the modified cross-entropy function will avoid such a problem with training a neural network system on unbalanced data, when the classification results can be biased towards a more common class.

2.2 Convolutional Neural Network System for Recognition of Pigmented Skin Lesions

In the field of diagnosing pigmented skin lesions, the most common types of data are dermoscopic images. To date, for the recognition of multidimensional visual data, the most optimal neural network architecture is CNN. For training, the input of the proposed

neural network system for the classification of pigmented skin lesions receives dermoscopic images of the $D_{(img)}$ and labels with a diagnosis of $j \in \{1, \dots, k\}$, where k is the number of diagnostic classes.

The input of the convolutional layer receives a dermoscopic image of the $D_{(img)}$, converted into a three-dimensional function $D_{(x,y,z)}$, where $0 \leq x < Q$, $0 \leq y < R$ and $0 \leq z < S$ are the spatial coordinates, where Q – rows, R – columns and S – color components. The amplitude D at any point with coordinates (x, y, z) is the intensity of the pixels at that point. Then the procedure for obtaining feature maps in the convolutional layer is as follows:

$$D_{(f)}(x, y) = h + \sum_{i=-\frac{v-1}{2}}^{\frac{v-1}{2}} \sum_{j=-\frac{v-1}{2}}^{\frac{v-1}{2}} \sum_{c=0}^{G-1} v_{ijc}^{(1)} B(x+i, y+j, c), \quad (4)$$

where $D_{(f)}$ is a feature map; $v_{ijc}^{(1)}$ is the filter coefficient of size $\alpha \times \alpha$ for processing B arrays; h is the offset factor.

The activation of the last layer of the neural network model for recognizing pigmented skin lesions is displayed through the softmax function. The output probability distribution of which class the pigment neoplasm belongs, which was obtained as a result of the operation of the softmax function, is compared with the original correct distribution. Categorical cross-entropy loss is used only with neural network systems with a softmax output function. The cross-entropy function indicates the distance between the output distribution and the original probability distribution. In this way, the predicted probabilities gradually remember the true vectors and there is a minimization of the loss during training.

3 Results

To simulate a neural network system for recognizing pigmented neoplasms, dermoscopic images from the open archive of The International Skin Imaging Collaboration (ISIC) [19] were used. The selected data included 32,454 dermoscopic photographs of various sizes, which were divided into two groups: benign neoplasms (21,939 images) and malignant neoplasms (10,515 images).

The simulation was carried out using the high-level programming language Python 3.8.8. All calculations were performed on a PC with an Intel (R) Core (TM) i5-8500 processor at 3.00 GHz with 16 GB of random-access memory (RAM) and a 64-bit Windows 10 operating system. CNN training was carried out using a graphics processing unit (GPU) based on the NVIDIA GeForce GTX 1050TI video chipset.

CNN AlexNet, Inception_v4, and ResNeXt were chosen to model the neural network system for recognizing malignant pigmented neoplasms, which were pre-trained on a set of ImageNet natural images. The neural network architectures Inception_v4 and ResNeXt are currently recognized as the most productive compared to a human [20]. At the same time, the AlexNet deep neural network architecture does not require specialized hardware and works well with a limited GPU; AlexNet learning is faster than other deeper architectures [21].

The most common size for dermoscopic images in the ISIC database is $450 \times 600 \times 3$, where 3 are color channels. For the AlexNet neural network architecture, images were converted to $227 \times 227 \times 3$, for Inception_v4 to $299 \times 299 \times 3$, and for ResNeXt to $256 \times 256 \times 3$. For further modeling, the database of dermoscopic photographs was divided into images for training and images for verification in a percentage ratio of 80 to 20.

Since the ISIC dermoscopic image database is highly unbalanced toward the category of “benign” neoplasms, a cost-sensitive learning method was used for training. For this, the cross-entropy loss function was modified. For the training set of images, the weight coefficients of each of the classes were calculated using the formula (2) and amounted to 0.73 for the class “benign” of neoplasms and 1.54 for the class “malignant” of neoplasms.

Table 1 presents the results of assessing the accuracy of recognition of dermoscopic images of pigmented skin lesions, as well as such methods for quantitative assessment of neural network systems as the loss function, F1-score, Matthew’s correlation coefficient (MCC), recall and sensitivity.

Table 1. Results of modeling neural network architectures for recognizing pigmented neoplasms on dermoscopic images

CNN architecture	Loss function weights	Loss function	Accuracy, %	F1-score	MCC	Recall	Sensitivity
AlexNet	Not used	0.41	81.60	0.82	0.63	0.83	0.72
	Used	0.28	82.70	0.85	0.65	0.85	0.85
Inception_v4	Not used	0.07	85.33	0.87	0.71	0.86	0.79
	Used	0.04	87.35	0.88	0.72	0.88	0.87
ResNeXt	Not used	0.12	87.30	0.89	0.77	0.89	0.82
	Used	0.09	88.12	0.91	0.79	0.91	0.90

The highest accuracy rate for recognizing pigmented skin lesions was achieved using CNN ResNeXt, trained using weight coefficients, and amounted to 88.12%. When training a CNN architecture using weights, the resulting recognition accuracy percentage was higher than when training the original CNNs without using weights. The increase in the accuracy of recognizing pigmented lesions during training of various neural network systems was 0.82–2.02%, depending on the architecture of the CNN. The least loss function was achieved when modeling a neural network system based on Inception_v4 using training weights and amounted to 0.04. The use of the modified cross-entropy for training made it possible to reduce the value of the loss function depending on the CNN architecture by 0.03–0.13.

F1 is a general score that is defined as the harmonic mean of model accuracy and recall and is used to evaluate binary classification systems. The best F1-score was obtained when testing a neural network architecture based on ResNeXt trained using weight coefficients and amounted to 0.91. The increase in F1-score when training with weights

for various CNN architectures was 0.01–0.03 compared to F1-scores for trained CNNs without weights.

A recall is used as a statistical measure when the cost of false negatives is high. In the case of false-negative recognition of pigmented neoplasms, skin cancer will be recognized as a benign skin lesion, which is more dangerous when diagnosed in dermatology. The increase in the recall of each neural network model using weights was 0.02 compared to neural network models trained without weights.

Sensitivity is a measure of a model’s ability to identify a pigmented neoplasm as “malignant”. The best sensitivity index was obtained for the neural network model based on the ResNeXt CNN architecture and amounted to 0.90. The use of weight coefficients in training allowed us to improve the sensitivity index for each neural network model by 0.08–0.13 compared to models trained without the use of weight coefficients.

However, the most common statistics can be overly optimistic and inflated, especially on imbalanced datasets. The Matthews Correlation Coefficient (MCC) is a more robust statistic that only scores high if the model performs well in all four categories of the confusion matrix, proportional to the size of the positive elements, and the size of negative elements in the data set [22]. The best value of the Matthews coefficient was obtained from a neural network model based on the ResNeXt CNN architecture and amounted to 0.79. The use of weight coefficients in training allowed us to improve the value of the Matthews coefficient for each neural network model by 0.01–0.02 compared to models trained without the use of weight coefficients.

Figures 1, 2 and 3 show the confusion matrix obtained from testing neural network models based on various CNN architectures that were trained with and without weights. As a result of the analysis of the presented confusion matrix, it can be concluded that the use of weight coefficients as a modification of the cross-entropy loss function when

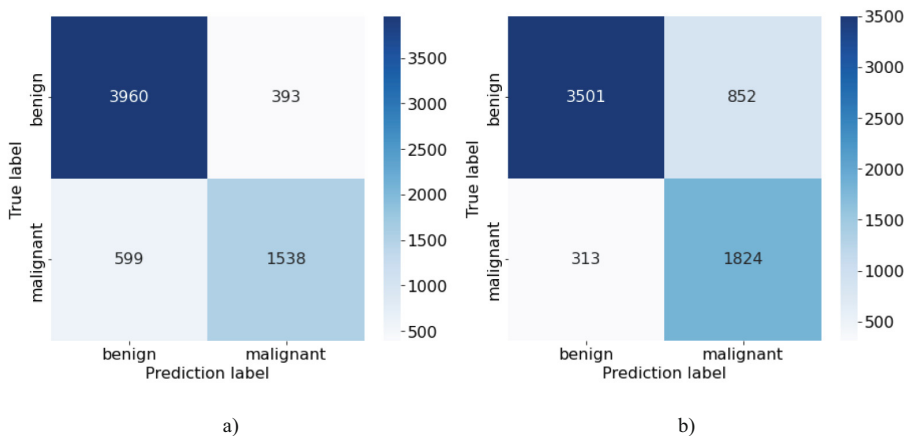


Fig. 1. Confusion matrix with the results of testing a neural network system for recognizing pigmented skin lesions based on CNN AlexNet: a) without the use of weight coefficients in training; b) using weight coefficients in training

training various CNN architectures improves the accuracy in recognizing a more significant class of “malignant” pigmented neoplasms. The use of weight coefficients makes it possible to reduce the number of false-negative recognitions of pigmented lesions.

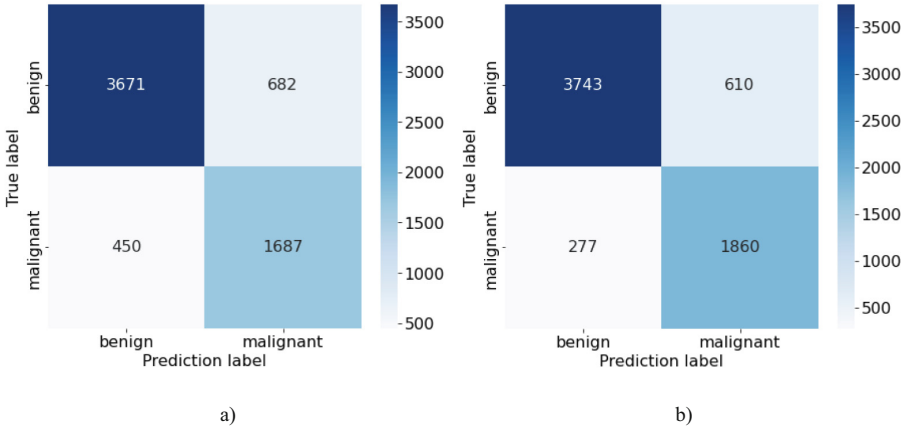


Fig. 2. Confusion matrix with the results of testing a neural network system for recognizing pigmented skin lesions based on CNN Inception_v4: a) without the use of weight coefficients in training; b) using weight coefficients in training

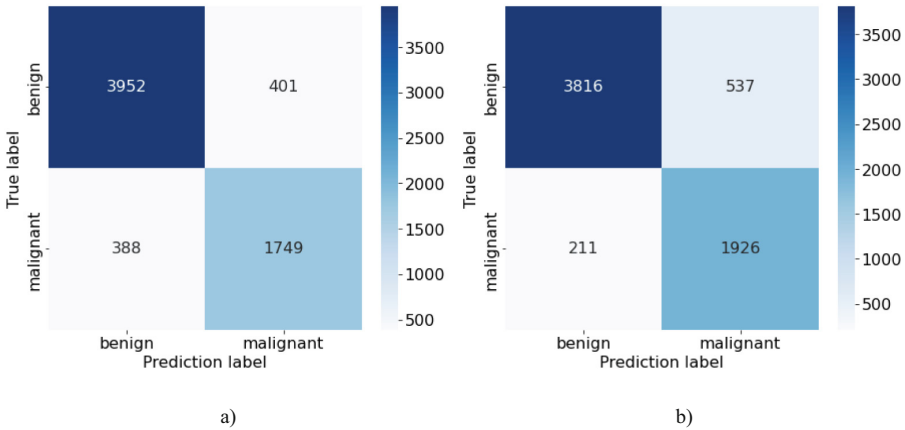


Fig. 3. Confusion matrix with the results of testing a neural network system for recognizing pigmented skin lesions based on CNN ResNeXt: a) without the use of weight coefficients in training; b) using weight coefficients in training

As a result of modeling taking into account the weighting coefficients of classes using the cross-entropy loss function, a neural network classifier was obtained that is sensitive to unbalanced dermoscopic images. At the same time, the best indicators of

accuracy, loss function, F1-score, Matthew's coefficient, recall, and sensitivity were obtained from models trained using weight coefficients. The best results in the accuracy of recognition of pigmented neoplasms, as well as in the recognition of malignant skin lesions, were obtained using the ResNeXt neural network architecture trained using weight coefficients. Modifying the cross-entropy loss function with weights showed that the neural network model is more punished for errors made in the minority class samples.

4 Discussion

As a result of training the neural network system using weight coefficients, the best recognition accuracy of malignant skin lesions was 88.12%. The use of a modified cross-entropy loss function increased the number of correctly classified malignant pigmented neoplasms and reduced the number of false-negative recognition cases.

Authors in [23] proposed a cost-effective CS-AF training method for unbalanced dermatological images. For training, cost matrices are used, which allow you to derive the objective weights of categories. The use of the CS-AF method made it possible to achieve an average accuracy of recognizing pigmented neoplasms on the ResNeXt neural network architecture of 79.20%, which is 8.92% lower than that of the proposed system based on a similar neural network architecture, which was trained using a modified cross-entropy function. For the Inception_v4 neural network architecture, the CS-AF method achieved an average accuracy of 76.40%, which is 10.95% lower than the accuracy of the proposed system based on a similar neural network architecture, which was trained using weight coefficients.

The work [8] presents a sensitive to unbalanced data (CoSen) deep neural network. To do this, the loss function was modified to include class-dependent costs during training. As a loss function, the CoSen function is used, which can be expressed as the average loss over the training sample. As a result of modeling on the Edinburgh Dermofit Image Library, the proposed CoSen CNN achieved an accuracy of 80.20%, which is 7.92% lower than the accuracy obtained by modeling a ResNeXt-based neural network system that was trained using a modified cross-entropy function.

The higher recognition accuracy of the proposed neural network system, which was trained using weight coefficients, in comparison with the results in [8, 23] is explained by the use of modified cross-entropy as a loss function. The cross-entropy loss function is the standard choice for high-performance neural networks.

The main limitation of using the proposed neural network system for recognizing pigmented lesions, trained using a modified loss function, is its use only as an additional diagnostic tool for a specialist. The proposed system cannot independently diagnose patients. A promising direction for further research is the construction of more complex neural network classification systems for pigmented skin lesions. The development of methods for preliminary processing of heterogeneous dermatological data will increase the accuracy and efficiency of recognition of pigmented neoplasms.

5 Conclusion

The article presents a neural network system for the recognition of malignant pigmented neoplasms, which was trained using a modified cross-entropy function. Modification of the cross-entropy loss function with the help of different weight coefficients has improved the accuracy and efficiency of recognition of pigmented neoplasms. When training CNN architectures using weighted cross-entropy loss, the accuracy improvement was 0.82–2.02% compared to the accuracy of CNN neural network architectures trained without using weights. The highest accuracy rate for recognizing pigmented skin lesions was achieved using CNN ResNeXt trained using weighting coefficients and amounted to 88.12%. The use of a modified cross-entropy function when training a neural network system made it possible to obtain a classifier that is sensitive to unbalanced dermoscopic images. Modification of the cross-entropy loss function by weight coefficients allowed to increase the number of recognized malignant neoplasms and reduce the number of false-negative classification errors.

The creation of systems for automated neural network recognition of pigmented skin lesions will reduce the consumption of financial and labor resources involved in the medical industry. At the same time, the creation of mobile systems for monitoring potentially dangerous skin tumors will help in making medical decisions and will automatically receive feedback on the condition of patients.

Acknowledgments. The research in section 3 was supported by the Council for grants of President of Russian Federation (project no. MK-3918.2021.1.6 and MK-371.2022.4). The authors would like to thank the North-Caucasus Federal University for supporting the contest of projects competition of scientific groups and individual scientists of the North-Caucasus Federal University.

References

1. Diepgen, T.L., Mahler, V.: The epidemiology of skin cancer. *Br. J. Dermatol.* **146**, 1–6 (2002). <https://doi.org/10.1046/J.1365-2133.146.S61.2.X>
2. Schadendorf, D., et al.: Melanoma. *The Lancet.* **392**, 971–984 (2018). [https://doi.org/10.1016/S0140-6736\(18\)31559-9](https://doi.org/10.1016/S0140-6736(18)31559-9)
3. Kittler, H., Pehamberger, H., Wolff, K., Binder, M.: Diagnostic accuracy of dermoscopy. *Lancet Oncol.* **3**, 159–165 (2002). [https://doi.org/10.1016/S1470-2045\(02\)00679-4](https://doi.org/10.1016/S1470-2045(02)00679-4)
4. Binder, M., et al.: Epiluminescence microscopy: a useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. *Arch. Dermatol.* **131**, 286–291 (1995). <https://doi.org/10.1001/ARCHDERM.1995.01690150050011>
5. Warshaw, E.M., et al.: Accuracy of teledermatology for pigmented neoplasms. *J. Am. Acad. Dermatol.* **61**, 753–765 (2009). <https://doi.org/10.1016/J.JAAD.2009.04.032>
6. Bratchenko, I.A., Alonova, M.V., Myakinin, O.O., Moryatov, A.A., Kozlov, S.V., Zakharov, V.P.: Hyperspectral visualization of skin pathologies in visible region. *Comput. Opt.* **40**, 240–248 (2016). <https://doi.org/10.18287/2412-6179-2016-40-2-240-248>
7. Liu, N., Li, X., Qi, E., Xu, M., Li, L., Gao, B.: A novel ensemble learning paradigm for medical diagnosis with imbalanced data. *IEEE Access.* **8**, 171263–171280 (2020). <https://doi.org/10.1109/ACCESS.2020.3014362>

8. Khan, S.H., Hayat, M., Bennamoun, M., Soheli, F.A., Togneri, R.: Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE Trans. Neural Netw. Learn. Syst.* **29**, 3573–3587 (2018). <https://doi.org/10.1109/TNNLS.2017.2732482>
9. Thai-Nghe, N., Gantner, Z., Schmidt-Thieme, L.: Cost-sensitive learning methods for imbalanced data. In: *Proceedings of the International Joint Conference on Neural Networks* (2010). <https://doi.org/10.1109/IJCNN.2010.5596486>
10. Zhang, L., Zhang, D.: Evolutionary cost-sensitive extreme learning machine. *IEEE Trans. Neural Netw. Learn. Syst.* **28**, 3045–3060 (2017). <https://doi.org/10.1109/TNNLS.2016.2607757>
11. He, H., Garcia, E.A.: Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **21**, 1263–1284 (2009). <https://doi.org/10.1109/TKDE.2008.239>
12. Zhang, C., Tan, K.C., Li, H., Hong, G.S.: A cost-sensitive deep belief network for imbalanced classification. *IEEE Trans. Neural Netw. Learn. Syst.* **30**, 109–122 (2019). <https://doi.org/10.1109/TNNLS.2018.2832648>
13. Mikołajczyk, A., Grochowski, M.: Data augmentation for improving deep learning in image classification problem. In: *2018 International Interdisciplinary PhD Workshop, IIPhDW 2018*, pp. 117–122 (2018). <https://doi.org/10.1109/IIPHDW.2018.8388338>
14. Perez, L., Wang, J.: *The Effectiveness of Data Augmentation in Image Classification using Deep Learning* (2017). <https://doi.org/10.48550/arxiv.1712.04621>
15. Jing, X.Y., et al.: Multiset feature learning for highly imbalanced data classification. *IEEE Trans. Pattern Anal. Mach. Intell.* **43**, 139–156 (2021). <https://doi.org/10.1109/TPAMI.2019.2929166>
16. Masnadi-Shirazi, H., Vasconcelos, N.: Cost-sensitive boosting. *IEEE Trans. Pattern Anal. Mach. Intell.* **33**, 294–309 (2011). <https://doi.org/10.1109/TPAMI.2010.71>
17. Yu, H., Sun, C., Yang, X., Zheng, S., Wang, Q., Xi, X.: LW-ELM: a fast and flexible cost-sensitive learning framework for classifying imbalanced data. *IEEE Access* **6**, 28488–28500 (2018). <https://doi.org/10.1109/ACCESS.2018.2839340>
18. Ryan Hoens, T., Chawla, N.V.: Imbalanced datasets: from sampling to classifiers. In: *Imbalanced Learning: Foundations, Algorithms, and Applications*, pp. 43–59 (2013). <https://doi.org/10.1002/9781118646106.CH3>
19. ISIC Archive. <https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main>
20. Alzubaidi, L., et al.: Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **8**(1), 1–74 (2021). <https://doi.org/10.1186/s40537-021-00444-8>
21. Hosny, K.M., Kassem, M.A., Foad, M.M.: Classification of skin lesions using transfer learning and augmentation with Alex-net. *PLoS ONE* **14**, e0217293 (2019). <https://doi.org/10.1371/JOURNAL.PONE.0217293>
22. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **21**, 1–13 (2020). <https://doi.org/10.1186/S12864-019-6413-7/TABLES/5>
23. Zhuang, D., Chen, K., Chang, J.M.: CS-AF: A cost-sensitive multi-classifier active fusion framework for skin lesion classification. *Neurocomputing* **491**, 206–216 (2022). <https://doi.org/10.1016/J.NEUCOM.2022.03.042>



Application of the SIFT Algorithm in the Architecture of a Convolutional Neural Network for Human Face Recognition

Diana Kalita^(✉)  and Parviz Almamedov 

North-Caucasus Federal University, Stavropol, Russia
diana.kalita@mail.ru

Abstract. Solving the problem of pattern recognition is one of the areas of research in the field of digital video signal processing. Recognition of a person's face in a real-time video data stream requires the use of advanced algorithms. Traditional recognition methods include neural network architectures for pattern recognition. To solve the problem of identifying singular points that characterize a person's face, this paper proposes a neural network architecture that includes the method of scale-invariant feature transformation. Experimental modeling showed an increase in recognition accuracy and a decrease in the time required for training in comparison with the known neural network architecture. Software simulation showed reliable recognition of a person's face at various angles of head rotation and overlapping of a person's face. The results obtained can be effectively applied in various video surveillance, control and other systems that require recognition of a person's face.

Keywords: face recognition · neural network · SIFT method · feature point descriptor · recognition accuracy

1 Introduction

The widespread use of video surveillance systems in various spheres of human activity entails an increase in the tasks associated with the processing of a digital video stream. One of the actual processing tasks is the task of analyzing a digital signal in video surveillance systems, which consists in visual tracking [1, 2], recognition [3, 4], control [5], etc.

In this regard, the tasks associated with the search for a moving object in the video data stream and face recognition are practically significant. For this reason, there is a search for new, better ones, as well as the improvement of known face recognition algorithms. The authors of [6] proposed a combination of Haar features and a classifier based on skin color for face detection. In [7], a real-time automatic face recognition system is being developed. The system uses the MTCNN algorithm for recognition, and uses a generative confrontation network to align faces from the effects of posture changes. Comparison of objects occurs with the help of cosine similarity. However, the

face detection and recognition algorithm can only extract low-level facial features from an image, and when processing a large number of images, it is difficult to extract a description of facial features with a higher degree of differentiation.

For this reason, a large number of researchers use convolutional neural networks as the main computing architecture to solve the task of recognition. The human face detection algorithm based on the Viola-Jones method and the convolutional neural network architecture is used in [8]. The authors of [9] use the DeepFace algorithm, which is first used in the architecture of convolutional neural networks for face recognition. The algorithm creates weights for each face element, optimizing intra-class differences with respect to inter-class similarity scores. In [10], a deep universal architecture is proposed that simultaneously performs the synthesis and recognition of faces of different ages. [11] propose a DeepID algorithm that first extracts deep facial features using a simple convolutional neural network structure and then improves the measurement of deep identification characteristics. Finally, two different methods are used in face verification. [12] propose a face recognition algorithm based on the CNN method and the TensorFlow deep learning platform for face recognition tasks with multiple poses and occlusion. However, most of the databases are small, which will cause the recognition accuracy of most face recognition algorithms to quickly decrease, and besides, it cannot meet the actual requirements of the application.

In 2004, Low's researcher proposed the SIFT algorithm [13] based on local features, which is an approach to extract distinctive invariant features from images. In recent years, an algorithm based on the invariant descriptor has been actively studied and successfully applied for object recognition, pose estimation, image search, etc. It has been proven that SIFT is the most locally invariant function descriptor, it is invariant to rotation, translation and scale changes. Since the first step in object recognition is to find image feature points, which are a kind of marks for subsequent stages of recognition, and taking into account the advantages of the SIFT algorithm, in this paper we apply this algorithm to solve the problem of human face recognition in a neural network architecture.

2 Materials and Methods

2.1 Background CNN

To solve the recognition problem, the developed neural network must be trained in two tasks:

1. Face Classification by Classes: In this article, this is treated as a binary classification problem and the cross entropy loss function is used in training.

$$D_j^{det} = -\left(z_j^{det}(p_j) + (1 - z_j^{det})(1 - \log(p_j))\right) \quad (1)$$

where p_j – the actual output value of the neural network, z_j^{det} – expected output value and classification loss.

2. Location of special points. At the first step of calculations, the descriptor of singular points is searched for all scales and positions of the image. The determination of potential key points that do not depend on scale and orientation is done using the Gaussian difference function. In this case, key points are selected based on their stability indicators. For stable keypoints, it is not sufficient to discard keypoints with low contrast. The Gaussian difference function will have a strong response at the edges. If key points are located on the edge, they are unstable. The difference detector detects points in Laplace scale space. The image scale space is defined as a function $L(x, y, \sigma)$ (1), which is obtained by convolving the variable scale Gaussian $G(x, y, \sigma)$, for the input image $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \quad (2)$$

The scale-space extrema, which are extracted from the Gaussian difference function convolved with the image, $F(x, y, \sigma)$ is defined as:

$$F(x, y, \sigma) = (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) = L(x, y, k\sigma) - L(x, y, \sigma) \quad (3)$$

where k is a constant multiplicative factor. After calculating the key point, the point in its neighborhood is compared, the key point is compared with its 26 neighboring points as $3 \times 3 \times 3$. If a pixel is a local maximum or minimum, it is selected as the key point.

To obtain descriptors that are invariant to the orientation of the keypoints, one or more orientations are assigned to each keypoint location based on the local gradient directions of the image. All subsequent operations are performed on image data that has been transformed with respect to the specified orientation, scale, and location for each object, thereby ensuring invariance to these transformations. The work uses the gradient value $m(x, y)$ and the orientation $\theta(x, y)$, which are calculated using the pixel difference:

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \quad (4)$$

$$\theta(x, y) = \tan^{-1}((L(x, y+1) - L(x, y-1)) / (L(x+1, y) - L(x-1, y))) \quad (5)$$

Based on the orientation of the sample points gradient in the area around the characteristic orientation point, a histogram is formed. The peaks in the orientation histogram correspond to the dominant directions of the local gradients.

Local image gradients are measured at the selected scale in the area around each key point. They are converted to a representation that allows for significant levels of local shape distortion and lighting changes. According to the above operation, the image location, scale and orientation are assigned to each key point. A key point descriptor is created by calculating the gradient value and orientation at each image sample point in the area around the key point location. To reduce the effect of lighting changes, the vector is normalized.

The structure of the developed network, which extracts the distinctive features of a person's face, is shown in Fig. 1 (a). The input data is a color image of the JPEG type, with a size of 48x48 pixels. Each image is divided into 3 channels: red, blue, green. Thus, we get 3 images with dimensions of 48x48 pixels. The first layers highlight the features of a person's face based on the SIFT method. The last layers of the neural network classify the image according to the neurons that determine the image class (Fig. 1(b)).

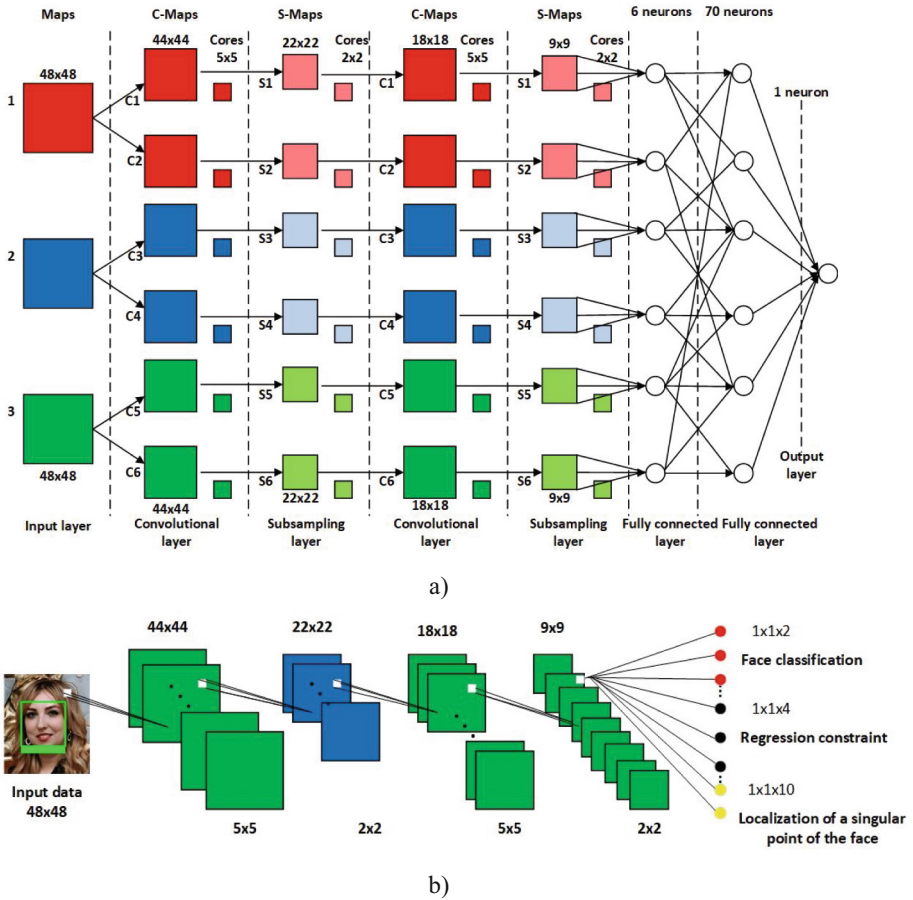


Fig. 1. Architecture of the proposed convolutional neural network a) network; b) the process of recognizing a person’s face in the architecture of the developed network structure

3 Results and Discussion

Based on the two identified tasks necessary to solve the problem of human face recognition, the general procedure of the algorithm proposed in this paper is shown in Fig. 2.

The entire algorithm for searching and recognizing a person’s face can be divided into three stages:

Stage 1: The neural network extracts the face image from the input images and then performs a face search.

Stage 2: The SIFT method searches for singular points, after which these points are written to the file in the form of a matrix.

Step 3: Face verification is done by comparing the total distance between the current input image and the one recorded in the file. For example, if the distance between the

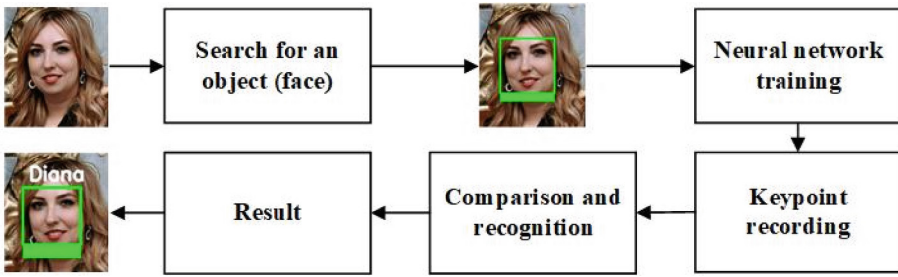


Fig. 2. Image recognition process in the developed neural network

joints is less than or equal to 0.85, which is the learning threshold, the system determines that the face belongs to the same person. On the contrary, if the distance between the joints is greater than 0.85, it is considered different people.

We will show the work of the developed neural network architecture using software simulation.

For software simulation, a neural network implemented in the Python 3.8 programming language was used. The training of the developed neural network consisted of positive and negative examples. In this case, from persons and “not persons”. The ratio of positive to negative examples is 4 to 1, 8000 positive and 2000 negative. The LFW3D database was used as a positive sample, which contains color images of frontal faces of the JPEG type, 90×90 pixels in size, in the amount of 13,000. The SUN397 database was used as negative training examples, it contains a total of 130,000 images. The database size is 37 GB. After training, the developed neural network determines faces with an accuracy of 95%.

Table 1 presents the results of the time required for training the developed neural network and the known neural network [12]. The results showed that the developed network requires less time for training at any resolution of the input images. In this case, there is a direct dependence on the resolution and training time for both neural networks.

Table 1. Dependence of training time on image resolution

Neural network architecture	Indicators	Image resolution			
		950×1280	960×1280	1620×2160	1920×1080
Proposed architecture	Studying time	0.56	0.86	1.23	1.58
	Recognition accuracy (%)	92	92.5	94	95
[12]	Studying time	1.19	2.64	3.59	4.81
	Recognition accuracy (%)	89	89.7	90.5	92

The Fig. 3 shows the dependence of the training time and the accuracy of the recognition at the input of the neural network of the image. Dependency plots show that training accuracy increases with image resolution, however, to achieve greater accuracy, it takes more time to train the neural network.

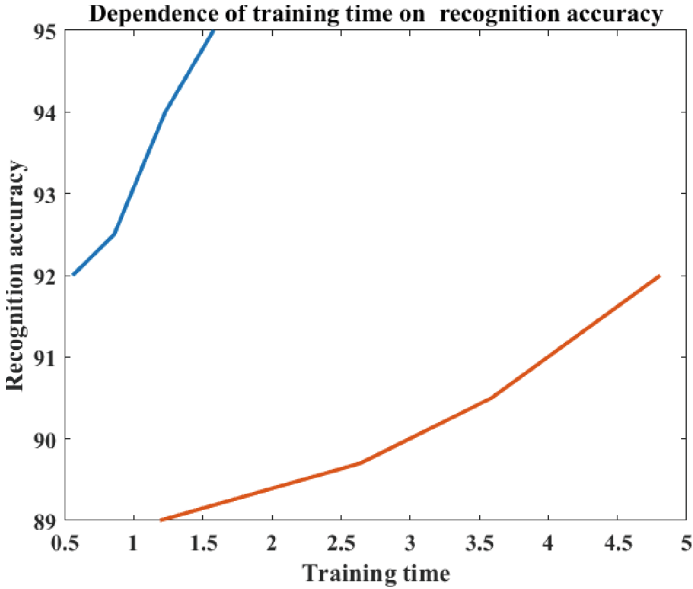


Fig. 3. Dependence of training time on recognition accuracy

Figure 4 shows the results of recognition of the human face “Diana” by the neural network based on the application of the algorithm for determining key points in the network structure.



Fig. 4. The result of recognition of the person’s face “Diana”

To improve the reliability of the neural network, this work used human face recognition under various conditions, such as rotation, mirroring, zooming. Figure 5 shows the results of detecting the face of one person in different poses (for example, straight face, look up, look down, face from the side) using the developed neural network. As can be seen from the results presented in the figure, the neural network accurately recognizes a person's face at all angles of face rotation.

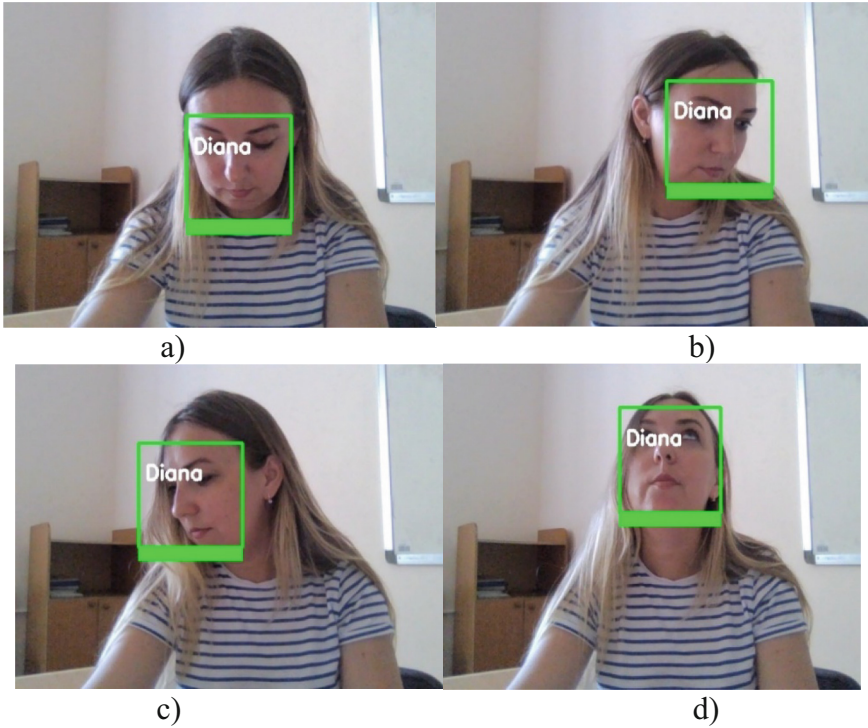


Fig. 5. Human face detection results at different head rotation angles a) straight face; b) look up; c) look down; d) face from the side

Figure 6 shows the results of human face detection at different percentages of face overlap (10/20/30/40%). The results show that the neural network can still accurately detect the face when part of the face is covered.

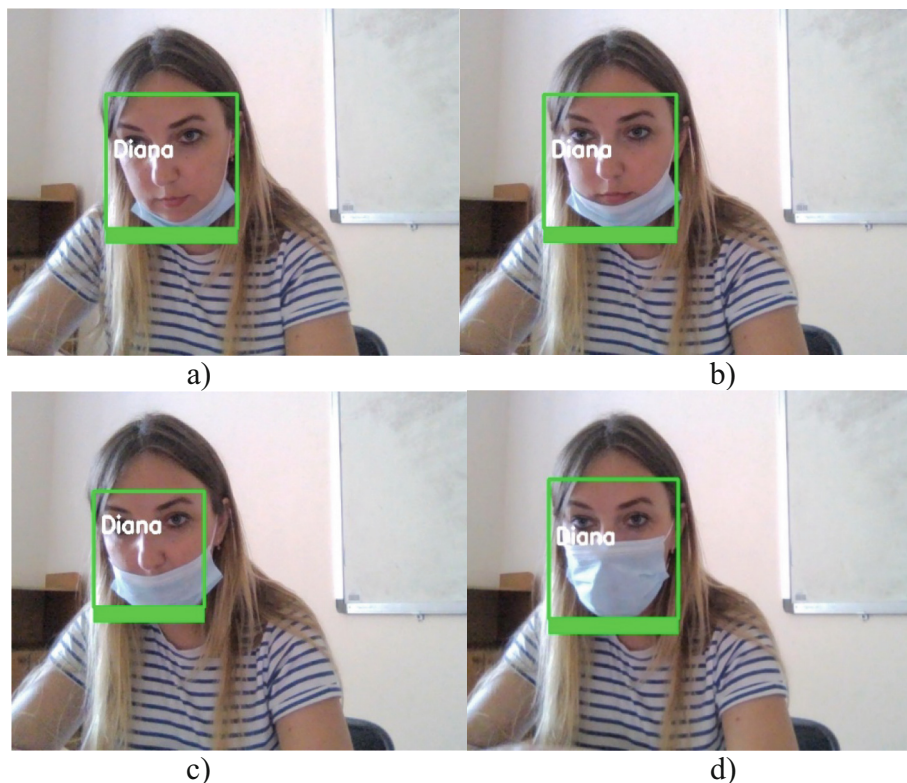


Fig. 6. Results of human face detection at various percentages of face overlap a) 10%; b) 20%; c) 30%; d) 40%

4 Conclusion

Face detection and recognition are important tasks that require solutions in video surveillance systems, access control systems and other systems that require recognition of a person's face. In order to solve the problem that traditional algorithms cannot fully extract facial features from an image with a complex background, this paper proposes a face detection algorithm based on the architecture of a convolutional neural network and the SIFT method for fast face detection from an image with a complex background.

Software simulation showed that the developed neural network is able to recognize a person's face under various conditions of face tilt angle, as well as under conditions of face overlap. Experimental results confirm that the recognition accuracy of the developed neural network is 3% higher compared to the known architecture. Moreover, the recognition speed of the algorithm proposed in this article is also faster than other algorithm, which is in line with real-time requirements. Thus, the algorithm proposed in this paper is of high practical value in video surveillance, access control, and other systems that are of great importance for improving security performance.

Further research will be directed to the development of a neural network that searches for a moving object in a video data stream with various noise components in the video stream.




Acknowledgments. The authors thank the North-Caucasus Federal University for supporting in the contest of projects competition of scientific groups and individual scientists of North-Caucasus Federal University.

References

1. Marvasti-Zadeh, S.M., Cheng, L., Ghanei-Yakhdan, H., Kasaei, S.: Deep learning for visual tracking: a comprehensive survey. *IEEE Trans. Intell. Transp. Syst.* **23**(5), 3943–3968 (2022)
2. Li, K., Kong, Y., Fu, Y.: Visual object tracking via multi-stream deep similarity learning Networks. *IEEE Trans. Image Process.* **29**, 3311–3322 (2020)
3. Zhang, X.-Y., Liu, C.-L., Suen, C.Y.: Towards robust pattern recognition: a review. *Proc. IEEE* **108**(6), 894–922 (2020)
4. He, L., Li, H., Zhang, Q., Sun, Z.: Dynamic feature matching for partial face recognition. *IEEE Trans. Image Process.* **28**(2), 791–802 (2019)
5. Ke-Qiang, X., Meng, X., Jun, W., Bao-Jun, L., Zhuo, C., Gan-Hua, L.: A new method for satellite control system fault pattern recognition combining multi-classification SVM with kernel principal component analysis. In: 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), pp. 794–797 (2019)
6. Tu, Y., Yi, F., Chen, G., Jiang, Sh., Huang, Zh.: Fast rotation invariant face detection in color image using multi-classifier combination method. In: 2010 International Conference on E-Health Networking Digital Ecosystems and Technologies (EDT), pp. 211–218 (2010)
7. Liu, Q.: Face matching system in multi-pose changing scene. In: 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), pp. 492–495 (2021)
8. Patil, P., Shinde, S.: Comparative analysis of facial recognition models using video for real time attendance monitoring system. In: 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 850–855 (2020)
9. Srisuk, S., Ongkittikul, S.: Robust face recognition based on weighted DeepFace. In: 2017 International Electrical Engineering Congress (iEECON), pp. 1–4 (2017)
10. Zhao, J., Yan, S., Feng, J.: Towards age-invariant face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(1), 474–487 (2022)
11. Yang, B., Cao, J., Ni, R., Zhang, Y.: Facial expression recognition using weighted mixture deep neural network based on double-channel facial images. *IEEE Access* **6**, 4630–4640 (2018)
12. Janahiraman, T., Subramaniam, P.: Gender classification based on Asian faces using deep learning. In: 2019 IEEE 9th International Conference on System Engineering and Technology (ICSET), pp. 84–89 (2019)
13. Chen, C., Mu, Z.: An improved image registration method based on SIFT and SC-RANSAC algorithm. In: 2018 Chinese Automation Congress (CAC), pp. 2933–2937 (2018)



High-Speed Wavelet Image Processing Using the Winograd Method

N. N. Nagornov¹ , N. F. Semyonova¹ , and A. S. Abdulsalyamova² 

¹ North-Caucasus Federal University, Stavropol, Russia
sparta1392@mail.ru

² North-Caucasus Center for Mathematical Research, Stavropol, Russia

Abstract. Wavelets are actively used for solving of image processing problems in various fields of science and technology. Modern imaging systems have not kept pace with the rapid growth in the amount of digital visual information that needs to be processed, stored, and transmitted. Many approaches are being developed and used to speed up computations in the implementation of various image processing methods. This paper proposes the Winograd method (WM) to speed up the wavelet image processing methods on modern microelectronic devices. The scheme for wavelet image filtering using WM has been developed. WM application reduced the computational complexity of wavelet filtering asymptotically to 72.9% compared to the direct implementation. An evaluation based on the unit-gate model showed that WM reduces the device delay to 66.9%, 73.6%, and 68.8% for 4-, 6-, and 8-tap wavelets, respectively. Revealed that the larger the processed image fragments size, the less time is spent on wavelet filtering, but the larger the transformation matrices size, the more difficult their compilation and WM design on modern microelectronic devices. The obtained results can be used to improve the performance of wavelet image processing devices for image compression and denoising. WM hardware implementation on a field-programmable gate arrays and an application-specific integrated circuits to accelerate wavelet image processing is a promising direction for further research.

Keywords: Wavelet Transform · Digital Filtering · Group Pixel Processing · Computational Complexity · High-Speed Calculations · High-Performance Computing

1 Introduction

Wavelets are actively used for solving image processing problems in various fields of science and technology such as denoising [1], color image processing [2], video analysis [3]. However, modern imaging systems have not kept pace with the rapid growth in the amount of digital visual information that needs to be processed, stored, and transmitted. Many approaches are being developed and used to speed up computations in the implementation of various image processing methods. The authors of [4] focus on the evolution and application of various hardware architectures. The fast decomposition algorithms based on a different representation called product convolution extension has

been proposed in [5]. This decomposition can be efficiently estimated by assuming that multiple operator impulse responses are available. The new simple adjacent sum method is developed in [6] for multidimensional wavelet constructing. This method provides a systematic way to build multidimensional non-separable wavelet filter banks from two 1D low-pass filters, one of which is interpolating to increase image processing speed. The authors of [7] describe an asymmetric 2D Haar transform and extend it to wavelet packets containing an exponentially large number of bases. A basis selection algorithm is also proposing for optimal basis finding in wavelet packets. Various modern GPU optimization strategies for the discrete wavelet transform implementation such as the use of shared memory, registers, warp shuffling instructions, and parallelism at the level of threads and instructions are presented in [8]. A mixed memory structure for the Haar transform is proposed in which a multilevel transform can be performed with a single launch of the combined kernel. The paper [9] proposes a new algorithm for 2D discrete wavelet transform of high-resolution images on low-cost visual sensors and nodes of the Internet of things. The reduction in computational complexity and power consumption compared to modern low-memory 2D discrete wavelet transform methods are the main advantages of the proposed segmented modified fractional wavelet filter. However, all of these methods are based on pixel-by-pixel image processing. The Winograd method (WM) reduce image processing time due to group pixel processing. The processed image is assembled from fragments of a certain size which reduces the multiplications number by increasing the additions number.

The purpose of this paper is to accelerate wavelet image processing using WM on modern microelectronic devices.

2 Wavelet Image Processing Using the Direct Implementation and the Winograd Method

Wavelet filtering using direct implementation (DI) has the form

$$I_2(x) = \sum_{i=0}^{f-1} I_1(x-i)K(i), \quad (1)$$

where I_1 and I_2 are the original and processed 2D images, respectively, x is the row number of the pixel processed by f -tap wavelet filter K . The wavelet transform extracts local information about the signal in both frequency and time. High computational complexity is a significant disadvantage of this transform. The scheme of 1D wavelet filtering of an image fragment using DI is shown in Fig. 1a, where S_I is the original image fragment, L and H are the low- and high-pass wavelet filters, P_A and P_D are the processed image pixels with approximate and detailing image information, respectively.

Image filtering using WM in matrix form [10] can be presented as

$$Z = A^T \left((GK) \odot (B^T S) \right), \quad (2)$$

where: Z is the processed image fragment of size $z \times 1$; K is the wavelet filter of size $f \times 1$; S is the original image fragment of size $s \times 1$, where $s = z + f - 1$; A^T , G , B^T are the transformation matrices of sizes $z \times s$, $s \times f$, $s \times s$, respectively; \odot is

the element-wise matrix multiplication. Algorithms for matrices A^T , G , B^T obtaining are described in [11]. WM is denoted as $F(z, f)$. Digital filtering is performed on two computational channels corresponding to low- and high-frequency wavelet filters during wavelet image processing. The products of GL and GH are calculated in advance when using a specific wavelet. The product of S and the transformation matrix B^T can be computed before splitting the calculations into two channels because does not depend on the wavelet choice. Next, the element-wise multiplications $B^T S$ by GL and GH and the products of the obtained results with the transformation matrix A^T are performed over two computational channels. The scheme of 1D wavelet filtering of an image fragment using WM is shown in Fig. 1b, where S is the original image fragment, L and H are the low- and high-frequency wavelet filters, B^T , A^T , G are the transformation matrices, S_A and S_D are the processed image fragments with approximate and detailing image information, respectively.

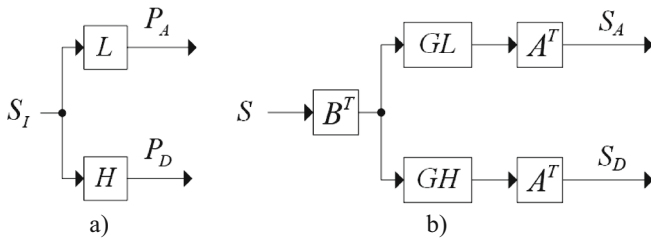


Fig. 1. The schemes of 1D wavelet filtering of an image fragment using: a) the direct implementation; b) the Winograd method

The results of increasing the speed of wavelet image processing using MW are presented below.

3 Acceleration of Wavelet Image Processing Using the Winograd Method

The computational complexity in time of wavelet filtering using WM $F(z, f)$ depends on the z and f and on the choice of points $s_0, s_1, \dots, s_{n-2}, s_{n-1}$. These values determine the form of transformation matrices A^T, G, B^T . The set of the Lagrange polynomial points $L = 0, 1, -1, 2, -2, 4, -4, \dots, 2^l, -2^l, 2^{l+1}, -2^{l+1}, \dots, \infty$ was used to construct the Vandermonde matrix V and matrices A^T, G, B^T [11]. The cases of using 4-, 6-, and 8-tap wavelets and processing of the original image fragments with size $z = 2, 3, 4, 5, 6, 7$ are considered. Table 1 is based on transformation matrices and contains the counting results of the multiplications and additions number required for wavelet filtering of images using DI and WM. The table values are obtained as follows.

1. The DI multiplications number is equal to the wavelet filters coefficients number.
2. The WM multiplications number is equal to twice the number of the processed image fragment pixels.

3. The DI additions number is equal to the number 2 less than the multiplications number.
4. The WM basic additions number is equal to the additions number of nonzero elements of matrices A^T (twice) and B^T by rows.
5. The WM complementary additions number is equal to the sum of the matrix element units in binary notation reduced by 1 for all elements of matrices A^T (twice) and B^T .
6. The total additions number is equal to the sum of basic and complementary additions.
7. WM receives several pixel values of the processed image in one iteration. Obtaining pixel brightness value requires the entire iteration as well as obtaining the entire fragment. Introduce the pixel specific value (PSV) for a correct comparison of the methods computational complexity. PSV is calculated as a quotient of the required operations number (multiplications or additions) divided by the number of pixels in the processed image fragment.

Table 1 shows that the greatest reduction in the specific weight of a pixel by multipliers is observed for 8-tap wavelet using WM $F(6, 8)$. The computational complexity decreases asymptotically by 72.9% compared to DI. The asymptotic estimate does not take into account addition operations since their complexity is an order of magnitude less than multiplication. This assessment is predominantly theoretical and may have a low correlation with the results obtained in the design of wavelet image processing devices in practice. Therefore, the unit-gate model (UGM) was used to calculate the operating time of a microelectronic device. UGM is a method for theoretical evaluation of device characteristics based on counting the number of the basic logical elements “and”, “or” [12]. The response time of one such element will be taken as a conventional unit (CU). Describe the principles of performing calculations in the theoretical estimation of the wavelet filtering devices delay according to the schemes in Fig. 1a and Fig. 1b for DI and WM, respectively. All multiplications are performed in parallel when using both methods.

Matrix multiplication operations can be replaced by shift and addition operations using the B^T and A^T matrices. The number of ones in the number binary representation for each element of the matrices A^T and B^T was calculated to determine the terms number in the rows of these matrices (Table 2). The products GL and GH are performed a priori. The products $B^T S$ on GL and GH are realized by element-wise multiplications. Multiplications and additions are implemented using a generalized multiplier (GM) and a multi-operand adder (MOA), respectively [13]. The delays of GM and MOA for k -bit numbers on computing devices are $6.8 \log_2 N + 2 \log_2 k + 4$ and $8.8 \log_2 k + 4$, respectively [14], where N is the largest number of elements in rows of matrices A^T and B^T , k is the image color depth and the coefficients bit-width of used wavelet filters. The calculations are performed for $k = 8$. The results of the device delay evaluation for wavelet image processing using DI and WM are presented in Table 2.

Table 1. The number of additions and multiplications in wavelet filtering of an image fragment using the direct implementation and the Winograd method

Tap	Method	Fragment pixels	For each fragment			Pixel specific value			
			Multiplications		Total	Additions		Total	
			Basic	Complementary		Basic	Complementary		
4	Direct	1	8	0	6	8.0	6.0	0.0	6.0
	$F(2,4)$	2	10	1	24	5.0	11.5	0.5	12.0
	$F(3,4)$	3	12	2	40	4.0	12.7	0.7	13.3
	$F(4,4)$	4	14	14	77	3.5	15.8	3.5	19.3
	$F(5,4)$	5	16	20	110	3.2	18.0	4.0	22.0
6	Direct	1	12	0	10	12.0	10.0	0.0	10.0
	$F(2,6)$	2	14	14	61	7.0	23.5	7.0	30.5
	$F(3,6)$	3	16	20	90	5.3	23.3	6.7	30.0
	$F(4,6)$	4	18	64	169	4.5	26.3	16.0	42.3
	$F(5,6)$	5	20	84	222	4.0	27.6	16.8	44.4
8	$F(6,6)$	6	22	197	382	3.7	30.8	32.8	63.7
	$F(7,6)$	7	24	248	478	3.4	32.9	35.4	68.3
	Direct	1	16	0	14	16.0	14.0	0.0	14.0
	$F(2,8)$	2	18	59	140	9.0	40.5	29.5	70.0
	$F(3,8)$	3	20	84	194	6.7	36.7	28.0	64.7
8	$F(4,8)$	4	22	197	350	5.5	38.3	49.3	87.5
	$F(5,8)$	5	24	248	442	4.8	38.8	49.6	88.4
	$F(6,8)$	6	26	505	756	4.3	41.8	84.2	126.0

Table 2. UGM-based evaluation results of the device delay for wavelet processing of 8-bit image using the direct implementation and the Winograd method

Tap	Method	Fragment pixels	The largest number of elements in a matrix row		Processing time according UGM	
			A^T	B^T	For each fragment	Pixel specific value
4	Direct	1	–	–	55.0	55.0
	$F(2, 4)$	2	4	4	78.6	39.3
	$F(3, 4)$	3	5	4	80.8	26.9
	$F(4, 4)$	4	6	9	90.5	22.6
	$F(5, 4)$	5	7	8	90.9	18.2
6	Direct	1	–	–	59.0	59.0
	$F(2, 6)$	2	6	9	90.5	45.3
	$F(3, 6)$	3	7	8	90.9	30.3
	$F(4, 6)$	4	8	18	100.2	25.0
	$F(5, 6)$	5	9	16	100.2	20.0
	$F(6, 6)$	6	10	34	108.6	18.1
	$F(7, 6)$	7	11	32	108.9	15.6
8	Direct	1	–	–	61.8	61.8
	$F(2, 8)$	2	8	18	100.2	50.1
	$F(3, 8)$	3	9	16	100.2	33.4
	$F(4, 8)$	4	10	34	108.6	27.2
	$F(5, 8)$	5	11	32	108.9	21.8
	$F(6, 8)$	6	12	58	115.6	19.3

The following conclusions are drawn based on the results in Table 2.

1. WM reduced the device delay of wavelet image processing to 66.9%, 73.6%, and 68.8% for 4-, 6-, and 8-tap wavelets, respectively, compared DI according to UGM.
2. The larger the processed image fragments size z , the less time is spent on wavelet filtering, but the larger the transformation matrices size, the more difficult their compilation and WM design on modern microelectronic devices.
3. The greatest reduction in device delay with an increase in the size of the resulting image fragments processed using WM is achieved at $z = 2$ and $z = 3$ according to UGM. For example, the device delay is reduced by $55.0 - 39.3 = 15.7$ CU and $39.3 - 26.9 = 12.4$ CU at $z = 2$ and $z = 3$, respectively, the device delay is reduced by 4.3 CU and 4.5 CU at $z = 4$ and $z = 5$, respectively, for 4-tap wavelet according to UGM.

4 Conclusion

The scheme for 1D wavelet image processing using WM has been developed. A comparative analysis of the image filtering time with DI was carried out. WM reduced the computational complexity of wavelet image processing asymptotically to 72.9% depending on the size of the filters used and fragments of the processed image. WM reduced the device delay of wavelet image processing to 66.9%, 73.6%, and 68.8% for 4-, 6-, and 8-tap wavelets, respectively, according to UGM. The larger the processed image fragments size z , the less time is spent on wavelet filtering, but the larger the transformation matrices size, the more difficult their compilation and WM design on modern microelectronic devices. The obtained results can be used to improve the performance of wavelet image processing devices for image compression and denoising. WM hardware implementation on FPGAs and ASICs to accelerate wavelet image processing is a promising direction for further research.

Acknowledgments. The work was supported by Russian Science Foundation, project № 22-71-00009.

References

1. Wu, Y., Gao, G., Cui, C.: Improved wavelet denoising by non-convex sparse regularization under double wavelet domains. *IEEE Access* **7**, 30659–30671 (2019)
2. Souillard, R., Carré, P.: Elliptical monogenic wavelets for the analysis and processing of color images. *IEEE Trans. Sig. Process.* **64**, 1535–1549 (2016)
3. Chen, Y., Li, D., Zhang, J.Q.: Complementary color wavelet: a novel tool for the color image/video analysis and processing. *IEEE Trans. Circuits Syst. Video Technol.* **29**, 12–27 (2019)
4. Alcaín, E., et al.: Hardware architectures for real-time medical imaging. *Electronics* **10**, 3118 (2021)
5. Escande, P., Weiss, P.: Fast wavelet decomposition of linear operators through product-convolution expansions. *IMA J. Numer. Anal.* **42**, 569–596 (2022)
6. Hur, Y., Zheng, F.: Prime coset sum: a systematic method for designing multi-d wavelet filter banks with fast algorithms. *IEEE Trans. Inf. Theory*, 7565569 (2016)
7. Ouyang, W., Zhao, T., Cham, W.K., Wei, L.: Fast full-search-equivalent pattern matching using asymmetric haar wavelet packets. *IEEE Trans. Circuits Syst. Video Technol.* **28**, 819–833 (2018)
8. Quan, T.M., Jeong, W.K.: A fast discrete wavelet transform using hybrid parallelism on GPUs. *IEEE Trans. Parallel Distrib. Syst.* **27**, 3088–3100 (2016)
9. Tausif, M., Khan, E., Hasan, M., Reisslein, M.: SMFrWF: segmented modified fractional wavelet filter: fast low-memory discrete wavelet transform (DWT). *IEEE Access* **7**, 84448–84467 (2019)
10. Winograd, S.: *Arithmetic Complexity of Computations* (1980)
11. Lyakhov, P., Abdulsalyamova, A., Semyonova, N., Nagornov, N.: On the computational complexity of 2D filtering by Winograd method. In: 2022 11th Mediterranean Conference on Embedded Computing (MECO), pp. 1–4 (2022)

12. Zimmermann, R.: Binary adder architectures for cell-based VLSI and their synthesis. Hartung-Gorre (1998)
13. Parhami, B.: Computer Arithmetic: Algorithms and Hardware Designs. Oxford University Press, Oxford (2010)
14. Lyakhov, P., Valueva, M., Valuev, G., Nagornov, N.: A method of increasing digital filter performance based on truncated multiply-accumulate units. *Appl. Sci.* **10** (2020)



Improving the Parallelism and Balance of RNS with Low-Cost and $2^k + 1$ Modules

Pavel Lyakhov^{1,2} 

¹ Department of Mathematical Modeling, North-Caucasus Federal University, Stavropol, Russia
ljahov@mail.ru

² North-Caucasus Center for Mathematical Research, North-Caucasus Federal University, Stavropol, Russia

Abstract. In this paper, extending the set of low-cost modules with modules of $2^k + 1$ form, which are considered next in the complexity of implementation after modules of $2^k - 1$ proposed to consider. An experiment was carried out on the construction of various RNSs for the ranges from 16 to 64 bits and the number of modules from 3 to 8. The results of the experiment showed a significant increase in parallelism and improvement in the balance of RNS with modules of the 2^k , $2^k - 1$ and $2^k + 1$ forms, compared to using only modules of the 2^k and $2^k - 1$ forms. In the discussion of the experiment, we raised the questions of more effective measurement of the balance of the RNS, as well as the problem of using the diminished-1 encoding. This problem is a subject to a thorough study both at the theoretical level, for example, using a unit-gate model, and practical tests on modern FPGA and ASIC microelectronic devices.

Keywords: Residue Number System · Parallel Computing · Digital Signal Processing · Computer Arithmetic

1 Introduction

The exponential growth of the amount of information processed in modern computing systems encourages researchers to look for fundamentally new approaches to the organization of calculations. One of the most promising alternatives to the traditional representation of information in the binary number system (BNS) is the use of the Residue Number System (RNS) [1]. The essence of RNS is to replace the multi-digit representation of numbers with a set of low-digit remainders from division into pairwise coprime RNS modules, thanks to the Chinese remainder theorem. This approach allows to significantly speed up the execution of addition, subtraction and multiplication operations at the cost of slowing down the comparison and division operations, as well as the need to add converters for converting data from BNS to RNS and back. Currently, research is underway on the use of RNS in the problems of digital signal processing [2], images and video [3], cryptography [4, 5], and machine learning [6].

Features of the number-theoretic representation of information in RNS give rise to the problems of finding effective solutions for converting data from the RNS to BNS [7],

as well as for implementing calculations in RNS [8, 9]. The need to calculate modular operations imposes additional requirements on hardware implementations of circuits in RNS. 2^k is the simplest module for implementing calculations since, in this case, it is enough to discard simply the most significant bits of the number during the calculations. The next complexity of the calculation's organization is the $2^k - 1$ module. Efficient methods of hardware implementation have been developed for such a module, based on the cyclic transfer of the most significant bits to the least significant bits of the number. The listed modules are called "low-cost" modules [10]. Unfortunately, the choice of low-cost modules is relatively small. In addition, the need for pairwise mutual simplicity leads to the fact that in the case of choosing a large number of RNS modules, they will all have different bit widths, and, as a result, different operation times, which leads to inefficient use of the computing system time resources. This problem is generated by the imbalance of RNS modules, the question of which was first raised in [11]. In this paper, extending the set of low-cost modules with modules of $2^k + 1$ form, which are considered next in the complexity of implementation after modules of the form $2^k - 1$ is proposed to consider. The research will show a significant improvement in the balance and level of parallelism of RNS due to such manipulation and will outline a plan for further research on the development of efficient hardware solutions based on such RNS.

2 Measuring the Balance of RNS with Low-Cost Modules and Modules of the Form $2^k + 1$

The major advantage of RNS, which makes it an attractive alternative to the traditional binary number system, is the possibility of parallel execution of arithmetic operations of addition, subtraction and multiplication for each of the system modules. In practice, this means that there is no information transfer between the channels responsible for different RNS modules when performing these operations. However, RNS modules cannot be chosen arbitrarily and must satisfy the basic requirement of pairwise mutual primality. If RNS is given by a set of modules $\{m_1, m_2, \dots, m_n\}$, then for all possible pairs (m_i, m_j) , $1 \leq i \leq n$, $1 \leq j \leq n$ the following conditions must be satisfied:

$$\text{GCD}(m_i, m_j) = 1, \quad (1)$$

where GCD means the greatest common divisor of numbers. Bit depth b_i of each RNS module m_i , $1 \leq i \leq n$ is calculated by the formula (2):

$$b_i = \log_2 m_i, \quad (2)$$

and is the most important characteristic of the computing channel for this RNS module, which determines the speed of calculations and hardware costs for the implementation of arithmetic units in it.

Let us assume that the computation time in the modulo m_i , $1 \leq i \leq n$, channel is equal to t_i . Let's say, for definiteness, RNS modules are ordered in such a way that $t_1 \leq t_2 \leq \dots \leq t_n$. Then, the computing system that implements a practical application in RNS (e.g., digital signal processing), and visually displays the problem of balance, can be depicted as shown in Fig. 1. Figure 1 shows that the elements of the system for

the m_1 module will stand idle $t_n - t_1$ time, for the m_2 module will stand idle $t_n - t_2$ time, etc. Thus, all computing channels, except for the last one, will be idle for some time, which leads to a decrease in the efficiency of using system resources. This circumstance is the problem of the imbalance of the RNS and is a negative factor for its use in practice.

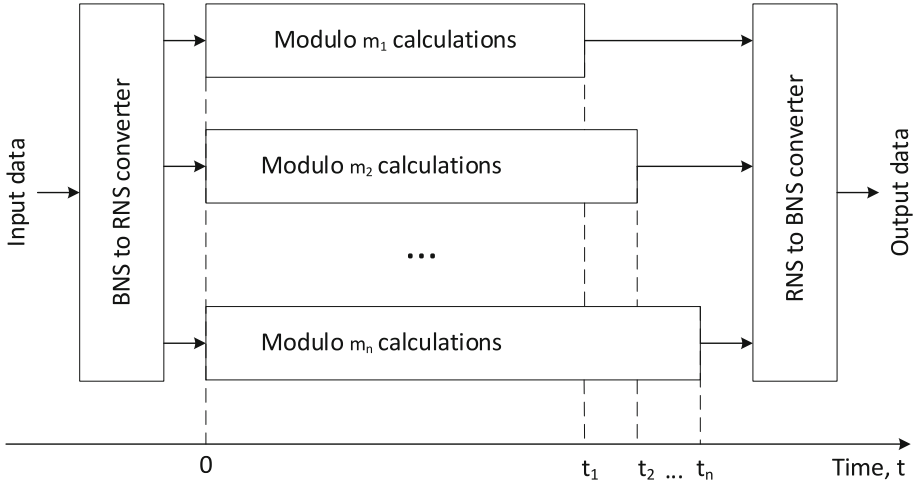


Fig. 1. Scheme of computations in RNS with the indication of computation time in computational channels for each module.

In [12], it was assumed that the value of t_i is directly proportional to the bit depth b_i of the $m_i, 1 \leq i \leq n$ module. To measure the balance of RNS, it is proposed to use the variance of the capacity of modules according to the formula:

$$\beta = \frac{1}{n} \sum_{i=1}^n (b_i - b_{av})^2, \tag{3}$$

where b_{av} is the average bit width of RNS modules, determined by the formula:

$$b_{av} = \frac{1}{n} \sum_{i=1}^n b_i. \tag{4}$$

Value β from formula (3) indicates the spread of bit depths b_i and, consequently, the spread of values t_i . Therefore, the smaller the value β , the more balanced is the RNS. In the case when $\beta = 0$, RNS can be considered perfectly balanced. It is possible to build an ideally balanced RNS by choosing modules of the same bit-width b , imposing the restriction $2^{b-1} < m_i \leq 2^b$ on all modules of the system. However, in practice, not all modules from the specified range are equivalent in the context of designing arithmetic units, except for the case of a tabular implementation of arithmetic operations, suitable for very small bit depths.

In [10], it was proposed to separate the modules of the form 2^k and $2^k - 1$ into a category called low-cost modules. For modules of the form 2^k , the implementation of

combinational adders and multipliers is identical to similar devices in the binary number system with the rejection of carries and bits, starting from the k -th. This allows for fast multi-input summation techniques based on Carry-Save Adders (CSA) and parallel-prefix adders, such as the Kogge-Stone Adder (KSA). For modules of the form $2^k - 1$, the implementation of combinational adders and multipliers is very close to similar devices in the BNS. The carry that occurs when summing the most significant bit modulo $2^k - 1$, is applied to the least significant bit of the device. This technique is called End-Around Carry (EAC) and can be used in combination with fast multi-input summation techniques based on carry-save adders (EAC-CSA) and Kogge-Stone adders (EAC-KSA) [13]. The condition $2^{b-1} < m_i \leq 2^b$ is satisfied by exactly two low-cost modules: 2^b and $2^b - 1$. In other words, a perfectly balanced RNS with low-cost modules can contain only two modules. If the number of RNS modules n is considered as the coefficient of its parallelism, then the system with $n = 2$ can be considered the least attractive from the point of view of the parallel organization of calculations. Adding a third low-cost module to the set $\{2^k, 2^k - 1\}$ inevitably leads to an increase in the measure of balance β . It is shown in [14] that the minimum possible $\beta = 2/9$ for three-module RNS with low-cost modules is achieved in the cases $\{2^k, 2^k - 1, 2^{k-1} - 1\}$ and $\{2^k, 2^k - 1, 2^{k+1} - 1\}$. The parallelism coefficient of such RNS is equal to $n = 3$, which, of course, would also be desirable to increase for efficient practical use of RNS. A further increase in the parallelism factor n will significantly increase the measure of balance β due to the well-known fact that $GCD(2^{k_1} - 1, 2^{k_2} - 1) = 1$ if and only if $GCD(k_1, k_2) = 1$ [10].

There is another type of RNS module, which, although it does not belong to the low-cost category, is quite promising. These are modules of $2^k + 1$ form. Modules of this kind allow the supply of an inverted carry, which occurs when the most significant bit is summed modulo $2^k + 1$, to the least significant bit of the device. This technique is called Inverted-End-Around Carry (IEAC) and can be used in conjunction with fast multi-input summation techniques based on carry-save adders (IEAC-CSA) and Kogge-Stone adder (IEAC-KSA), but requires modulo $2^k + 1$ channel conversion in RNS in diminished-1 format [15]. Adding $2^k + 1$ modules to low-cost modules provides several important benefits in terms of balance and concurrency. First, it is possible to construct perfectly balanced three-module RNSs of the form $\{2^k, 2^k - 1, 2^{k-1} + 1\}$ under the condition $GCD(2^k - 1, 2^{k-1} + 1) = 1$. Secondly, an increase in the parallelism coefficient $n > 3$ will not lead to such a sharp increase in the measure of balance β due to the greater variability of module combinations. Further, the influence of these factors in the construction of RNS for solving applied problems will be experimentally demonstrated.

3 Analysis of the Parallelism and Balance of RNS with Low-Cost and $2^k + 1$ Modules

To demonstrate the advantages of greater balance and parallelism of the RNS, consisting of low-cost modules and modules of the $2^k + 1$ type, an experiment was carried out, described by the following parameters.

1. Ranges were selected that require representation in RNS and characterize typically applied tasks in the field of digital signal processing. For this purpose, the ranges from 16 to 64 bits in 8-bit increments were used, that is, 16, 24, 32, ..., 64 bits.

2. When constructing possible RNS, it was assumed that they would contain 3 to 8 modules.
3. Three sets of possible modules were formed for subsequent enumeration. The first set included modules of the 2^k , $1 \leq k \leq 12$ forms. The second set consisted of modules of the $2^k - 1$, $2 \leq k \leq 12$ forms. The third set contained modules of the $2^k + 1$, $2 \leq k \leq 12$ forms. Thus, the first two sets contained possible low-cost RNS modules, and the third set contained possible modules of the $2^k + 1$ form.
4. For each selected range from 16 to 64 bits, a complete enumeration of all possible RNS was carried out with modules from the sets described above and a range no less than the required one and not exceeding the required value by more than 2 bits. That is, for the 16-bit range, RNSs with ranges of 16 to 18 bits were selected. For the 24-bit range, RNSs with ranges of 24 to 26 bits were selected, etc. The RNS, with the minimum measure of balance β was considered the best and was entered in Table 1. In the case of the presence of several possible RNS with the required parameters, and the same measure of balance β , the RNS with a smaller range was chosen. For example, for a 24-bit range, two different RNSs with low-cost modules and a measure of balance $\beta \approx 0.33$ can be chosen. The first case contains modules $\{255, 256, 511\}$, the range of such RNS is 33358080 and its bit depth is $\log_2 33358080 \approx 24.99$. The second case contains modules $\{255, 511, 512\}$, the range of such RNS is 66716160 and its bit depth is $\log_2 66716160 \approx 25.99$. According to the above criterion, Table 1 is filled with RNS $\{255, 256, 511\}$. The absence in Table 1 of some rows corresponding to the combinations of the bit depth of the range and the number of RNS modules, as well as dashes, means that it was not possible to find the RNS with the required parameters for the specified cases.

Table 1. Results of the experiment on the enumeration of RNS, consisting of low-cost modules and modules of the $2^k + 1$ form

Required limits			RNS obtained during the experiment			
Required range bit depth	Number of RNS modules	Type of RNS modules	Values of RNS modules	Range bit depth	Average bit width of RNS modules b_{av}	Measure of balance β
16	3	low-cost	31,63,64	16.93	5.67	0.33
		all	17,63,64	16.06	5.67	0.33
	4	low-cost	7,15,31,32	16.67	4.25	0.92
		all	9,17,31,32	17.21	4.75	0.25
	5	low-cost	-	-	-	-

(continued)

Table 1. (continued)

Required limits			RNS obtained during the experiment			
Required range bit depth	Number of RNS modules	Type of RNS modules	Values of RNS modules	Range bit depth	Average bit width of RNS modules b_{av}	Measure of balance β
		all	5,7,9,16,17	16.39	3.80	0.70
	6	low-cost	-	-	-	-
		all	3,4,5,7,17,31	17.76	3.33	1.86
24	3	low-cost	255,256,511	24.99	8.33	0.33
		all	129,256,511	24.01	8.33	0.33
	4	low-cost	31,63,127,128	24.92	6.25	0.92
		all	33,65,127,128	25.06	6.75	0.25
	5	low-cost	7,15,31,64,127	24.66	5.00	2.50
		all	17,31,32,33,65	25.11	5.60	0.80
	6	low-cost	-	-	-	-
		all	5,7,17,31,32,33	24.22	4.50	1.50
	7	low-cost	-	-	-	-
		all	3,5,7,8,17,31,127	25.74	4.00	3.00
32	3	low-cost	2047,2048,4095	34.00	11.33	0.33
		all	1025,2047,2048	32.00	11.00	0.00
	4	low-cost	127,255,511,512	32.98	8.25	0.92
		all	129,257,511,512	33.01	8.75	0.25
	5	low-cost	31,32,127,255,511	33.93	6.80	3.20
		all	33,65,127,128,257	33.06	7.20	1.20
	6	low-cost	7,15,16,31,127,2047	33.66	5.67	8.67
		all	17,31,33,64,65,127	33.10	6.00	0.80
	7	low-cost	-	-	-	-
		all	7,9,17,31,32,65,127	33.03	5.14	2.14
	8	low-cost	-	-	-	-
		all	3,5,7,8,17,31,127,257	33.75	4.63	5.70
40	4	low-cost	511,1023,2047,2048	41.00	10.25	0.92
		all	513,1025,2047,2048	41.00	10.75	0.25
	5	low-cost	127,128,255,511,2047	41.98	8.40	2.80
		all	127,129,257,511,512	40.00	8.40	0.80

(continued)

Table 1. (continued)

Required limits			RNS obtained during the experiment			
Required range bit depth	Number of RNS modules	Type of RNS modules	Values of RNS modules	Range bit depth	Average bit width of RNS modules b_{av}	Measure of balance β
	6	low-cost	15,31,64,127,511,2047	41.85	7.00	6.80
		all	17,65,127,128,129,511	40.11	7.17	1.76
	7	low-cost	-	-	-	-
		all	17,31,33,64,65,127,257	41.10	6.43	1.95
	8	low-cost	-	-	-	-
		all	7,16,17,31,33,65,127,257	41.91	5.75	3.64
48	5	low-cost	127,511,1023,2047,4096	48.98	9.80	3.70
		all	257,511,1023,1024,2047	48.00	9.80	0.70
	6	low-cost	31,127,255,511,512,2047	48.93	8.17	4.17
		all	65,127,256,257,511,1023	48.01	8.33	1.47
	7	low-cost	-	-	-	-
		all	17,65,127,128,129,257,511	48.11	7.43	1.95
	8	low-cost	-	-	-	-
		all	17,31,32,33,65,127,257,511	49.10	6.62	2.84
56	6	low-cost	-	-	-	-
		all	257,511,513,1024,1025,2047	57.01	10.00	0.80
	7	low-cost	-	-	-	-
		all	65,127,129,256,257,511,2047	56.02	8.43	1.95
	8	low-cost	-	-	-	-
		All	17,31,65,127,256,257,511,513	56.06	7.50	3.43
64	7	low-cost	-	-	-	-
		All	127,257,511,513,1025,2047,2048	65.00	9.71	2.24
	8	low-cost	-	-	-	-
		All	17,65,127,257,511,512,1023,2047	64.10	8.38	3.70

In the next section, the results of the experiment shown in Table 1 will be discussed, and some directions for further research on the application of RNS with special sets of modules in practice will be proposed.

4 Results and Discussion

According to the numerical results given in Table 1, some conclusions can be drawn. First, as expected, the greater variability of possible RNS modules due to the addition of modules of the $2^k + 1$ form made it possible to apply such RNS more widely, both in the required range and in the number of modules in the system. The ranges of 56 and 64 bits could not be represented at all in RNS with low-cost modules, while in RNS with the addition of modules of the $2^k + 1$ form, it was possible to obtain combinations of six modules (for a range of 56 bits), as well as seven- and eight-module sets. For all other ranges, adding modules of the $2^k + 1$ form made it possible to obtain an RNS with an increased parallelism coefficient. Secondly, the addition of modules of the $2^k + 1$ form made it possible in most cases to improve significantly the balance of the RNS, which can be seen from the decrease in the measure of balance β in almost all comparable cases in Table 1. The average bit width of RNS modules b_{av} remained practically unchanged when moving from low-cost modules to extended sets in comparable cases in Table 1. An increase in the number of modules for the same range proportionally reduces the average bit width of RNS modules and has a pronounced tendency to increase the measure of balance β .

To discuss the efficiency of the practical use of the constructed RNS containing both low-cost modules and modules of the $2^k + 1$ form, suppose that to solve some applied problem (for example, the implementation of a FIR filter), an RNS with modules $\{2^{k_1}, 2^{k_2} - 1, \dots, 2^{k_r} - 1, 2^{k_{r+1}} + 1, \dots, 2^{k_n} + 1\}$, whose exponents are $k_i, 1 \leq i \leq n$ can be repeated. The need to use diminished-1 encoding to organize calculations by modules $2^{k_{r+1}} + 1, \dots, 2^{k_n} + 1$ requires the organization of calculations in such RNS according to the scheme shown in Fig. 2.

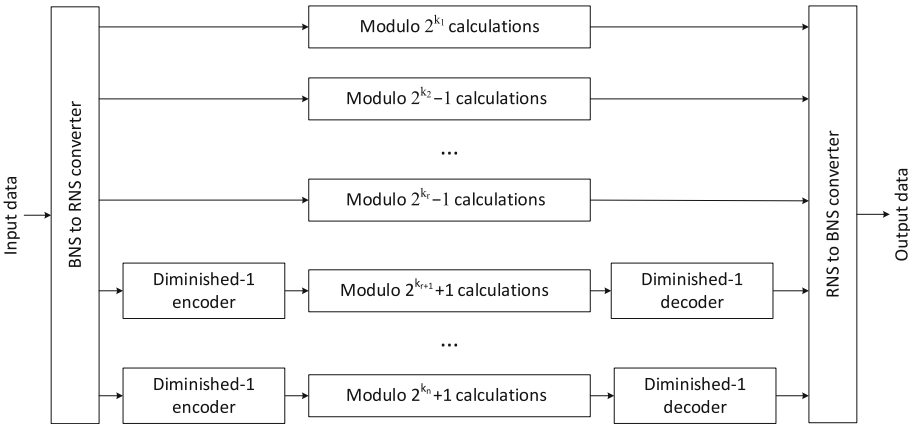


Fig. 2. Scheme of calculations in RNS built on low-cost modules with the addition of modules of the $2^k + 1$ form.

If we follow the notation introduced earlier and consider the execution time of operations for the i -th module equal to $t_i, 1 \leq i \leq n$, then for the module 2^{k_1} the computation

time will be equal to the total operation time of all involved CSA-trees and KSA. For modules $2^{k_2} - 1, \dots, 2^{k_r} - 1$ the computation time will be equal to the total running time of all involved EAC-CSA-trees and EAC-KSA. For modules $2^{k_{r+1}} + 1, \dots, 2^{k_n} + 1$, the computation time will be equal to the total running time of all involved IEAC-CSA-trees and IEAC-KSA, as well as the diminished-1 encoder and decoder. The unique nature of the formation of the values $t_i, 1 \leq i \leq n$ encourages the search for improved approaches to the theoretical measurement of the RNS balance. As an alternative to calculations using formulas (3)–(4), we can consider the average calculation time t_{av} in the RNS channels:

$$t_{av} = \frac{1}{n} \sum_{i=1}^n t_i \quad (5)$$

and time dispersion:

$$\tau = \frac{1}{n} \sum_{i=1}^n (t_i - t_{av})^2. \quad (6)$$

Of course, calculating the values $t_i, 1 \leq i \leq n, t_{av}$ and τ is a much more difficult task than calculating the values $b_i, 1 \leq i \leq n, b_{av}$ and β , as it depends on the way the logical implementation of adders and multipliers and other devices is involved. In addition, it is necessary to take into account the direct technical characteristics and features of the target devices for computing, such as FPGA and ASIC. The well-known unit-gate model [16] can be considered as a first approximation for further theoretical analysis of RNS balance according to formulas (5)–(6). This direction may be an interesting topic for further research.

5 Conclusion

The paper showed a significant increase in parallelism and improvement in the balance of RNS with modules of the $2^k, 2^k - 1$ and $2^k + 1$ forms, compared to using only modules of the 2^k and $2^k - 1$ forms. This fact can help in the development of more efficient solutions in digital signal processing and cryptography based on RNS. A serious issue for further research in this area is to study the effect of additional converters on the diminished-1 code and vice versa for modules of the $2^k + 1$ form compared to the improvement in the RNS characteristics shown in the experimental part. This problem is subject to a thorough study both at the theoretical level, for example, using a unit-gate model, and practical tests on modern FPGA and ASIC microelectronic devices.

Acknowledgments. The research was supported by Russian Science Foundation, project 21-71-00017.


References

1. Sousa, L.: Nonconventional computer arithmetic circuits, systems and applications. IEEE Circuits Syst. Mag. **21**, 6–40 (2021). <https://doi.org/10.1109/MCAS.2020.3027425>

2. Cardarilli, G.C., di Nunzio, L., Fazzolari, R., Nannarelli, A., Petricca, M., Re, M.: Design space exploration based methodology for residue number system digital filters implementation. *IEEE Trans. Emerg. Top. Comput.* **10**, 186–198 (2022). <https://doi.org/10.1109/TETC.2020.2997067>
3. Vayalil, N.C., Paul, M., Kong, Y.: A Residue number system hardware design of fast-search variable-motion-estimation accelerator for HEVC/H.265. *IEEE Trans. Circuits and Syst. Video Technol.* **29**, 572–581 (2019). <https://doi.org/10.1109/TCSVT.2017.2787194>
4. Vennos, A., George, K., Michaels, A.: Attacks and defenses for single-stage residue number system PRNGs. *IoT* **2**, 375–400 (2021). <https://doi.org/10.3390/iot2030020>
5. Bajard, J.C., Eynard, J., Merkiche, N.: Montgomery reduction within the context of residue number system arithmetic. *J. Cryptogr. Eng.* **8**, 189–200 (2018). <https://doi.org/10.1007/S13389-017-0154-9/FIGURES/1>
6. Samimi, N., Kamal, M., Afzali-Kusha, A., Pedram, M.: Res-DNN: a residue number system-based DNN accelerator unit. *IEEE Trans. Circuits Syst. I Regul. Pap.* **67**, 658–671 (2020). <https://doi.org/10.1109/TCSI.2019.2951083>
7. Zarandi, A.A.E., Molahosseini, A.S., Hosseinzadeh, M., Sorouri, S., Antao, S., Sousa, L.: Reverse converter design via parallel-prefix adders: novel components, methodology, and implementations. *IEEE Trans. VLSI Syst.* **23**, 374–378 (2015). <https://doi.org/10.1109/TVLSI.2014.2305392>
8. Gorodecky, D., Villa, T.: Efficient hardware operations for the residue number system by boolean minimization. In: Drechsler, R., Soeken, M. (eds.) *Advanced Boolean Techniques*, pp. 237–258. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-20323-8_11
9. Hiasat, A.: New residue number system scaler for the three-moduli set $\{2n + 1 - 1, 2n, 2n - 1\}$. *Computers* **7**, 46 (2018). <https://doi.org/10.3390/computers7030046>
10. Parhami, B.: *Computer Arithmetic: Algorithm and Hardware Designs* (2009)
11. Chaves, R., Sousa, L.: Improving residue number system multiplication with more balanced moduli sets and enhanced modular arithmetic structures. *IET Comput. Digital Tech.* **1**, 472–480 (2007). <https://doi.org/10.1049/IET-CDT:20060059>
12. Boyvalenkov, P., et al.: Classification of moduli sets for residue number system with special diagonal functions. *IEEE Access.* **8**, 156104–156116 (2020). <https://doi.org/10.1109/ACCESS.2020.3019452>
13. Mohan, P.A.: *Residue Number Systems: Theory and Applications*. Birkhäuser (2016). <https://doi.org/10.1007/978-3-319-41385-3>
14. Lyakhov, P., Bergerman, M., Semyonova, N., Kaplun, D., Voznesensky, A.: Design reverse converter for balanced RNS with three low-cost modules. In: *2021 10th Mediterranean Conference on Embedded Computing, MECO 2021* (2021). <https://doi.org/10.1109/MECO52532.2021.9460200>
15. Živaljević, D., Stamenković, N., Stojanović, V.: Digital filter implementation based on the RNS with diminished-1 encoded channel. In: *2012 35th International Conference on Telecommunications and Signal Processing, TSP 2012 – Proceedings*, PP. 662–666 (2012). <https://doi.org/10.1109/TSP.2012.6256380>
16. Zimmermann, R.: *Binary Adder Architectures for Cell-Based VLSI and Their Synthesis*. Hartung-Gorre, Konstanz (1998)



Uncoded Video Data Stream Denoising in a Binary Symmetric Channel

Anzor R. Orazhev^(✉) 

North-Caucasus Center for Mathematical Research, North-Caucasus Federal University,
Stavropol, Russia
aorazhev@ncfu.ru

Abstract. In this paper, a method for denoising an uncoded video stream in a binary symmetric channel is considered. When a bit of information is damaged, noise similar to impulse noise occurs with a certain probability. This work considers a model for transmitting visual data through a binary symmetric channel, where the noise characteristic corresponds to impulse noise distributed over the image with random values. Both random variables are distributed uniformly, both in brightness and in spatial location. In the proposed method, the distorted pixel is detected by comparing pixels inside the filter mask. Pixels are compared by their brightness value, and the remoteness of pixels within the detector area is also taken into account. The distance between pixels is calculated using the Euclidean metric. The local area of the filter takes into account pixels from the previous and next frames. Video frame recovery is performed using adaptive median filtering. A comparison was made with known methods. Based on the mean square error (MSE) and structural similarity index (SSIM) characteristics, it was shown that the proposed method copes with the task of denoising visual data better than the known methods.

Keywords: Binary Symmetric Chanel · Impulse noise · Median filter · Image processing · Adaptive filtering

1 Introduction

Images, as one of the forms of information presentation, which are transmitted as messages over communication channels, are subject to interference [1]. When transmitting uncoded images over a communication channel, for example, represented as a binary symmetric channel model, each bit can be distorted with a certain probability [2]. If at least one bit of the image is distorted, interference occurs that distorts the pixel values. Pixel distortion with a certain probability in a binary symmetric channel is similar to impulse noise. The probability of the appearance of distorted pixels increases with a video data stream, which is a series of images (frames) that follow each other with a certain frequency. Noisy frames can negatively affect the operation of various digital image processing algorithms, for example, real-time object recognition, as well as incorrectly displaying data received from medical or seismological sensors [3].

The task of restoring images and videos from impulse noise consists of the task of finding an impulse and the task of restoring a distorted pixel. One of the simple and effective methods for removing impulse noise from images and videos is median filters [4]. But the standard median filter leads to blurring of the image, so modifications of the median filter have been proposed by various authors. One of the modifications that greatly reduced the negative effect of blurring is adaptive median filtering [5]. In adaptive median filters, pixels that do not impulse noise remain unchanged, changes are made only for pixels that have been identified as noisy [6].

To denoising video from impulse noise, much fewer approaches have been proposed that would take into account the features of video data. In [7], an adaptive median filter with a mask is proposed that takes into account pixels from the previous and next video frames, and the brightness of the reconstructed pixel is calculated using the Lorentz function, which takes into account the distance between pixels inside the filter mask.

The complexity of the pixel detection task depends on the impulse noise model. In the “salt and pepper” impulse noise model, where the distorted pixels take on two values: an impulse with minimum and maximum brightness, the task of detection is usually not worth it at all. Currently, a number of methods for cleaning and detecting impulse noise are known. In the work [8] for denoising the image from impulse and Gaussian noise, a method was developed, which is a modification of the bilateral filter for determining distorted impulses [9]. [10] describes a method that offers an improvement on the [8] method and uses a logarithmic function and threshold transformations for this. [11] also describes a comparison between the [10] and [8] methods and proposes a method that is another modification of the [8] method. The method introduces a new statistic, the Local Consensus Index (LCI), which is calculated by summing up all pixel similarity values in its neighborhood and finding the value of the central element.

In this work, the model of video data transmission through a binary symmetric channel (BSC) will be considered. It will be shown that the characteristics of the noise that occurs in the BSC correspond to the random-valued impulse noise model. The detector of distorted pixels in a video frame is based on the score of the difference between pixels.

2 Binary Symmetric Channel Model

A channel with binary input and a binary output, where the probabilities of error and correct transmission are equal, is called a binary symmetric channel. Since each output binary symbol of a channel depends only on the corresponding input binary symbol, we say that this channel without memory [12]. Signals can be transmitted over a binary channel, for example, 0 or 1. Transmission in such a communication channel is not ideal, because of this, the receiving signal with a certain probability may receive an error, which consists in replacing the sign of 1 with 0 or 0 with 1.

Figure 1 shows a scheme of a binary symmetric channel p is the probability of error, $1 - p$ is the probability of correct transmission. BSC has input and output signal $X \in \{0, 1\}$ and $Y \in \{0, 1\}$, hence.

$$p(X|Y) = \begin{cases} 1 - p, & \text{if } X = Y \\ p, & \text{if } X \neq Y \end{cases} \quad (1)$$

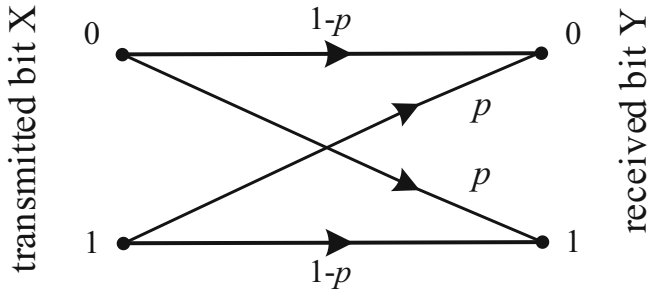


Fig. 1. Scheme of a binary symmetric channel

Distorted signals in the BSC appear as a result of interference. Interference is understood as any random influence on the signal in the communication channel that prevents the correct reception of signals. In communication channels, there are both additive interferences, i.e., random processes superimposed on the transmitted signals, and multiplicative interference, expressed in random changes in the channel characteristics [13].

Additive interference contains three components: concentrated in frequency (harmonic), concentrated in time (impulse) and fluctuation. Impulse interference is a sequence of short-term pulses separated by intervals exceeding the time of transients in the channel. The causes of impulse interference are: the influence of lightning discharges on communication lines; the influence of power lines on communication lines; poor contacts in transmission and power equipment; shortcomings in the development and manufacture of equipment; operational reasons, etc. Deficiencies in the design and manufacture of equipment lead to the fact that impulse noise occurs during voltage surges in the supply network or switching from the main elements to the reserve ones. Digital data is often transmitted as a sequence of binary numbers (bits of information). During transmission, noise can distort the original message. The model consists of a transmitter capable of sending a binary signal and a receiver.

Data transmission in the BSC can be described by the Bernoulli scheme. Let Δ be a random variable that counts the number of failures. Then, according to the Bernoulli scheme, the probability of m errors occurring when transmitting n bits through the BSC is

$$p(\Delta = m) = \binom{n}{m} p^{n-m} q^m \tag{2}$$

where n is the bit depth of a video frame pixel, p is the probability of one bit distortion in a binary symmetric channel. Based on (2), the density of impulse noise in the image in accordance with the bit depth of the image and the probability of bit distortion is

$$\rho = 1 - p^{n-m} \tag{3}$$

Table 1 shows the density of impulse noise ρ on the image in accordance with the probability of bit distortion p .

Table 1. Impulse noise density ρ on the image

Bit corruption probability, p	Pixel depth, n			
	8	12	16	24
0.01	0.0772	0,1136	0,1485	0,2143
0.05	0.3366	0,4596	0,5599	0,7080
0.10	0.5695	0,7176	0,8147	0,9202

Consider the case of broadcasting from a video surveillance camera, where errors occur in the communication channel. Figure 2 shows video frames, where each bit of the frame is distorted in the images with probability $p = 0.01, 0.05, 0.1$. In this case, video frames are 8-bit.

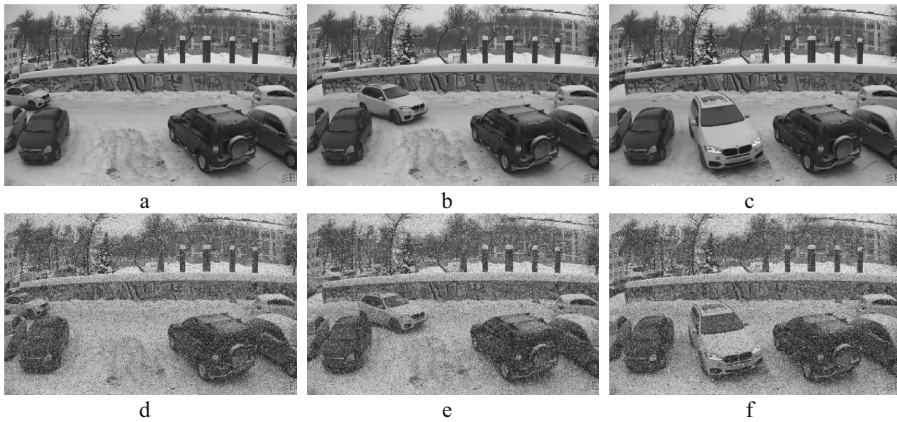


Fig. 2. Frames from the video used in the modeling a) 1st; b) 50th; c) 100th; d) distorted 1st video frame; e) distorted 50th video frame; f) distorted 100th video frame;

Figure 3 shows the distribution of pixels an uncoded video frame transmitted via BSC. The figure shows that the distribution of distorted pixels and their brightness is close to uniform. Random in value and location in the frame noise that is uniformly distributed corresponds to the characteristics of random-valued impulse noise. To eliminate this type of noise, denoising methods based on median filtering are used.

3 Method for Uncoded Video Data Stream Denoising in a Binary Symmetric Channel

Let digital videos be represented by a set of pixels with intensity values $x_{i,j,k}$ whose coordinates (i, j) change over some subset Z^2 , where Z is a set of integers, and k is the frame number in the video.

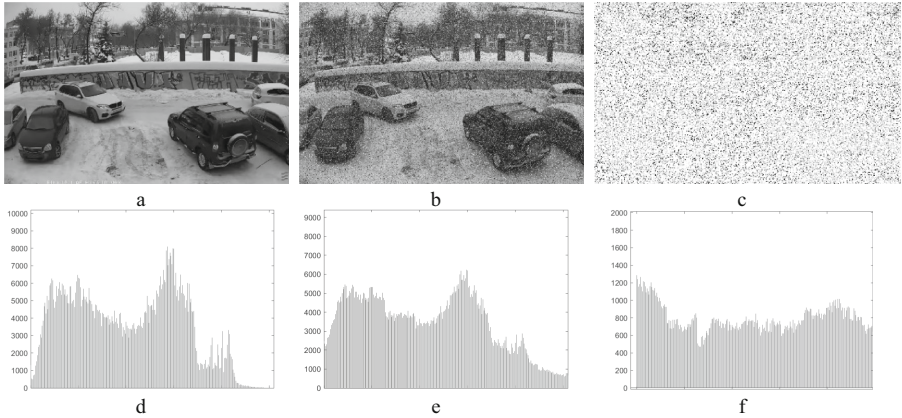


Fig. 3. An example of an uncoded video frame distortion transmitted via BSC: a) original undistorted frame; b) frame corrupted by BSC; c) the location of impulse noise in the frame; d) distribution of pixel brightnesses on an undistorted frame; e) distribution of pixel brightnesses on a distorted frame; f) the distribution of the brightness of impulse noise in the frame

In the proposed method, on the frame of a noisy video, it is necessary to determine whether a bit of an image pixel has been distorted. To do this, a score of the difference between pixels in the local window is introduced. The score of the difference is based on two parameters:

- 1) The pixel brightness difference parameter, which we propose to calculate by the formula

$$\gamma(i, j, k) = 1 + \{\log_2 |x_{i,j,k} - x_{a,b,c}| - 8, -8\}, x_{a,b,c} \in \Omega_{x_{i,j,k}}, \quad (4)$$

Next, sort U and sum the first $m/2$ elements, where m is the number of elements in the local window Ω :

$$\alpha_m(x_{i,j,k}) = \sum_{l=1}^m \gamma_l(x_{i,j,k}). \quad (5)$$

- 2) Geometric distance parameter, based on the Euclidean metric, which determines the difference between pixels in the local window Ω

$$\beta(x_{i,j,k}, x_{a,b,c}) = \exp\left(-\|x_{i,j,k} - x_{a,b,c}\|^2 / (2\psi_\beta^2)\right), \quad (6)$$

where (i, j) and (a, b) denote the pixel coordinates, k and c are the frame number in the video. The ψ parameter controls $\beta(x_{i,j,k}, x_{a,b,c})$ with respect to the geometric distance.

As a result, the similarity parameters between two pixels are obtained, based on the geometric distance and the difference in the brightness of the pixels in the detector window, with which a score of the difference between the pixels M can be obtained:

$$M(x_{i,j,k}, x_{a,b,c}) = \alpha(x_{i,j,k}, x_{a,b,c}) \cdot \beta(x_{i,j,k}, x_{a,b,c}). \quad (7)$$

The similarity score under formula (7) forms an array of values, where, using a certain threshold T , it is possible to determine whether a pixel is an impulse. In the proposed method, the optimal threshold value for $T = 20$. Therefore, if the values in the array M are greater than the threshold value, then the image pixel is an impulse.

Given the features of the video, you can use pixels from other video frames in the Ω filter mask. We propose to use the filter mask of the following type, which is shown in Fig. 4. The distance between pixels in the local window is proposed to be determined by the Euclidean metric (L_2) [14]. The distance $R(x_{ijk}, x_{abc})$ between pixels x_{ijk} and x_{abc} in the metric is determined by the formula

$$R(x_{ijk}, x_{abc}) = \sqrt{(i - a)^2 + (j - b)^2 + (k - c)^2}.$$

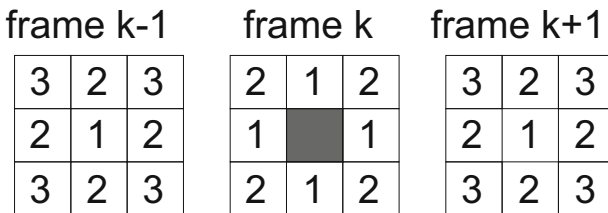


Fig. 4. Local window size Ω . The squares of distances defined by the Euclidean metric L_2 are inscribed in the pixel cells.

For pixels defined as distorted in the local area Ω , an array of undistorted pixels is formed, in which the median is calculated. The resulting median value is assigned to the distorted pixel.

Figure 5 shows the scheme of the proposed method, which was described above.

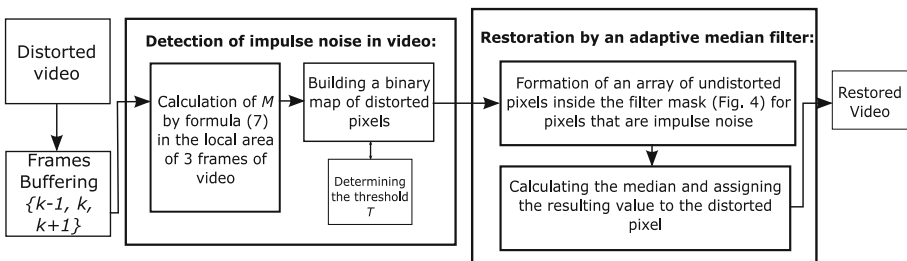


Fig. 5. Scheme of the method for denoising of uncoded the video data stream transmitted through a binary symmetric channel.

For the first and last frames of the video, the buffer is only two frames: for the first it is the current and next frames, for the last it is the previous and current frames.

4 Modeling of Uncoded Video Data Stream Denoising in a Binary Symmetric Channel

An 8-bit grayscale video was used for simulation. The frames of this video are shown in Fig. 2. The video consists of 100 frames. The frame rate is 24 frames per second. Resolution 1228 by 720 pixels.

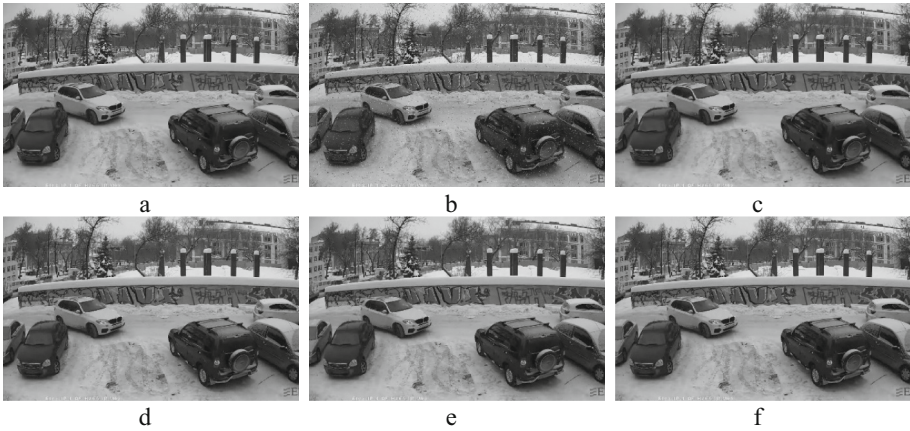


Fig. 6. Modeling results by various methods: a) 50th frame of the original video; b) 50th frame of video transmitted via BSC ($p = 0.01$); c) the result of restoration by the method [11]; d) the result of restoration by the method [8]; e) the result of restoration by the method [10]; f) the result of recovery by the proposed method

In the video, each bit is distorted with probabilities $p = 0.01, 0.05, 0.1, \dots$. To determine the quality of video processing, the mean square error (MSE) [15] and the structural similarity index (SSIM) [16] were used.

The results of the modeling are presented in Tables 2 and 3. Figures 6, 7 and 8 show the 50th frame of the original, distorted and restored video.

The simulation results showed that the proposed method demonstrates the best processing results for frames that were transmitted through the BSC, as can be seen in Tables 2 and 3. In Table 2, lower MSE values are obtained for all bit distortion probabilities. In Table 3, for the bit distortion probabilities $p = 0.01$ and $p = 0.10$, the highest SSIM values were obtained, but for $p = 0.05$ the [11] method received the best processing quality.

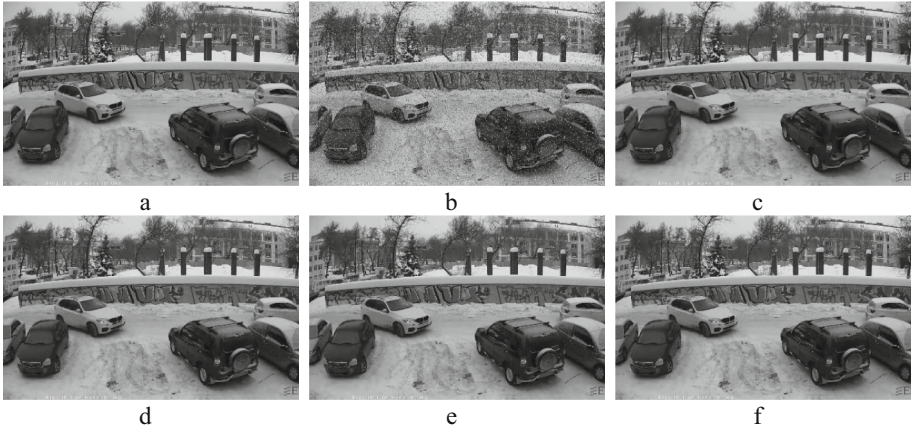


Fig. 7. Modeling results by various methods: a) 50th frame of the original video; b) 50th frame of video transmitted via BSC ($p = 0.05$); c) the result of restoration by the method [11]; d) the result of restoration by the method [8]; e) the result of restoration by the method [10]; f) the result of recovery by the proposed method

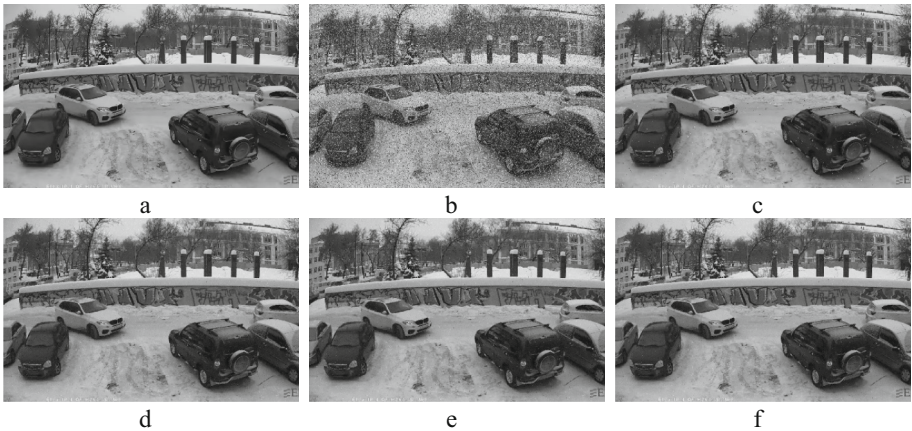


Fig. 8. Modeling results by various methods: a) 50th frame of the original video; b) 50th frame of video transmitted via BSC ($p = 0.10$); c) the result of restoration by the method [11]; d) the result of restoration by the method [8]; e) the result of restoration by the method [10]; f) the result of recovery by the proposed method

Table 2. MSE values for various methods of video data removal from impulse

bit corruption probability, p	known methods			proposed method
	[8]	[10]	[11]	
0.01	93.8194	86.1825	102.9058	49.4638
0.05	129.8409	124.4533	122.9748	92.9278
0.10	179.4645	183.0262	215.4525	151.3833

Table 3. SSIM values for various methods of video data removal from impulse

bit corruption probability, p	known methods			proposed method
	[8]	[10]	[11]	
0.01	0.9272	0.9319	0.9150	0.9545
0.05	0.8211	0.8312	0.8954	0.8593
0.10	0.7062	0.7007	0.7290	0.7428

5 Conclusion

The paper considers the case of video transmission through a binary symmetric channel, where with a certain probability each bit of the image was distorted. The distribution of distortions on video frames showed that the random noise in terms of value and location in the frame, which is uniformly distributed, corresponds to the characteristics of random-valued impulse noise. A method was proposed for detecting and denoising distorted pixels in video frames, which is based on score of the difference between pixels in terms of brightness and geometric distance of pixels in the local window. The distances in the local window are defined by the Euclidean metric.

In simulation, using the characteristics of SSIM and MSE, it was shown that the proposed method coped with the task of denoising in the best way. The proposed method can be used, for example, in video surveillance systems, where interference often occurs in communication channels. And also, in places where weather conditions distort the signals transmitted via communication channels.





Acknowledgments. The authors would like to thank the North Caucasus Federal University for supporting the contest of projects competition of scientific groups and individual scientists of the North Caucasus Federal University. The work is supported by the North-Caucasus Center for Mathematical Research under agreement № 075-02-2021-1749 with the Ministry of Science and Higher Education of the Russian Federation and by Russian Foundation for Basic Research project 19-07-00130.

References

1. Gonzales, R.C., Woods, R.E.: Digital Image Processing, Fourth edn. (2018)
2. Stone, J.V.: Information Theory: A Tutorial Introduction (2015)
3. Kuruvilla, J., Sukumaran, D., Sankar, A., Joy, S.P.: A review on image processing and image segmentation. In: Proceedings of 2016 International Conference on Data Mining and Advanced Computing, SAPIENCE 2016 (2016)
4. Petrou, M., García Sevilla, P.: Image Processing: Dealing with Texture (2006)
5. Chervyakov, N., Lyakhov, P., Orazaev, A.: Two methods of adaptive median filtering of impulse noise in images. *Comput. Optics* **42**(4), 667–678 (2018). <https://doi.org/10.18287/2412-6179-2018-42-4-667-678>
6. Verma, K., Kumar Singh, B., Thoke, A.S.: An enhancement in adaptive median filter for edge preservation. *Procedia Comput. Sci.* **48**, 29–36 (2015)
7. Chervyakov, N.I., Lyakhov, P.A., Orazaev, A.R.: 3D-generalization of impulse noise removal method for video data processing. *Comput. Optics* **44**, 92–100 (2020). <https://doi.org/10.18287/2412-6179-CO-577>
8. Garnett, R., Huegerich, T., Chui, C., He, Wenjie: A universal noise removal algorithm with an impulse detector. *IEEE Trans. Image Process.* **14**(11), 1747–1754 (2005). <https://doi.org/10.1109/TIP.2005.857261>
9. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: Proceedings of the IEEE International Conference on Computer Vision (1998)
10. Dong, Y., Chan, R.H., Shufang, X.: A detection statistic for random-valued impulse noise. *IEEE Trans. Image Process.* **16**, 11112–11120 (2007). <https://doi.org/10.1109/TIP.2006.891348>
11. Xiao, X., Xiong, N.N., Lai, J., Wang, C.-D., Sun, Z., Yan, J.: A local consensus index scheme for random-valued impulse noise detection systems. *IEEE Trans. Syst., Man, Cybern.: Syst.* **51**(6), 3412–3428 (2021). <https://doi.org/10.1109/TSMC.2019.2925886>
12. Timme, N.M., Lapish, C.: A tutorial for information theory in neuroscience. *eneuro* **5**(3), ENEURO.0052-18.2018 (2018). <https://doi.org/10.1523/ENEURO.0052-18.2018>
13. Witten, E.: A mini-introduction to information theory. *La Rivista del Nuovo Cimento* **43**, 187–227 (2020). <https://doi.org/10.1007/s40766-020-00004-5>
14. Jähne, B.: Digital Image Processing and Image Formation (2018)
15. Karunasingha, D.S.K.: Root mean square error or mean absolute error? Use their ratio as well. *Inform. Sci.* **585**, 609–629 (2022). <https://doi.org/10.1016/j.ins.2021.11.036>
16. Sara, U., Akter, M., Uddin, M.S.: Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *J. Comput. Commun.* **07**, 8–18 (2019). <https://doi.org/10.4236/jcc.2019.73002>



Cloud-Based Service for Recognizing Pigmented Skin Lesions Using a Multimodal Neural Network System

Ulyana Alekseevna Lyakhova^(✉) , Daria Nikolaevna Bondarenko ,
Emiliya Evgenevna Boyarskaya , and Nikolay Nikolaevich Nagornov 

Department of Mathematical Modeling, North-Caucasus Federal University, 355017 Stavropol,
Russia

uljahovs@mail.ru

Abstract. Skin cancer is the most common cancer in humans today and is usually caused by exposure to ultraviolet radiation. There are many diagnostic methods for visual analysis of pigmented neoplasms. However, most of these methods are subjective and largely dependent on the experience of the clinician. To minimize the influence of the human factor, it is proposed to introduce artificial intelligence technologies that have made it possible to reach new heights in terms of the accuracy of classifying medical data, including in the field of dermatology. Artificial intelligence technologies can equal and even surpass the capabilities of a dermatologist in terms of the accuracy of visual diagnostics. The article proposes a web application based on a multimodal neural network system for recognizing pigmented skin lesions as an additional auxiliary tool for oncologist. The system combines and analyzes heterogeneous dermatological data, which are images of pigmented neoplasms and such statistical information about the patient as age, gender, and localization of pigmented skin lesions. The recognition accuracy of the proposed web application was 85.65%. The use of the proposed web application as an auxiliary diagnostic method will expand the possibilities of early detection of skin cancer and minimize the impact of the human factor.

Keywords: Multimodal Neural Networks · Heterogeneous Data · Cloud-based System · Pigmented Skin Lesions · Skin Cancer · Melanoma

1 Introduction

According to the World Health Organization (WHO), skin cancer is one of the leading oncological diseases in the world [1]. The most common method for recognizing malignant pigmented neoplasms in dermatologists-oncologists is visual diagnostics using dermatoscopy. Dermatoscopy is a non-invasive diagnostic method that makes it possible to identify the morphological features of a pigmented formation [2]. The main limitation of this method is the dependence of its effectiveness on the experience of the dermatologist [3]. The average accuracy of visual diagnosis of skin cancer is 65–75% [4].

A more accurate and reliable method for diagnosing skin cancer is a biopsy. This method is invasive. A portion of the pigmented lesion is scraped off and sent to a laboratory to confirm the tumor status [5]. The procedure is painful, time-consuming, and requires healthcare funding. The main problem with using the biopsy method is that laboratory sampling often causes inflammation or even the spread of the lesion. A primary potential malignant pigmented neoplasm may pave the way for further foci of melanoma and non-melanoma skin cancer [6].

To overcome the problem of diagnosing skin cancer, new strategies have been proposed to provide greater “objectivity” in the early detection of suspicious pigmented lesions, especially for dermatologists who do not have sufficient experience in dermoscopy. Among them is the introduction of intelligent systems for analyzing images and statistical data on patients, which are an auxiliary tool for the early diagnosis of skin cancer.

Automatic analysis of pigmented skin lesions is a hot topic of research aimed at developing tools for the computerized diagnosis of skin cancer [7]. To date, artificial intelligence algorithms have been able to surpass the classification accuracy of dermatologists [8]. However, neural network systems do not provide reliable estimates of predicted pigmented skin lesions, since they cannot be used as an independent diagnostic tool without the help of a doctor. At the same time, artificial intelligence algorithms can be applied in web applications, mobile applications, and software as a telemedicine tool, as well as an auxiliary diagnostic system for doctors. Computerized diagnostics is a growing need due to the growing number of emerging malignant neoplasms, subjectivity of the procedure, time, and financial costs [9].

The work [10] presents a cloud-based system for diagnosing skin lesions using convolutional neural networks. The cloud system only processes visual data. The accuracy of the proposed system is 77.4%. The mobile application proposed in [11] for recognizing pigmented skin neoplasms based on convolutional neural networks achieved an accuracy of 75.2%. The studies published to date have used various software based on intelligent visual data classification algorithms. However, there are no software-based intelligent automated systems for analyzing heterogeneous dermatological data, which are images and statistical data of patients. Combining and analyzing heterogeneous dermatological data can significantly improve the accuracy of recognizing pigmented skin lesions.

This study aims to develop a cloud-based web application based on a multimodal neural network system for the analysis of heterogeneous dermatological data. The rest of the work is structured as follows. Section 2 describes the structure of a cloud-based web application based on a multimodal neural network system for processing statistical data and dermoscopic images of pigmented skin lesions. Section 3 presents the practical development of the proposed cloud-based web application based on a multimodal neural network architecture as an auxiliary system for the classification of pigmented neoplasms.

2 Materials and Methods

2.1 The Architecture of a Cloud-Based Web Application Based on a Multimodal Neural Network System for Recognizing Pigmented Skin Lesions

The proposed supporting diagnostic tool is a cloud-based neural network system that recognizes pigmented skin lesions using digital imaging and artificial intelligence methods. Figure 1 shows a proposed system consisting of a client device, a website for downloading dermatological data, and a server connected to both a cloud database and a script that downloads a neural network. To work correctly with the proposed system, the user uploads an image and statistical data from the device to the site. The neural network server is notified to load and process data using the neural network script. As a result of neural network recognition, the user is presented with classification data for 10 diagnostically significant categories of pigmented neoplasms.

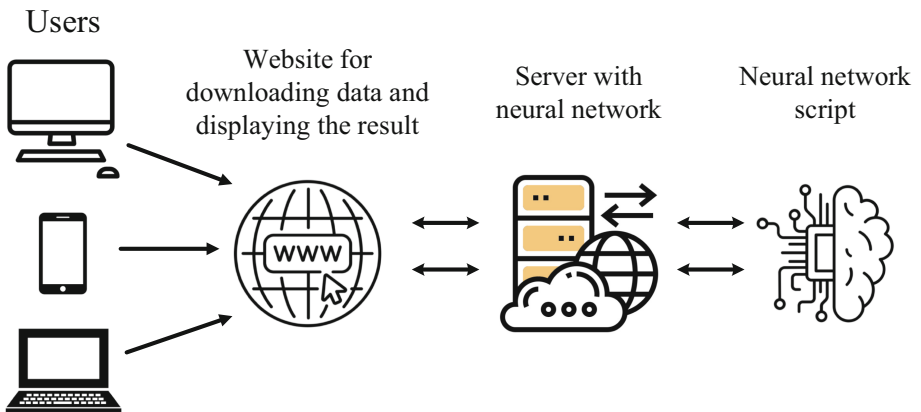


Fig. 1. Architecture of a cloud system for recognition of pigmented skin neoplasms

The website the user is accessing is a single-listing web page for downloading dermatological data. The only requirements for the website are to connect to a cloud database and resize the site depending on the device. The architecture of the cloud-based system for recognition of pigmented neoplasms of the skin was designed in such a way as to minimize memory and processor requirements and be available for a large number of devices with different characteristics. As a result, users with different types of devices can easily use the system when they have access to the Internet.

The task of the cloud server is to establish a connection between client devices and a script with a pre-trained neural network. Also need to store and maintain data coming from both directions. The server must be able to connect to a large number of user devices, store lesion images and statistics uploaded by users, and capture neural network recognition results. When a user uploads an image of a pigmented lesion and fills in the statistical data in the forms provided, a new catalog is created on the server with the image and patient data in a table format. In the next step, the server sends a script launch notification with a pre-trained neural network.

A script with a neural network is a loading of a trained multimodal neural network system for processing dermatological heterogeneous data. Combining heterogeneous data allows you to obtain additional information and improve the efficiency of medical neural network systems for analyzing and classifying heterogeneous data.

The proposed multimodal neural network system consists of a pre-trained neural network for processing visual data and a linear neural network for processing patient statistics. The input of the proposed system receives an image of a pigmented neoplasm I_d and statistical data of the patient C_d . The multimodal pigmented skin lesion recognition system is a fully connected neural network and requires input data encoded as a feature vector. For each image, a corresponding vector of statistical data is created, which depends on the amount and type of information. As a result, the patient statistics are hot-coded into a binary feature vector \vec{C}_d . The \vec{C}_d statistical feature vector is processed by a multilayer linear neural network architecture to obtain an output \vec{D} data vector.

The most optimal neural network architecture for the recognition of multidimensional visual data is CNN [12]. The main difference between this type of neural network is the convolution operation, which results in the receipt of image feature maps. As a result of passing through all the layers of the used pre-trained CNN, the image of the I_d pigment neoplasm is transformed into a I_m feature map. The input of the concatenation layer is the feature map of the I_m image and the \vec{D} data vector. The operation of combining heterogeneous data on the concatenation layer can be represented as follows:

$$fd = \sum_i \sum_j \sum_l k_{ijl} v_{ijld}^{(1)} + \sum_{i=1}^m D_i v_{id}^{(2)}, \tag{1}$$

where $v_{ijld}^{(1)}$ is a set of weights for processing feature maps of images and $v_{id}^{(2)}$ is a set of weights for processing \vec{D} data vector.

The last layer of the multimodal neural network is activated using the softmax function. The architecture of a multimodal neural network system for recognizing pigmented skin lesions based on CNN AlexNet is shown in Fig. 2.

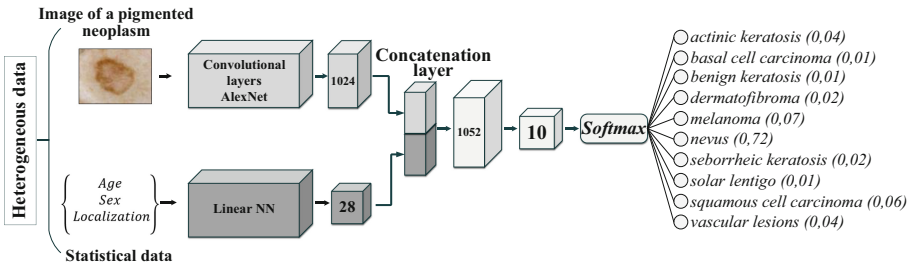




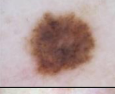

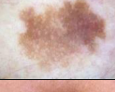


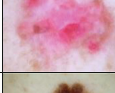


Fig. 2. Multimodal neural network architecture for classification of pigmented skin lesions based on AlexNet

3 Results

The data set used for training the multimodal neural network system was taken from the open archive of The International Skin Imaging Collaboration (ISIC) [13]. The dataset consists of 41,725 images of pigmented skin lesions of various sizes divided into the 10 most significant diagnostic categories. The dataset also contains statistical information about the patient's age group in increments of five years, the anatomical localization of the pigmented neoplasm in eight areas of the body, and gender. The rationale for using different types of data in training databases is that combining heterogeneous data can provide additional information and improve the efficiency of neural network analysis and classification systems. The use of heterogeneous dermatological data, such as images and statistical information about the patient, when training multimodal neural network systems can improve the accuracy of diagnostics by searching for links between visual objects of research and statistical metadata.

The statistical variable "Age" was divided into four groups by the classification of ages adopted by the World Health Organization. The first group of "young age" is represented by patients under the age of 44 years. The second group of "middle age" is represented by patients aged 45 to 59 years. The third group of "elderly age" is represented by patients aged 60 to 75 years. The fourth group of "senile age" is represented by patients over the age of 76 years. The database was divided into heterogeneous data for training and heterogeneous data for validation at a percentage of 80 to 20. Examples of images of pigmented skin lesions, as well as statistical data for these images from the modeling database, are presented in Table 1.

Table 1. Examples of images of pigmented skin lesions, as well as statistical data for these images from the selected database for neural network modeling

№	Image example	Category name	Statistics example			№	Image example	Category name	Statistics example		
			Age	Location	Gender				Age	Location	Gender
1.		vascular lesions	55	anterior torso	male	6.		benign keratosis	45	anterior torso	female
2.		nevus	35	lower extremity	female	7.		actinic keratosis	75	head	male
3.		solar lentigo	45	head	male	8.		basal cell carcinoma	85	anterior torso	male
4.		dermatofibroma	65	lower extremity	female	9.		squamous cell carcinoma	75	posterior torso	male
5.		seborrheic keratosis	75	upper extremity	male	10.		melanoma	65	upper extremity	male

Modeling of a multimodal neural network system for recognizing pigmented skin lesions was carried out using the high-level programming language Python 3.10.5. All calculations were carried out on a PC with an Intel(R) Core (TM) i5–8500 processor with a clock speed of 3.00 GHz, 16 GB of random-access memory (RAM) and a processor with a 64-bit Windows 10 operating system. Training multimodal neural network system was carried out using a graphics processing unit (GPU) based on an NVIDIA video chipset GeForce GTX 1050TI.

The choice of AlexNet as a CNN for processing visual data is because the neural network architecture does not require specialized hardware and works well with a limited GPU. The neural network architecture makes it possible to achieve high levels of image recognition accuracy by using a larger number of filters on each layer. In this case, the pooling layer follows each convolutional layer, and ReLU is used as the activation function, which reduces the probability of the disappearance of the gradient [14]. To train the architecture of a convolutional neural network based on AlexNet, images were converted to a size of $227 \times 227 \times 3$, where 3 is the number of color channels since the visual data is in RGB format.

Table 2 presents the results of assessing the recognition accuracy of dermoscopic data when testing a multimodal neural network system, as well as such quantification methods as loss function, F1-score, Matthew’s correlation coefficient (MCC), Jaccard-score, and specificity. The accuracy of recognition of pigmented skin lesions using CNN AlexNet was 85.65%. The loss function index when testing a multimodal neural network system was 0.12. The F1 score when testing the neural network architecture based on AlexNet was 0.86. The Jaccard index is used to compare a set of predicted labels. The Jaccard index for a multimodal neural network system based on CNN AlexNet was 0.75. The Matthews Correlation Coefficient (MCC) is the most reliable statistic showing how well a model performs in all four categories of the confusion matrix in proportion to the size of the positive elements and the size of the negative elements in the dataset. The value of the Matthews coefficient was 0.72. The specificity index for the neural network model was 0.98.

Table 2. Results of modeling multimodal neural network system for recognizing pigmented neoplasms

CNN architecture	Loss function	Accuracy, %	F1- score	MCC	Jaccard-score	Specificity
AlexNet	0.12	85.65	0.86	0.72	0.75	0.98

The Flask framework was used to create a web application for recognizing pigmented skin lesions. Figure 3 shows the appearance of the dermatological data download page of the proposed web application, as well as the page with the result displayed to the user after the neural network recognition of a pigmented skin neoplasm.

Because the proposed web application should not include a large number of functionality, as well as the need for compatibility with the PyTorch, OpenCV, and Torchvision modules used, the Flask framework is the most optimal framework for creating web

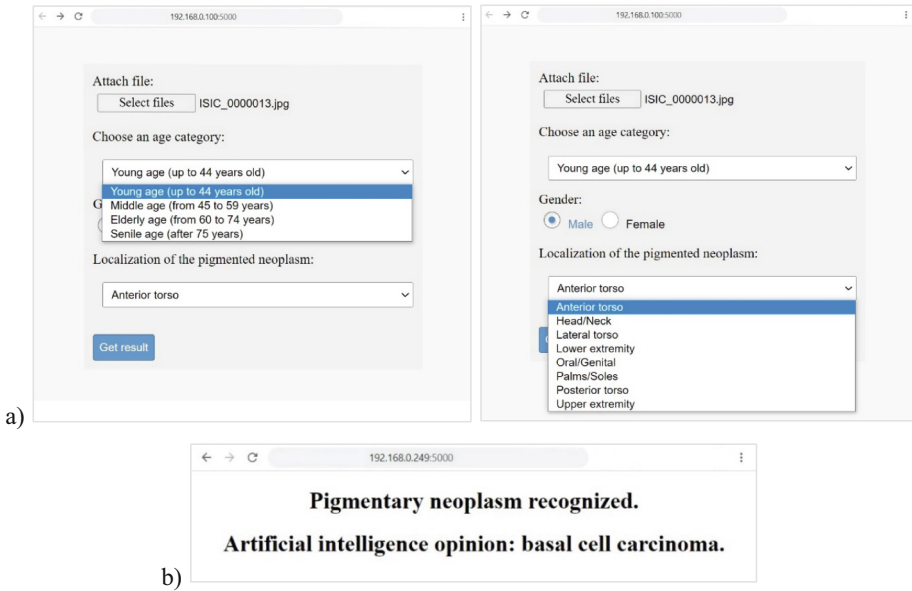


Fig. 3. Appearance of the proposed web application: a) page for downloading dermatological data; b) the result displayed to the user

applications. The HTML markup language and CSS Cascading Style Sheets were used to create the frontend. The page structure of a web application is HTML files located on the server. Files are issued to the browser through a request to the server with HTML templates. The appearance of the pages is described using CSS.

The introduction of the proposed web application as an auxiliary diagnostic tool can significantly change tactics in the prevention of skin cancer and early detection of melanoma. Unlike other types of cancer, which are localized within the body, the location of pigmented neoplasms allows diagnosis using non-invasive approaches. The use of the proposed web application as an auxiliary diagnostic tool opens up prospects for accelerated diagnosis of the patient, and detection of skin cancer at an early stage.

4 Conclusion

As a result of testing the proposed web application based on a multimodal neural network system with CNN AlexNet, the accuracy of recognition of skin neoplasms was 85.65%. The proposed system can be used as an effective auxiliary diagnostic tool for a dermatologist-oncologist.

Efficiency in the implementation of applications based on artificial intelligence in the field of medicine has been demonstrated in recent studies. Applications based on intelligent systems in the field of dermatology are becoming increasingly important [15] and can be especially effective for auxiliary diagnostics of a large number of patients. The introduction of artificial intelligence approaches as an auxiliary diagnostic tool makes

it possible to increase the efficiency of skin cancer recognition in comparison with the visual diagnostic methods of dermatologists-oncologists [16].

The proposed web application demonstrates more accurate results in recognizing pigmented skin lesions compared to existing similar intelligent diagnostic systems. The cloud system presented in [10] visual data of pigmented neoplasms with an accuracy of up to 77.40%, which is 8.25% lower than the accuracy of the proposed web application based on a multimodal neural network system. The work [11] presents a mobile application for recognizing pigmented skin lesions with an accuracy of up to 75.20%, which is 10.45% lower than the accuracy of the proposed web application. The main distinguishing characteristic of the proposed auxiliary web application is the use of heterogeneous data based on visual data and statistical information about patients. Combining and analyzing heterogeneous information can significantly improve the accuracy of neural network recognition by searching for additional links between a diagnostic category and available dermatological data.

The main limitation in the implementation of the proposed web application based on a multimodal neural network system for recognizing pigmented lesions is the possibility of using it only as an additional diagnostic tool for physicians and specialists. The proposed web application cannot independently diagnose patients and is not a full-fledged certified medical tool. A promising direction for further research is the construction of more complex neural network systems for the classification of pigmented skin lesions. The development and implementation of methods for segmenting pigmented neoplasms in images, as well as methods for cleaning noise and hair structures as additional functions in the proposed web application, will improve the accuracy and efficiency of recognizing pigmented neoplasms.

Acknowledgments. The research in Sect. 3 was supported by Russian Science Foundation, project 22-71-00009. The authors would like to thank the North-Caucasus Federal University for supporting the contest of projects competition of scientific groups and individual scientists of the North-Caucasus Federal University.

References

1. Health consequences of excessive solar UV radiation: <https://www.who.int/news/item/25-07-2006-health-consequences-of-excessive-solar-uv-radiation>
2. Mihm, M.C., Clark, W.H., From, L.: The clinical diagnosis, classification and histogenetic concepts of the early stages of cutaneous malignant melanomas. *N. Engl. J. Med.* **284**, 1078–1082 (1971). <https://doi.org/10.1056/NEJM197105132841907>
3. Tromme, I., et al.: Availability of digital dermoscopy in daily practice dramatically reduces the number of excised melanocytic lesions: results from an observational study. *Br. J. Dermatol.* **167**, 778–786 (2012). <https://doi.org/10.1111/J.1365-2133.2012.11042.X>
4. Nami, N., Giannini, E., Burrioni, M., Fimiani, M., Rubegni, P.: Teledermatology: state-of-the-art and future perspectives. *Expert Rev. Dermatol.* **7**(1), 1–3 (2014). <https://doi.org/10.1586/edm.11.79>
5. Swanson, N.A., Lee, K.K., Gorman, A., Lee, H.N.: Biopsy techniques: diagnosis of melanoma. *Dermatol. Clin.* **20**, 677–680 (2002). [https://doi.org/10.1016/S0733-8635\(02\)00025-6](https://doi.org/10.1016/S0733-8635(02)00025-6)

6. Bong, J.L., Herd, R.M., Hunter, J.A.A.: Incisional biopsy and melanoma prognosis. *J. Am. Acad. Dermatol.* **46**, 690–694 (2002). <https://doi.org/10.1067/MJD.2002.123488>
7. Korotkov, K., Garcia, R.: Computerized analysis of pigmented skin lesions: a review. *Artif. Intell. Med.* **56**, 69–90 (2012). <https://doi.org/10.1016/J.ARTMED.2012.08.002>
8. Brinker, T.J., Hekler, A., Enk, A.H., Klode, J., Hauschild, A.: Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur. J. Cancer* **113**, 47–54 (2019). <https://doi.org/10.1016/J.EJCA.2019.04.001>
9. Amelard, R., Glaister, J., Wong, A., Clausi, D.A.: High-Level Intuitive Features (HLIFs) for intuitive skin lesion description. *IEEE Trans. Biomed. Eng.* **62**, 820–831 (2015). <https://doi.org/10.1109/TBME.2014.2365518>
10. Akar, E., Marques, O., Andrews, W.A., Furht, B.: Cloud-based skin lesion diagnosis system using convolutional neural networks. *Adv. Intell. Syst. Comput.* **997**, 982–1000 (2019). https://doi.org/10.1007/978-3-030-22871-2_70/FIGURES/13
11. Dai, X., Spasic, I., Meyer, B., Chapman, S., Andres, F.: Machine learning on mobile: an on-device inference app for skin cancer detection. In: 2019 4th International Conference on Fog and Mobile Edge Computing, FMEC 2019, pp. 301–305 (2019). <https://doi.org/10.1109/FMEC.2019.8795362>
12. Li, H., Pan, Y., Zhao, J., Zhang, L.: Skin disease diagnosis with deep learning: a review. *Neurocomputing* **464**, 364–393 (2021). <https://doi.org/10.1016/J.NEUCOM.2021.08.096>
13. ISIC Archive: <https://www.isic-archive.com/#!/topWithHeader/wideContentTop/main>
14. Hosny, K.M., Kassem, M.A., Foad, M.M.: Classification of skin lesions using transfer learning and augmentation with Alex-net. *PLoS ONE* **14**, e0217293 (2019). <https://doi.org/10.1371/JOURNAL.PONE.0217293>
15. Koh, U., et al.: Consumer acceptance and expectations of a mobile health application to photograph skin lesions for early detection of melanoma. *Dermatology* **235**, 4–10 (2019). <https://doi.org/10.1159/000493728>
16. Rat, C., Hild, S., Sérandour, J.R., Gaultier, A., Quereux, G., Dreno, B., Nguyen, J.-M.: Use of smartphones for early detection of melanoma: systematic review. *J. Med. Internet Res.* **20**(4), e135 (2018). <https://doi.org/10.2196/jmir.9392>



Temperature Profile Nowcasting Using Temporal Convolutional Network

Nikolay Baranov^(✉)

Dorodnicyn Computing Centre, FRC CSC RAS, Moscow 119333, Russia
baranov@ians.aero

Abstract. The paper deals with the short-term forecasting problem of the temperature profile based on observational data. The MTP-5 temperature profiler is the observational data source. This remote sensor provides measurement of the temperature profile in the surface layer of the atmosphere with a high spatiotemporal resolution. Measurement data is considered as a multivariate time series. We use a temporal convolutional neural network (TCN) to prediction such a series. A quality analysis of the temperature profiles forecast for several hours using TCN is presented.

Keywords: forecasting · temperature profile · temporal convolutional network

1 Introduction

The problem of forecasting the state of the atmosphere is always relevant and is solved using high-resolution numerical weather prediction models. These models have different scales and are based on solving equations in partial derivatives of atmospheric dynamics. One of the problems of numerical prediction models is the low accuracy of determining the profiles of meteorological parameters of the surface layer of the atmosphere, in particular, temperature profiles. At the same time, the vertical distribution of temperature in the lower layer of the atmosphere is critical for practical purposes. In particular, it is important for the ecology of cities, since it determines the dispersion of pollutants. It is known that temperature inversions contribute to the pollutants accumulation in the atmosphere surface layer [1, 2].

Mesoscale numerical forecasting models do not always effectively determine the inversions in the surface layer of the atmosphere. As an example, Fig. 1 shows a comparison of the forecasting data from the Global Forecast System model and the measurement results of the MTP-5 temperature profiler. The Global Forecast System (GFS) is a National Centers for Environmental Prediction (NCEP) weather forecast model. Comparison data are presented for a point with coordinates 56.143°N, 34.989°E with a three-hour step during the day. We see that the numerical prediction model predicts an adiabatic temperature profile during the day. At the same time, observational data show a stable elevated inversion

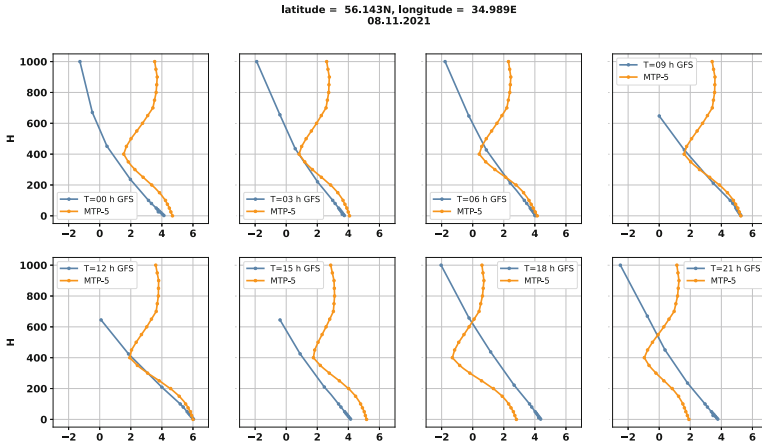


Fig. 1. Comparison of forecasting and observation data for temperature profiles

at 400m. The presented results clearly show that the numerical forecast model does not allow predicting the temperature inversion at low altitudes.

In this regard, the problem of the temperature inversions nowcasting using observational data is topical. In particular, this issue was the subject of a paper [3] in which the prognostic temperature at each altitude is calculated on the basis of the weighted-average trend. In its turn, the value of the weighted-average trend is calculated as a combination of the observational data trend and the modeling data trend taken with their weights. Weight coefficients are calculated based on the analysis of the quality of the forecast for the previous observation period.

In this paper, we will consider the use of neural networks to solve the problem of short-term forecasting of temperature profiles based on observational data. In contrast to [1], here we use only observational data for nowcasting. As is known, convolutional neural networks (CNN) allow extracting spatial characteristics. This property determines their wide application for classification problems. On the other hand, long short-term memory (LSTM) neural networks can learn temporal characteristics. Therefore, there are many options for combining CNN and LSTM networks. In practice, it is better to study temporal characteristics from data sequences with a long history due to the fact that the patterns of trend change appear over large time intervals. However, LSTM is less efficient when processing a long sequence of history [4]. To solve this issue, this work uses a temporal convolution network combining causal filters with extended convolutions to increase the length of the network's receptive field.

The temporal convolution network (TCN) were first proposed in the work [5] for video-based action segmentation and have been rapidly developed [6, 7]. In 2020, a paper [8] was published on the use of TCN in weather forecasting problems. Based on the comparison of TCN and LSTM, the authors concluded that TCN performs well in time series forecasting tasks. We also note the works [9, 10] devoted to various problems of meteorological forecasting based on TCN.

Let us list the advantages of TCN, following work [6].

- **Parallelism.** A long input sequence can be processed as a whole in TCN, instead of sequentially as in RNN.
- **Flexible receptive field size.** A TCN can change its receptive field size in multiple ways: stacking more dilated (causal) convolutional layers, using larger dilation factors, or increasing the filter size. TCNs thus afford better control of the model’s memory size, and are easy to adapt to different domains.
- **Stable gradients.** TCN avoids the problem of exploding/vanishing gradients, which is a major issue for RNNs.
- **Low memory requirement for training.**
- **Variable length inputs.** Just like RNNs, TCNs can also take in inputs of arbitrary lengths by sliding the 1D convolutional kernels.

2 Temporal Convolution Network

We give a brief description of the TCN architecture following the work [6]. Let the input sequence be given

$$x_0, x_1, \dots, x_T. \tag{1}$$

The problem is to predict corresponding outputs at each time:

$$\hat{y}_0, \hat{y}_1, \dots, \hat{y}_T. \tag{2}$$

The TCN uses 1-dimensional fully-convolutional network architecture. Each hidden layer is the same length as the input layer, and zero padding of length (kernel size - 1) is added to keep subsequent layers the same length as previous ones. To achieve the second point, the TCN uses causal convolutions, convolutions where an output at time t is convolved only with elements from time t and earlier in the previous layer (the blue line connections at the Fig. 2).

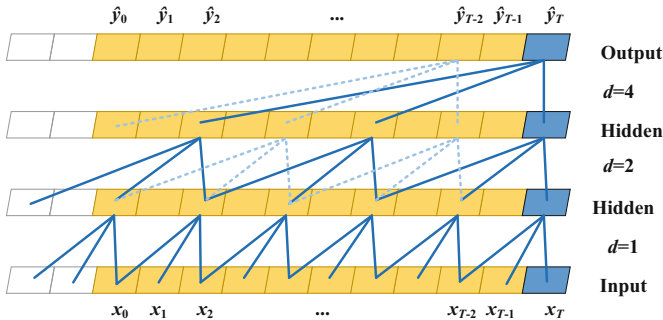


Fig. 2. A dilated causal convolution with dilation factors $d = 1, 2, 4$ and filter size $k = 3$ [6]

A simple causal convolution is only able to look back at a history with size linear in the depth of the network. Therefore, TCN uses dilated convolutions that enable an exponentially large receptive field. More formally, for a 1-D sequence input $x \in \mathbb{R}^n$ and a filter $f: \{0, \dots, k-1\} \rightarrow \mathbb{R}$, the dilated convolution operation F on element s of the sequence is defined as

$$F_s = x_{*d} f(s) = \sum_{i=0}^{k-1} f(i) \cdot x_{s-d \cdot i}, \quad (3)$$

where d is the dilation factor, k is the filter size.

The residual block allows us to consider the output as a modification of the identity transformation:

$$O = \text{activation function}(x + F(x)) \quad (4)$$

Calculating the increment of the input is more efficient than directly calculating the value of the output, especially when the dimension of the input is large. As activation function we use the rectified linear unit *ReLU*:

$$\text{ReLU}(x) = \max(0, x). \quad (5)$$

3 Problem Statement

The problem of forecasting temperature profiles is formulated as follows.

At the moment of time t_0 there are time-series of observations of temperature values $\{T_k^{(i)}\}$ at given heights h_k , $k = 1, \dots, n$, $i = 0, -1, \dots, -m$, where $T_k^{(i)}$ is a temperature value at altitude h_k at time moment t_i . Observations are carried out with a constant time interval:

$$t_i = t_0 + i \cdot \Delta t. \quad (6)$$

We will assume that a fixed forecasting interval $\Delta\tau$ is given.

It is required for each height h_k to calculate the temperature estimate $\hat{T}_k(\Delta\tau)$ at the time moment $t_0 + \Delta\tau$ as a function of observational data:

$$\hat{T}_k(\Delta\tau) = F_k^{(\Delta\tau)}(T_1^{(0)}, \dots, T_n^{(0)}, T_1^{(-1)}, \dots, T_n^{(-1)}, \dots, T_1^{(-m)}, \dots, T_n^{(-m)}). \quad (7)$$

Note that, in accordance with the problem statement, we assume that the forecast for a fixed height h_k is calculated from observational data at all heights h_l , $l = 1, \dots, n$.

MTP-5 measures the temperature at the altitudes:

$$\begin{aligned} h_k &= 25 \cdot (k - 1), \text{ if } k < 5, \\ h_k &= 100 + 50 \cdot (k - 5), \text{ if } 5 \leq k \leq 23. \end{aligned} \quad (8)$$

The total number of measurement heights is 23. We are testing the altitude range up to 600m, which is most to temperature anomalies.

The nowcasting time is equal to 2h.

The receptive field size of TCN is defined by next parameters:

- k - the kernel size to use in each convolutional layer;
- d - the number of filters to use in the convolutional layers;
- n_r - the number of stacks of residual blocks to use.

The receptive field for $k = 2, d = 4, n_r = 1$ is shown on Fig. 3. The yellow circles correspond to the input sequence. The red circles correspond to the output of the residual blocks. The blue circles show the elements involved in the formation of the output of the residual block. Accordingly, the receptive field for $k = 2, k = 4, n_r = 2$ is shown on Fig. 4.

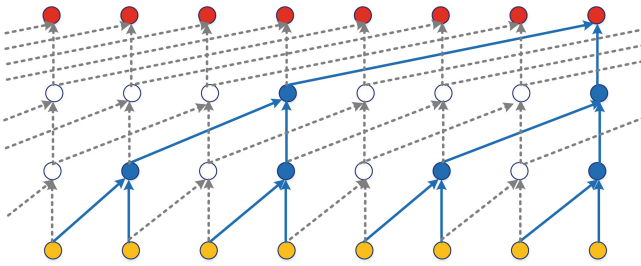


Fig. 3. The receptive field for $k = 2, d = 4, n_r = 1$

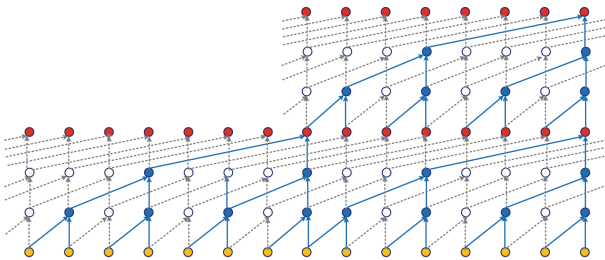


Fig. 4. The receptive field for $k = 2, d = 4, n_r = 2$

The size of the convolution kernel determines the local temperature trend. Small kernel sizes ($k = 2, 3, 4$) highlight random measurement fluctuations, so it makes no sense to use them. We use the size of the convolution kernel equal to 7, which corresponds to an observation interval of 30 min. In turn, the dilation factor determines the time interval on which we identify the dynamics of a larger time scale, for example, daily. With an expansion factor of $d=8$, TCN with a single residual block analyzes observations over approximately 4.5 h. With $d = 32$, the observation interval will expand to 18.5 h, and with $d = 64$, respectively, up to a day and a half.

If the length of the input sequence is less than required, the missing elements are padding with zeros.

For training, we use measurement data for the period july-august 2019, performed in Novosibirsk. Testing of the trained neural network is performed on measurement data for the same period of 2020.

4 Results and Discussion

We use TCN with $k = 7, d = 32, n_r = 1$. The convergence of the learning process is shown in Fig. 5. As an accuracy metrics, we use the mean absolute error. Figures 6, 7, 8 and 9 show examples of nowcasting using a trained TCN for some typical temperature stratification situations.

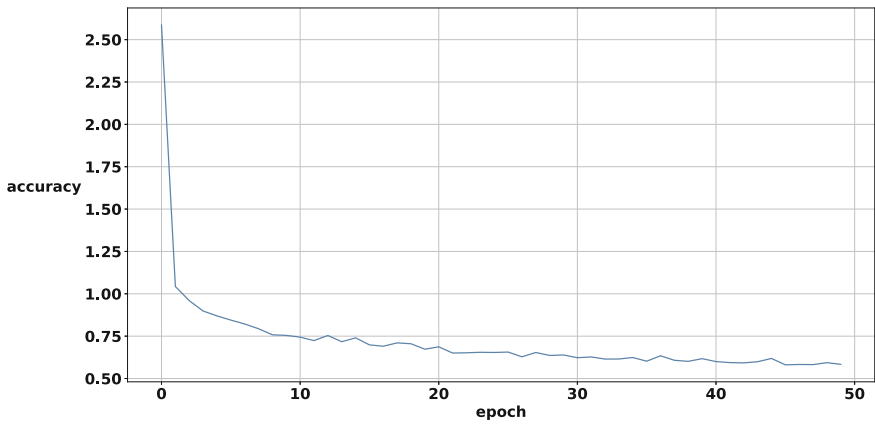


Fig. 5. The convergence of the learning process

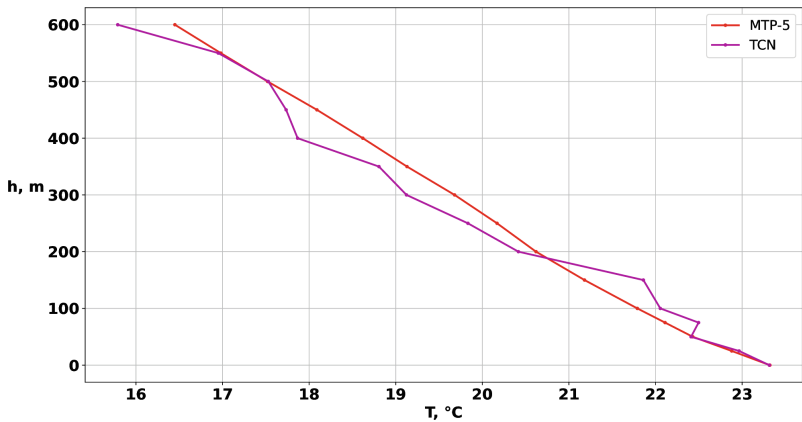


Fig. 6. Nowcasting of adiabatic temperature profile

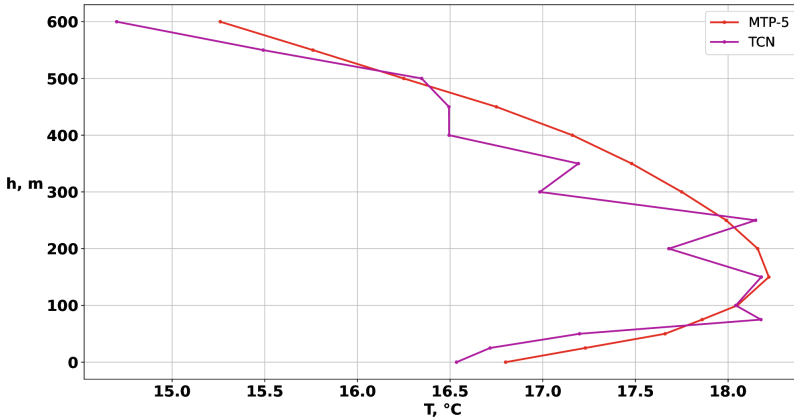


Fig. 7. Nowcasting of temperature inversion

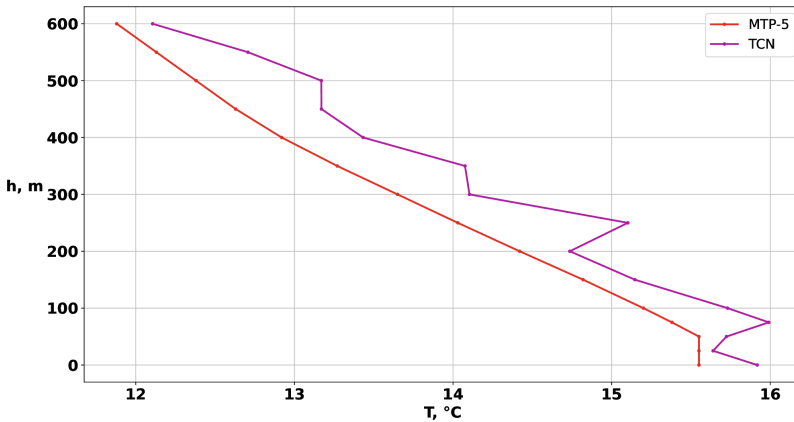


Fig. 8. Nowcasting of temperature stratification with surface isotherm

It can be seen that, in general, TCN predicts the type of temperature stratification quite well. At the same time, the greatest quantitative difference between the forecast and observational data is observed close to the earth's surface. It can be seen that, in general, TCN predicts the type of temperature stratification quite well. In this case, the greatest quantitative difference between the forecast and observational data is observed close to the earth's surface. As the height increases, the forecast is more consistent with the observational data. A higher forecasting error at low altitudes is explained by the influence of the process of heating the earth's surface on the evolution of air temperature.

Statistical analysis of the results of forecasting and observation confirms the comments made. Figures 10, 11, 12 and 13 show histograms of the forecasting error distribution for different heights. It can be seen that at higher altitudes the range of forecasting errors is becoming smaller. If in the range of heights up

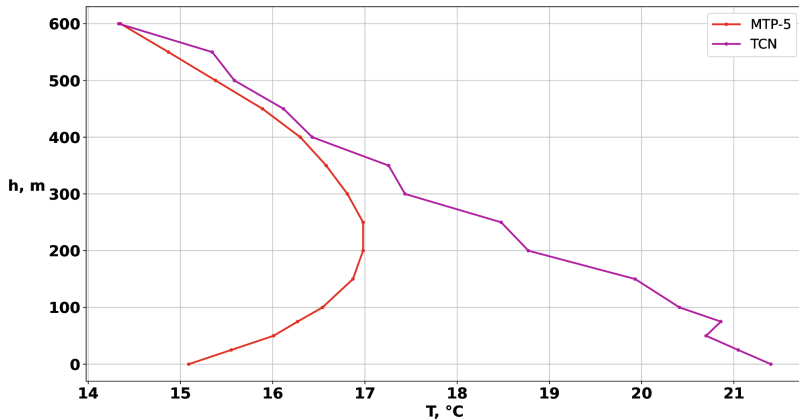


Fig. 9. Incorrect prediction of temperature inversion

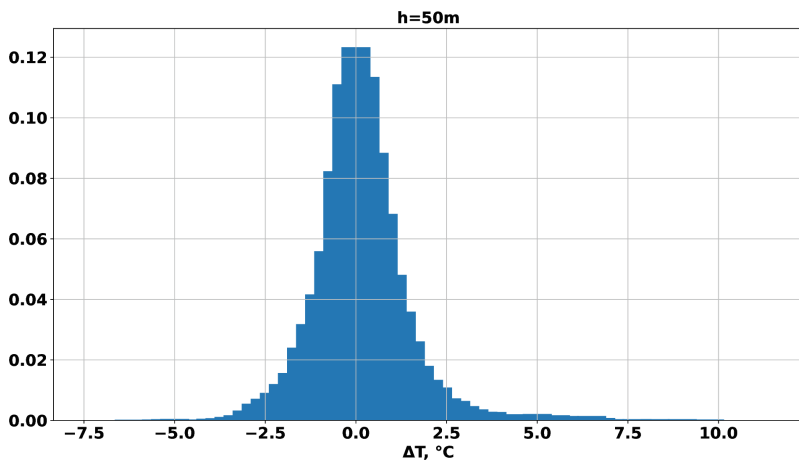


Fig. 10. Histogram of the forecasting error distribution for $h = 50$ m

to 100 m the root-mean-square forecasting error is 1.3...1.8 °C, then at heights over 400 m it is less than 1 °C (Fig. 14).

As you can see from the results presented, the forecasting temperature profile calculated using TCN is characterized by temperature fluctuations with heights. This is because TCN analyzes the temperature profile as a collection of independent time series. The absence of spatial convolution of the multidimensional input sequence in height does not allow taking into account the correlation of temperature changes at neighboring heights.

In this work, we did not optimize the network parameters, choosing its characteristics from qualitative considerations. However, the presented results show that TCN can be an effective tool for short-term prediction of temperature stratification of the surface layer of the atmosphere.

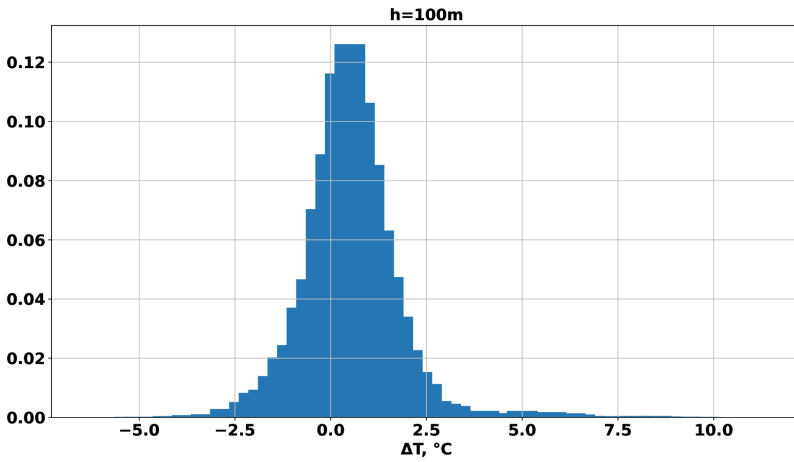


Fig. 11. Histogram of the forecasting error distribution for $h=100m$

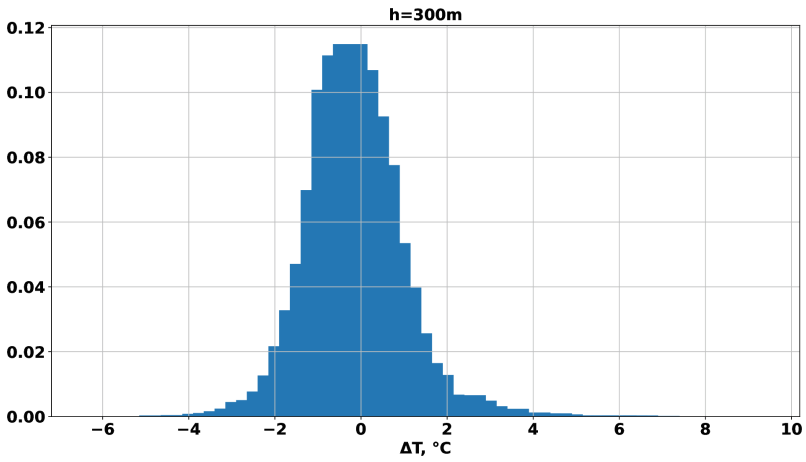


Fig. 12. Histogram of the forecasting error distribution for $h = 300 m$

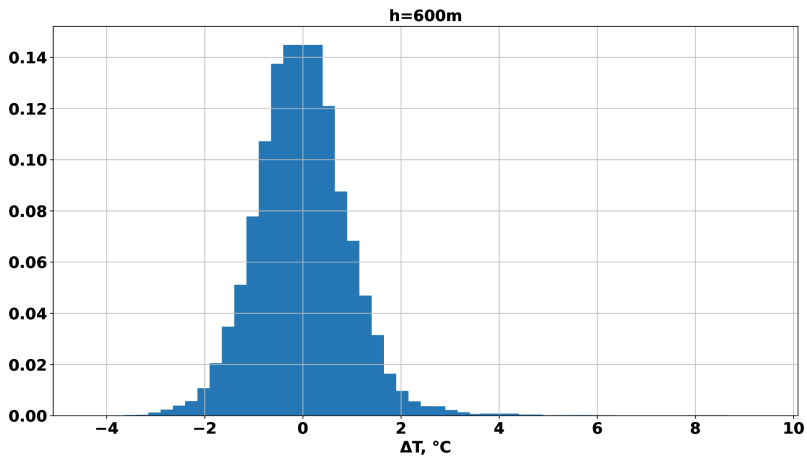


Fig. 13. Histogram of the forecasting error distribution for $h = 600$ m

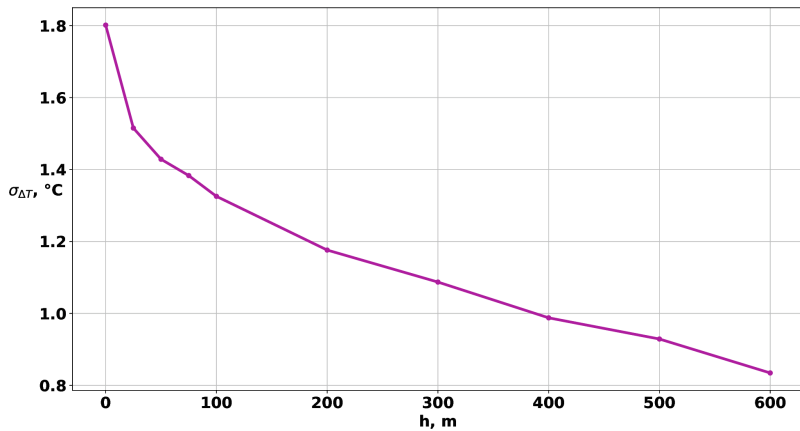


Fig. 14. Root-mean-square forecasting error as function of altitude

References

1. Rendón, A.M., Salazar, J.F., Palacio, C.A., Wirth, V.: Temperature inversion breakup with impacts on air quality in urban valleys influenced by topographic shading. *J. Appl. Meteorol. Climatol.* **54**(2), 302–321 (2015). <https://doi.org/10.1175/JAMC-D-14-0111.1>
2. Trinh, T.T., Trinh, T.T., Le, T.T., Nguyen, T.D.H., Tu, B.M.: Temperature inversion and air pollution relationship, and its effects on human health in Hanoi City, Vietnam. *Environ. Geochem. Health* **41**(2), 929–937 (2018). <https://doi.org/10.1007/s10653-018-0190-0>
3. Baranov, N.A., Lemishchenko, E.V.: Forecasting temperature profile based on blending of measurement data and numerical prediction models. *Int. J. Circuits Syst. Signal Process.* **12**, 235–239 (2018)

4. Gers, F.: Long short-term memory in recurrent neural networks. EPFL Theses. 2366 (2001). <https://doi.org/10.5075/epfl-thesis-2366>
5. Lea, C., Flynn, M. D., Vidal, R., Reiter, A., Hager, G. D.: Temporal convolutional networks for action segmentation and detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1003–1012 (2017). <https://doi.org/10.1109/CVPR.2017.113>
6. Bai, S., Kolter, J.Z., Koltun, V.: An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. [arXiv:1803.01271v2](https://arxiv.org/abs/1803.01271v2). <https://doi.org/10.48550/arXiv.1803.01271>
7. Li, Y., Song, L., Zhang, S., Kraus, L., Adcox, T., Willardson, R., Lu, N.: A TCN-based Spatial-Temporal PV Forecasting Framework with Automated Detector Network Selection. [arXiv:2111.08809](https://arxiv.org/abs/2111.08809). <https://doi.org/10.48550/arXiv.2111.08809>
8. Yan, J., Mu, L., Wang, L., et al.: Temporal convolutional networks for the advance prediction of ENSO. *Sci. Rep.* **10**, 8055 (2020). <https://doi.org/10.1038/s41598-020-65070-5>
9. Jiang, Y., Zhao, M., Zhao, W., Qin, H., Qi, H., Wang, K., Wang, C.: Prediction of sea temperature using temporal convolutional network and LSTM-GRU network. *Complex Eng. Syst.* **1**, 6 (2021). <https://doi.org/10.20517/ces.2021.03>
10. Hewage, P., Behera, A., Trovati, M., Pereira, E., Ghahremani, M., Palmieri, F., Liu, Y.: Temporal convolutional neural (TCN) network for an effective weather forecasting using time-series data from the local weather station. *Soft. Comput.* **24**(21), 16453–16482 (2020). <https://doi.org/10.1007/s00500-020-04954-0>



Application of Bidirectional LSTM Neural Networks and Noise Pre-cleaning for Feature Extraction on an Electrocardiogram

Mariya Kiladze^(✉) 

North-Caucasus Federal University, Stavropol, Russia
merchali@mail.ru

Abstract. Cardiac diseases are one of the most common diseases on the planet. Thousands of people die from this disease every year. For prompt diagnosis, an automated system for processing electrocardiograms is required. The standard model of an automated system consists of signal preprocessing, feature extraction, and classification. In this article, unidirectional and bidirectional network models with long short-term memory were considered for the classification of electrocardiogram signals. The simulation results showed that the use of both methods without preliminary signal processing and feature extraction on them is not advisable. Also, the simulation result showed that models that include the removal of noise from electrocardiograms have more accurate training results for bidirectional networks with a long short-term memory. The simulation was carried out in the MatLab 2020b mathematical environment based on the PhysioNet Computing in Cardiology Challenge 2017 database, taken from an open source. The best result was obtained in the classification of atrial fibrillation.

Keywords: Electrocardiogram · Long Shot-Term Memory · neural network · signal noise reduction · feature extraction

1 Introduction

The number of people suffering from cardiac diseases has increased 4.5 times from 2000 to 2019 [1]. Electrocardiography (ECG) allows timely diagnosis of these diseases. An increase in the volume of electrocardiograms requires automation of the process of their decoding. For example, automatic detection of atrial fibrillation by ECG signals will allow doctors to quickly pay attention to a particular patient.

The most common method for identifying signs of cardiac dysfunction on ECG is wavelet analysis followed by training a neural network for classification. For example, in [2], the authors highlight the P-peak on the ECG using the Daubechies wavelet transform. This is one of the first works devoted to the selection of signs of atrial fibrillation. In [3], the authors use an adaptive wavelet transform based on the Shannon method and the Poincret section.

Pacemakers use techniques to process signals in real time. The method based on the use of biorthogonal wavelet transform is described in [4].

One of the methods for detecting atrial fibrillation [5] is based on the use of an eight-layer convolutional network, which does not require either pre-processing of the ECG signal or extracting features from it. In contrast to it, there is a method [6], in which classification is carried out by a supervised learning algorithm, and feature extraction is based on several feature sets (Adreotti, Zabiha feature set, aggregated feature set, and Dutt feature set).

In [7], we described a method for determining atrial fibrillation in an ECG stream, consisting of signal preprocessing, noise removal from the signal using wavelet transform, detection of atrial fibrillation signs based on instantaneous frequency and spectral analysis of the signal, followed by the use of Long Shot-Term Memory (LSTM) networks for ECG classification. The question arises: is it advisable to remove noise from ECG signals and use spectral analysis to extract features, or is it enough to use a single LSTM classifier. This article is dedicated to this issue.

2 Materials and Methods

2.1 Model Description

In [7], we describe our proposed method for classifying ECG signals, which consists of the following algorithm:

1. Signal preprocessing.
2. Noise removal using wavelet transform.
3. Identification of the sign P - peak using spectral analysis.
4. Signal classification using LSTM network.

The question arose about the advisability of applying paragraphs 2 and 3 to solve this problem. In this article, we will compare the simulation results of this method using different LSTM networks and different number of classifiers without applying steps 2 and 3.

Thus, the general form of the modeling algorithm will take the form:

1. Signal preprocessing.
2. Feature extraction (this item is not present in all models)
3. Signal classification by different LSTM networks.

The modeling was carried out according to the following schemes: unidirectional LSTM network (Fig. 1(a)), bidirectional LSTM network (Fig. 1(b)), unidirectional LSTM network with preliminary feature extraction on the ECG signal (Fig. 1(c)), bidirectional LSTM network with preliminary feature extraction on the signal ECG (Fig. 1(d)).

2.2 Electrocardiogram Signals

Electrocardiogram signals are the result of electrocardiography, a technique for recording and studying electric fields generated during the work of the heart [8, 9]. The potential difference is measured by electrodes located on different parts of the body: suction cups

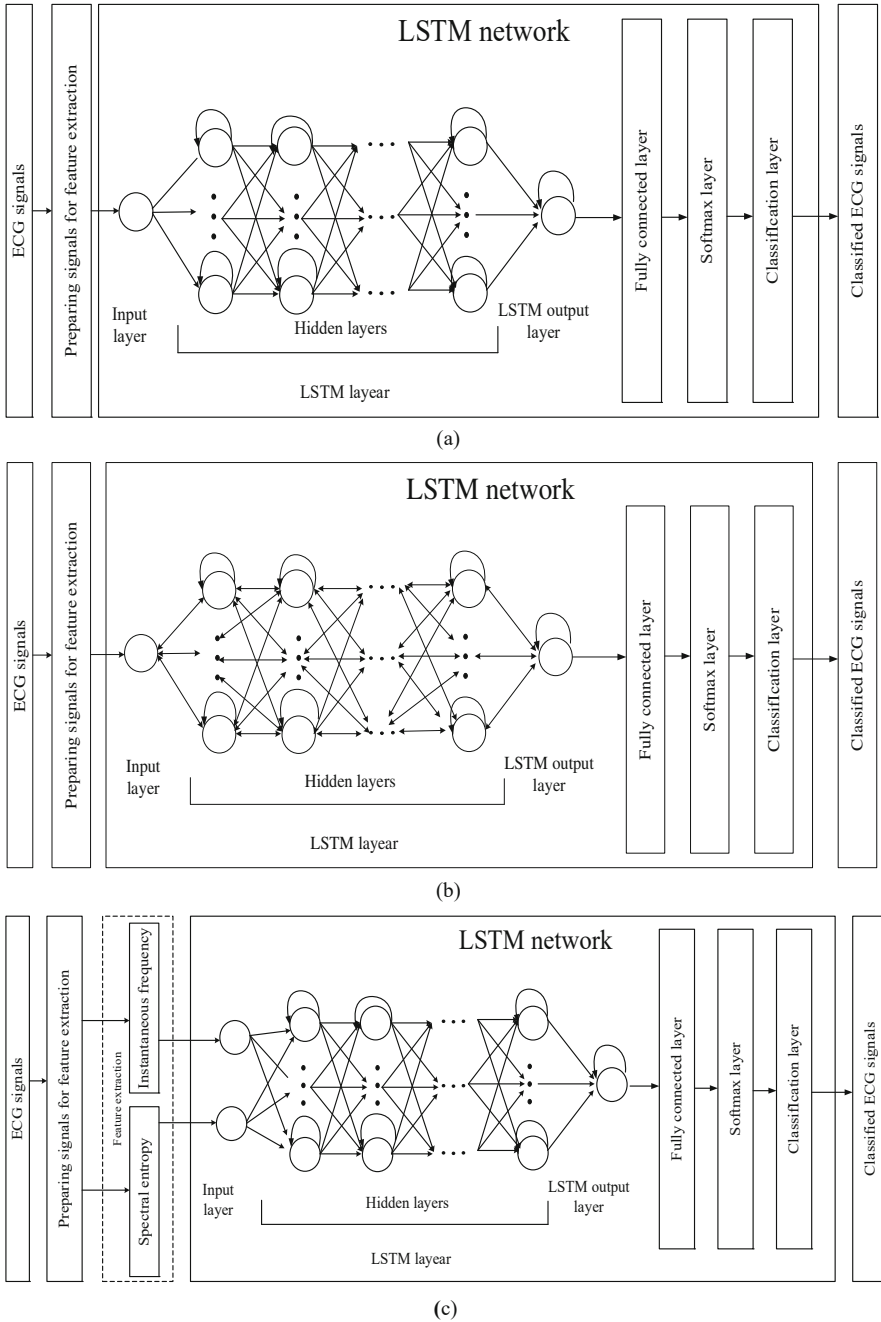


Fig. 1. LSTM network: a) Unidirectional, b) Bidirectional, c) Unidirectional LSTM network with pre-feature extraction, d) Bidirectional LSTM network with pre-feature extraction.

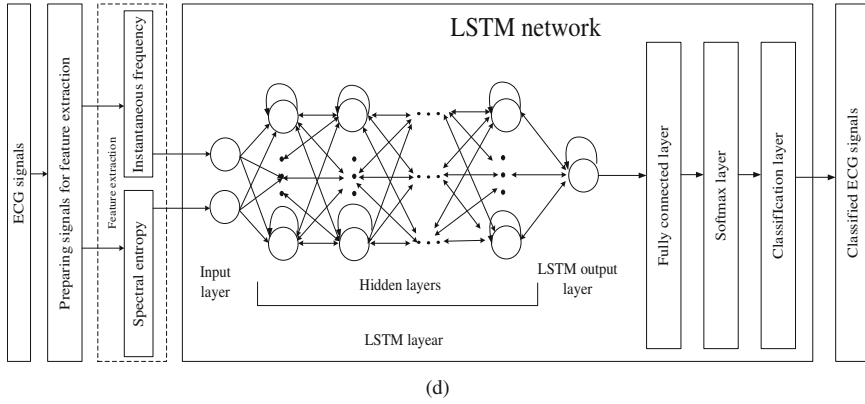


Fig. 1. (continued)

are placed on the chest, plastic tweezers-clamps are on the limbs. To reduce signal noise, a conductive gel is applied to the skin at the points of contact [8, 9]. Leads I, II and III are superimposed on the limbs, leads V1 - V9 are located on the chest according to Fig. 2 [10].

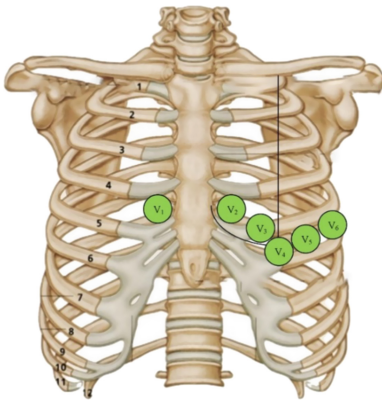


Fig. 2. Location of chest leads [10].

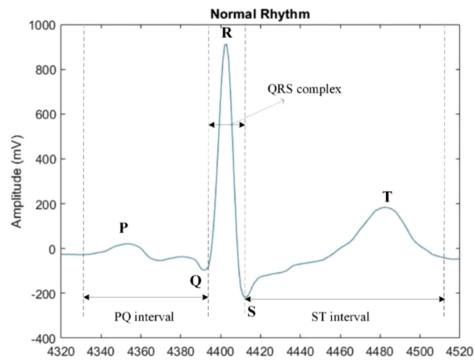


Fig. 3. ECG waves

In modern electrocardiographs for primary noise reduction, low-frequency, notch and anti-tremor filters are built in [9].

On Fig. 3 shows 5 ECG waves: P, Q, R, S, T. Depolarization of the atrial myocardium is described by the P wave, ventricular depolarization by the QRS complex, repolarization processes by the ST segment and the T wave.

The absence of P waves on the ECG, the presence of atrial fibrillation waves, different R-R intervals, the presence of atrial fibrillation waves, different R-R intervals, the heart rate is unchanged or accelerated - the main signs of atrial fibrillation. In [7], we determined the presence of P waves and R-R intervals, having previously cleared the signal from noise by a discrete wavelet transform.

2.3 Signal Preprocessing

Preprocessing is necessary for correct classification of ECG signals. The number of ECG signals must be the same in each stream. To equalize the number of counts of ECG signals, you must:

1. Calculate the number of ECG signals with the same number of readings.
2. Select a group consisting of the largest number of ECG signals with the same number of readings x .
3. Groups of signals consisting of signals with the number of samples less than in the selected group are deleted. Signal groups consisting of signals with more samples than in the selected group are divided into signals with the number of samples from the reference group, the rest are removed [7].

2.4 LSTM Network

The LSTM network is a recurrent neural network with a long short-term memory, suitable for processing time-dependent data such as physiological signals. In the schemes presented by us, unidirectional LSTM networks (Fig. 1(a) and (c)) and bidirectional LSTM networks (Fig. 1(b) and (d)) are used.

Unidirectional networks during training move only forward through the layers and elements of the network, while bidirectional networks can return to previous layers and elements.

Each element of the LSTM network goes through four stages: determining the required information (sigmoid layer), determining the relevance of information and setting a new candidate vector (sigmoid layer and hyperbolic tangent layer), saving a new candidate vector and determining the necessary information for its subsequent output from the element (sigmoid layer and the layer of hyperbolic tangent) [11].

For circuits without signal preprocessing, the LSTM network element looks like in Fig. 4(a). The signal enters a one-dimensional input vector and further training takes place. For circuits with signal preprocessing, the LSTM network element looks like in Fig. 4(b). The input two-dimensional vector receives two signals of the same length with pre-selected features. Further training proceeds according to the same scheme, where h_{t-1} and h_t - weekend vectors ($h_0 = 0$), C_{t-1} and C_t - state vectors ($C_0 = 0$), σ - sigmoidal activation function, \tanh - activation function based on hyperbolic tangent, \otimes - multiplication operator, \oplus - addition operator.

The vector f_t determines whether information is needed from the input vector by the formula (1).

$$f_t = \sigma(W_f[h_{t-1}, x] + b_f), \quad (1)$$

where W_f - parameter matrix, b_f - parameter vector.

The vector i_t checks the relevance of the vector value by the formula (2)

$$i_t = \sigma(W_i[h_{t-1}, x] + b_i), \quad (2)$$

where W_i - parameter matrix, b_i - parameter vector.

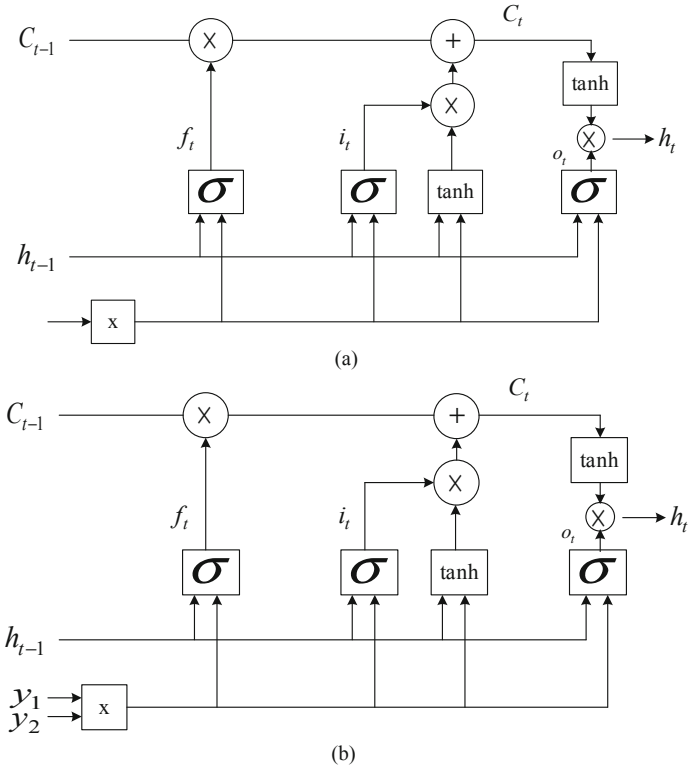


Fig. 4. The structure of the LSTM element: a) without signal preprocessing, b) with signal preprocessing [7].

The assignment of a new state vector is carried out according to the formula (3)

$$C_t = f_t \times C_{t-1} + i_t \times \tanh(W_c[h_{t-1}, x] + b_c), \tag{3}$$

where W_c - parameter matrix, b_c - parameter vector, \times - multiplication operator.

The candidate for the output vector o_t is determined by the expression (4)

$$o_t = \sigma(W_o[h_{t-1}, x] + b_o), \tag{4}$$

where W_o - parameter matrix, b_o - parameter vector.

The output vector of the LSTM element h_t is determined by the formula (5)

$$h_t = o_t \times \sigma(C_t). \tag{5}$$

Further classification of signals takes place according to three standard layers of neural networks presented in the simulation diagrams (Fig. 1 a-d) in the “LSTM network” block.

2.5 Feature Extraction

To extract features, the instantaneous frequency and spectral entropy calculated by formulas (6) and (7), respectively, were used:

$$f_i(t) = \frac{1}{2\pi} \frac{dl(t)}{dt}, \quad (6)$$

$$S = - \sum_{i=1}^N p_i \log n_i, \quad (7)$$

where $l(t)$ - ECG signal, S - amount of information, N - number of possible events, n_i - value of the i -th count of the ECG signal. A detailed description of this feature extraction method and justification for its use is given in [7].

3 Results

The simulation of the circuits Figs. 1, 2, 3, and 4 was carried out in the MatLab 2020b environment on the PhysioNet Computing in Cardiology Challenge 2017 database, taken from an open source [12]. The database includes 4 types of single-channel signals: signals with atrial fibrillation (A), noisy signals (~), signals with other pathologies and signals of work, a healthy heart (N). 9. To simulate the proposed method, we used part of the database [12] (1000 signals) (Fig. 5).

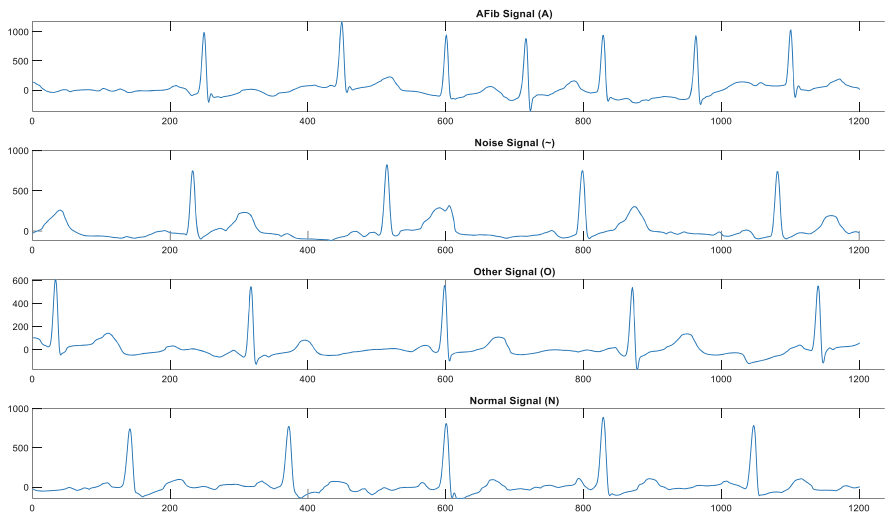


Fig. 5. Examples of ECG signals from the database [12]

90% of the signals from the database were used to train the network, and 10% were used for testing. The training was carried out on the base using four types of signals.

In models with a one-dimensional vector, testing and training were carried out on four types of signals, and in models with a two-dimensional input signal, training was carried out on the basis of two types of signals, and testing on all types of signals. The simulation results are presented in Table 1.

Table 1. Simulation results

Type of neural network	Preliminary feature extraction	Signals included in the database	Signals for Feature Extraction	Result
Unidirectional LSTM	No	A, O, N, ~	A, O, N, ~	40%
	Yes	A, O, N, ~	A, O, N, ~	58%
	Yes	A, O, N, ~	A, N	83%
	Yes	A, O, N, ~	O, ~	79%
Bidirectional LSTM	No	A, O, N, ~	A, O, N, ~	46%
	Yes	A, O, N, ~	A, O, N, ~	46%
	Yes	A, O, N, ~	A, N	31%
	Yes	A, O, N, ~	O, ~	50%
Bidirectional LSTM [7] with preliminary signal denoising	Yes	A, N	A, N	87.5%

4 Discussion

Using a unidirectional LSTM network to detect the type of ECG signal showed the following results: this type of neural network gives the maximum result for determining atrial fibrillation based on four types of signals with a preliminary feature extraction of 83%. The result of 45% in the classification of noisy signals and signals with other pathologies is due to inappropriate methods of preliminary feature extraction. The result of 31% for classification without preliminary feature extraction is understandable by the visual similarity of ECG signals.

The results of modeling the bidirectional LSTM network showed a significantly worse result, which is due to the lack of preliminary cleaning of the ECG signal from noise. Such a conclusion allows us to draw the results of modeling from [7], where before a similar simulation, ECG signals were cleaned from noise using a discrete wavelet transform.

5 Conclusion

The use of a unidirectional LSTM network for classifying ECG signals is possible without preliminary signal denoising. Using a bidirectional LSTM network to classify ECG signals without prior noise reduction is not practical.

The data obtained in this study allow us to continue the work on the automatic classification of ECG signals in the direction of searching for a universal method of noise suppression and preliminary extraction of features on ECG signals.

Acknowledgments. The author are grateful to the North Caucasus Federal University for supporting the competition of scientific groups and individual scientists of the North Caucasus Federal University.

References

1. World Health Organization. Trends In Noncommunicable Disease Mortality And Risk Factors, And Deaths From Injuries And Violence. World health statistics 2020: monitoring health for the SDGs, sustainable development goals, pp. 12–18 (2020)
2. Anant, K.S., Dowla, F.U., Rodrigue, G.H.: Detection of the electrocardiogram P-wave using wavelet analysis. *Int. Soc. Opt. Photonics* **1994**(2242), 744–750 (1994)
3. Yang, H., Bukkapatnam, S., Komanduri, R.: Nonlinear adaptive wavelet analysis of electrocardiogram signals. *Phys. Rev. E* **76**, 026214 (2007). <https://doi.org/10.1103/PhysRevE.76.026214>
4. Kumar, A., Komaragiri, R., Kumar, M.: Design of wavelet transform based electrocardiogram monitoring system. *ISA Trans.* **2018**(80), 381–398 (2018)
5. Fujita, H., Cimr, D.: Computer aided detection for fibrillations and flutters using deep convolutional neural network. *Inf. Sci.* **486**, 231–239 (2019). <https://doi.org/10.1016/j.ins.2019.02.065>
6. Lippi, G., Sanchis-Gomar, F., Cervellin, G.: Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge. *Int. J. Stroke* **2020**(16), 217–221 (2020)
7. Lyakhov, P., Kiladze, M., Lyakhova, U.: System for neural network determination of atrial fibrillation on ECG signals with wavelet-based preprocessing. *Appl. Sci.* 2021, **11**, 7213 (2021). <https://doi.org/10.3390/app11167213>
8. Hallhuber, M.J., Günther, R., Ciresa, M.: Technique of ECG recording. In: *ECG—An Introductory Course A Practical Introduction to Clinical Electrocardiography*, pp. 141–145. Springer, Heidelberg (1979). https://doi.org/10.1007/978-3-642-67280-4_15
9. Berbari, E.J., Lander, P.: The methods of recording and analysis of the signal averaged ECG. In: *Signal Averaged Electrocardiography*, pp. 49–68. Springer, Dordrecht (1998). https://doi.org/10.1007/978-94-011-0894-2_4
10. British Cardiovascular Society, Guidelines for recording a standard 12-lead electrocardiogram (2013)
11. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Comput*, pp. 1735–1780 (1997)
12. PhysioNet Computing in Cardiology Challenge 2017 (CinC Challenge). <http://physionet.org/challenge/2017/>. Accessed on 23 August 2022



Trust Monitoring in a Cyber-Physical System for Security Analysis Based on Distributed Computing

Elena Basan¹  , Maria Lapina² , Alexander Lesnikov¹, Anatoly Basyuk¹, and Anton Mogilny¹

¹ Southern Federal University, Taganrog, Russian Federation
ebasan@sfnfedu.ru

² North Caucasus Federal University, Stavropol, Russian Federation

Abstract. Cyber-physical systems are widely used. Nevertheless, security issues are quite acute for them. First of all, because the system must work constantly without downtime and failures. The Cyber-Physical System (CPS) must quickly transfer the parameters to the monitoring system, but if the system is not flexible enough, fast and optimal, then collisions and additional loads on the CPS may occur. This study proposes a system for monitoring and detecting anomalies for CPS based on the principles of trust, which allows you to verify the correctness of the system and detect possible anomalies. In our study, we focus on traffic analysis and analysis of the CPU operation, since these parameters are the most critical in the operation of the CPS itself. The technique is based on computationally simple algorithms and allows to analyze the basic parameters that are typical for most CPS. These factors make it highly scalable and applicable to various types of CPS, despite the fragmentation and a large number of architectures. A distributed application architecture was developed for monitoring and analyzing trust in the CPS. The calculation results show the possibility of detecting the consequences of the influences of denial-of-service attacks or CPS. In this case, three basic parameters are sufficient for detection. Thus, one of the features of the system is reflexivity in detecting anomalies, that is, we force devices to independently analyze their behavior and make a decision about the presence of anomalies.

Keywords: Trust · Reflection · Anomaly Detection · Attacks · Denial of Service · Monitoring

1 Introduction

A Cyber-Physical System (CPS) provides a tight link between the cyber and physical domains by embedding cyber processes (e.g., communications, computing, or control) into physical devices. Intrusion detection systems are designed to detect anomalous behavior or unexpected activities in networks by automatically analyzing their behavior based on a given hypothesis and/or policies that are governed by the network's security

rules [1]. The system monitors system configuration, data files, and/or network transmissions to check for an attack. Thus, this system is an important first step in preventing any covert/overt actions aimed at exploiting security vulnerabilities to crash or hijack the system. Such misuse can be defined as any undesirable action that could cause any harm in terms of performance or security of the entire group. Attacks exploit vulnerabilities in CPS that can result from network misconfiguration, implementation errors, design and/or protocol failures [2].

The issues of CPS safety monitoring are discussed in many works of the authors. To ensure the uninterrupted operation of cyber-physical systems, decision-making systems are used based on information received by the information security management system from the monitoring system. In this regard, we single out the monitoring system as an important step in the operation of the CPS information security management system. Modern information security management systems (ISMS) are devoted to a large number of research works, from architectural solutions [3–8] to the search for methods for solving security problems [9, 10]. The authors of the paper [11] highlight the problem of choosing the most appropriate set of methods for solving security problems for a particular CPS configuration. To solve this problem, the authors present a method for managing an adaptive information security monitoring system. The method consists in solving the problem of multiobjective discrete optimization under Pareto optimality conditions when the available data, methods, or external requirements change. An experimental study was carried out on the example of intrusion detection in a smart home system. As a result, the information security monitoring system acquires the property of adaptability to changing tasks and available data. As the number and complexity of cyberattacks have increased, machine learning (ML) has been actively used to detect cyberattacks and malicious activity. Cyber-Physical Systems (CPS) combined calculations with physical procedures. An embedded computer and network monitor and control physical processes, usually with feedback. Normally physical procedures affect computations, and ML approaches have been vulnerable to data poisoning attacks. Improving network security and achieving the reliability of network schemes defined by ML have been critical issues in the growth of attacks and the size of the CPS. In the paper [12] authors develop a new stochastic fractal search algorithm with a deep learning based intrusion detection system (SFSA-DLIDS) for the CPS cloud environment. The presented SFSA-DLIDS approach primarily implements a minimum-maximum data normalization approach to transform input data into a compatible format. Monitoring systems are necessary for the analysis and control of the CPS behavior. CPS are associated with real-time constraints and physical phenomena that are usually not taken into account in typical information systems. In the paper presented by the authors of paper [13], the CPS-MT system is shown, aimed at creating a universal tool for monitoring CPS in real time from a security point of view.

Thus, the problem of security monitoring in CPS is quite relevant. The authors propose a large number of different solutions. At the same time, there is no specifics on which parameters are analyzed and whether they are universal. In addition, the concept of collecting information about parameters has not been fully developed. The CPS must quickly transfer the parameters to the monitoring system, but if the system is not flexible enough, fast and optimal, then collisions and additional loads on the CPS may occur. This study proposes a system for monitoring and detecting anomalies for CPS based

on the principles of trust, which allows you to verify the correctness of the system and detect possible anomalies. In our study, we focus on traffic analysis and analysis of the CPU operation, since these parameters are the most critical in the operation of the CPS itself.

2 Materials and Methods

2.1 Basic Concept of a Cyber-Physical System based on State Analysis

These layers are important for the features of the interaction of the system components and understanding which components interact directly and which through intermediaries. This understanding is important when modeling attack vectors on a system. As a rule, an attacker acts through communication channels if he does not have direct access to the system. Physical layer components may not have network interfaces but be connected to other components through Low-Level Management components.

$$CPS = \{HLC\} \cup \{NLC\} \cup \{LLC\} \cap \{PH\}, \quad (1)$$

where $\{HLC\}$ - set high-level management components, $\{NLC\}$ - set of network layer components, $\{LLC\}$ - set of Low-Level Management components, $\{PH\}$ - set of Physical layer components.

Moreover, the components of the set of high-level, network and low-level management do not intersect, but the components of the low-level management and the physical layer can intersect, since their properties intersect. This will be proven below.

Physical layer components include sets of sensors $S = \{s_0, \dots, s_n\}$, detectors $D = \{d_0, \dots, d_j\}$, actuators $A = \{a_0, \dots, a_m\}$, power supplies $PS = \{ps_0, \dots, ps_i\}$ and other additional devices that ensure the functioning of the CPS. In this case, the number of elements of the set may differ. Thus, the set $\{PH\}$ has 4 subsets, and none of these subsets is a subset of the other. The characteristics of the system can be:

- indications of various sensors (cyber-physical parameters),
- state of cyber-physical objects.

$\{PH\}$ is characterized by a set of cyber-physical parameters $CP = \{cp_0, \dots, cp_i\}$ that can be obtained from sensors. This set depends on the sets $\{S\}$, $\{D\}$, $\{A\}$, $\{PS\}$. In particular, the more sensors, transmitters, etc. are installed, the more data can be obtained from them, and the more cyber-physical parameters can be processed.

$$|CP_i| = |CPS_i| + |CPD_i| + |CPA_i| + |CPPS_i| \quad (2)$$

where $|CP_i|$ is the number of all elements in the finite set of the given CPS_i , CPS_i is the number of all cyber-physical parameters received from the sensor system, CPD_i is the number of all cyber-physical parameters received from sensors, CPA_i is the number of all cyber-physical parameters received from actuators, $CPPS_i$ is from the accumulator and other peripheral devices.

In the case of a cyber-physical system, a microcontroller can be used to control sensors, sensors, and actuators. Accordingly, the data from the sensors comes to a higher

level from the microcontroller. A microcontroller belongs to a set of microcontrollers $MC = \{mc_0, \dots, mc_i\}$ that are a subset of the LLC (low-level control), so the following is true:

$$mc_{fc,i} \in \{LLC\} \quad (3)$$

Since the microcontroller essentially includes a set of sensors that are connected to it and from which it receives information that it transmits further, and it can also transmit information to actuators, it can be said that the sets of low-level control and physical level objects can intersect and an element such as a microcontroller belongs to both sets:

$$LLC \cap PH = \{mc_{fc,i} | mc_{fc,i} \in LLC, mc_{fc,i} \in PH\} \quad (4)$$

Accordingly, we will assume that the controller is also characterized by a set of cyber-physical parameters that it gives to a higher level. Moreover, it can receive cyber-physical parameters from several elements of the PH set at once. Thus, a vector of parameters is formed, which may include a set of cyber-physical parameters of the controller. Thus, the evaluation of trust in a cyber-physical system is reduced to an assessment of trust in the quality of changes in cyber-physical parameters.

2.2 Trust-Based Verification Method of CPS Operation

To determine the degree of confidence in the current state of the CPS, we define a set of states. The trusted state is such a state when the change in the CPS parameters does not exceed the allowable values and corresponds to the expected values, which are trustworthy and allow for the smooth operation of the CPS.

An untrusted state is a state when the change in the parameters of the CPS exceeds the allowable confidence intervals, or does not reach the minimum values, which leads to failures in the operation of the CPS. In this study, we have focused only on active malicious activities and attacks that can damage the integrity and availability of the system. Let's define a set of metrics that are used to assess the state of the CPS (Table 1).

Table 1. A set of metrics for monitoring the state of the system

Metric	Formula	Note
Reliability of the functions performed in the current state		
1. Confidence limits		
CP_i^{\min} lower limit of confidence interval	$CP_{i,st}^{\min} = CP_{i,st} - \sigma_i^{st}$ $CP_{i,s_{n-1}}^{\min} = CP_{S_{n-1}} - \sigma_i$ $CP_{i,current}^{\min} = \overline{CP}_i - t \cdot \sigma_{omi}^{current} / \sqrt{n}$	$CP_{i,st}^{\min}$ - minimum value in the presence of targets, $CP_{S_{n-1}}$ - the value of the cyber-physical parameter from the previous state, $t * \sigma / \sqrt{n}$ - estimation accuracy, t - argument of the Laplace function, where $(t) = \frac{\alpha}{2}$, α - given reliability, σ_i - allowable deviation, $CP_{i,current}^{\min}$ - the minimum value based on the collected parameters for the previous time intervals

(continued)

Table 1. (continued)

Metric	Formula	Note
$CP_{i,ST}^{max}$ - upper limit of the confidence interval	$CP_{i,ST}^{max} = CP_{i,ST} - \sigma_i^{st},$ $CP_{i,SN-1}^{max} = CP_{SN-1} + \sigma_i,$ $CP_{i,current}^{max} = \overline{CP}_i + t \cdot \sigma_{omi}^{current} / \sqrt{n}$	$CP_{i,ST}^{max}$ - the maximum value in the presence of target, $CP_{i,current}^{max}$ - the maximum value based on collected data for previous time intervals
2. Estimation of the probability of going beyond the confidence interval		
Cumulative function for the Poisson distribution f_{pois}	$f_{pois}(CP_i \overline{CP}_i) = \sum_{j=1}^{CP_i^n} \frac{\overline{CP}_i^{CP_i} CP_i e^{-\overline{CP}_i}}{CP_i!},$ $f_{pois,min}(CP_i CP_{i,min}) =$ $\sum_{j=1}^{CP_i^n} \frac{CP_{i,min}^{CP_i} e^{-CP_{i,min}}}{CP_i!}$ $f_{pois,max}(CP_i CP_{i,max}) =$ $\sum_{j=1}^{CP_i^n} \frac{CP_{i,max}^{CP_i} e^{-CP_{i,max}}}{CP_i!}$	The cumulative Poisson probability is related to the probability that the random Poisson frequency is greater than a given limit and less than a given upper limit $CP_{i,max}$ - upper redistribution of the value of the cyber-physical parameter, $CP_{i,min}$ - the lower limit of the cyber-physical parameter
The average value of the cyber-physical parameter in the range of the sliding window	$\overline{CP}_i = \frac{1}{n} \sum_{j=1}^n P_i \Delta w_{ij}$	n is the sample size, P_i is the values of the sample parameters, Δw is the sliding window for a given time interval of values, equal to n

Metric 1. Boundary of the confidence interval.

Boundary of the confidence interval in the presence of targets. Target indicators are those values of cyber-physical parameters that the cyber-physical system must achieve. In the case when the CPS operates in an autonomous mode, such indicators can be taken from the technical certificate, as well as during an expert assessment of normal indicators.

Boundary based on knowledge of CPS. The value of the parameter can be used to define the limit of the indicator, which is obtained in the idle state, when the system is not performing active actions, but is already enabled. Values for determining the boundary of the confidence interval can be taken from the previous state of the system operation.

Confidence interval bound based on previous values from the sample. If there is no input information about the normal performance of the system, as well as information about the reference behavior of the system, then it is possible to calculate the boundaries of the interval dynamically. To do this, it is necessary to build a confidence interval based on the previous behavior of the system. In this case, the confidence limits of the interval will be regulated by the standard deviation. In any mode of functioning of a cyber-physical system, which is included in the set of normal states of the system, the change in cyber-physical parameters should occur smoothly. Even if the growth of the function is observed, in order for it not to have a critical impact on the system, it must be smooth.

Metric 2. Reliability of performed functions in the current state.

This metric allows you to determine how much the current parameters of the cyber-physical system correspond to the given boundaries. That is, it is necessary to evaluate whether the running process goes beyond the allowable interval. Such an assessment is possible only for those parameters for which the normal or preset values are reliably known. The reliability of the functions performed is understood as the degree of confidence (or the probability that) in the i -th function or action of the process, which is described by a change in cyber-physical parameters such that their current change does not go beyond the confidence interval. Thus, this metric allows you to control that the process being executed does not go beyond the confidence limits.

To determine whether the system can be trusted, it is necessary to determine whether the current technological process of the CPS goes beyond the permissible range. To do this, the boundaries of the confidence interval are calculated and the degree of exceeding or reaching them is estimated. Thus, if values close to 1 are observed, there is a correspondingly high probability that the system does not correspond to the given boundaries.

3 Results

3.1 Software Module for Monitoring and Analyzing Trust in CPS

Figure 2 shows the general architecture for data collection, monitoring and determining the state of the CPS. One of the main modules of the project. Engaged in obtaining and normalizing all monitored cyber-physical parameters of the device according to certain algorithms. Since the module and its functionality is very extensive, it is divided into several subroutines: Subroutine for logging errors; Subroutine for logging network sniffer errors; A subroutine for logging server errors; Subroutine for logging errors in monitoring the cyber-physical parameters of the device; Subroutine for logging errors during normalization and saving parameters; Subroutine for reading the configuration file. Designed to read the configuration file and initialize critical parameters for the module to work; TCP server initialization routine; Network sniffer initialization routine. Designed to track and monitor the state of the network; Subprogram for initialization of the module for monitoring the cyber-physical parameters of the device. It is designed to obtain the cyber-physical parameters of the device, for example: CPU load, CPU temperature, RAM load; Subroutine for initializing the normalization module and saving the received parameters. It is intended for constructing series of parameter values according to the given settings, normalizing the obtained series and providing normalized information for analysis on the corresponding module.

There are a collection of basic parameters that may be relevant for most CPSs: CPU load, CPU temperature, network traffic load (number of packets for each protocol), RAM load.

Transferring current device settings to monitoring website. The data for connection is obtained by the parameter collection module from the configuration file. Further, when accessing the API of the monitoring site, the current parameters of the device are transferred. The analysis module is designed to analyze normalized series in order to obtain confidence values. Recording the received rows of information in the database on the

device (low-level control). The monitoring module implemented using web technologies is located at the high management and receives data from the low management layer. Carrying out calculations at a low control level, on the one hand, loads the microprocessor or microcontroller, but it allows the low-level device to independently detect an anomaly and make trust decisions in a distributed network. Between the analysis module and the monitoring module, data is transmitted over the network, which is a vulnerability. If an attack is carried out on a communication channel, then the data will be lost and the system will not respond in time. In the proposed architecture, the monitoring module is only informative. All decisions are made by a distributed system consisting of small computers.

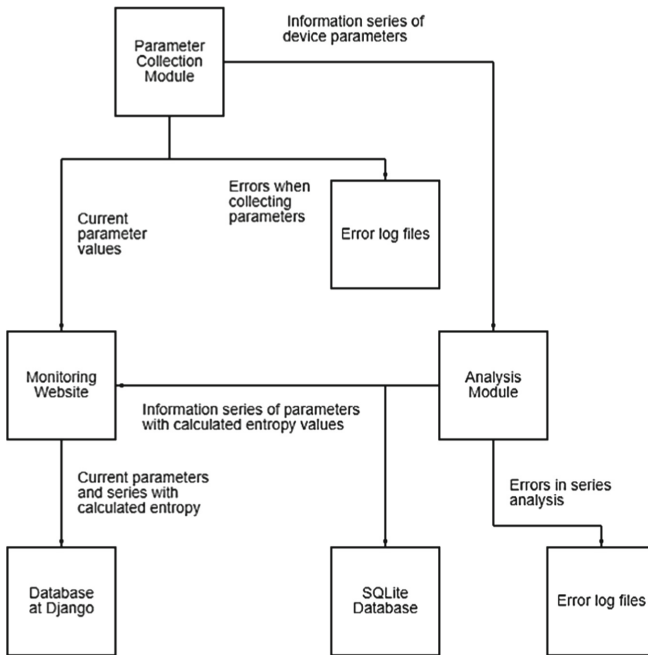


Fig. 2. Overall System Architecture

3.2 Results of Calculating the Probability of Values Going Beyond the Confidence Interval

An experimental study was carried out for the CPS model, which was built following the example of a full-scale model of an automated plant [14]. At the same time, 4 types of malicious impact were carried out: low-intensity TCP flood attack, medium-intensity TCP flood attack, high-intensity UDP flood attack and high-intensity ICMP flood attack [15–17]. The intensity of the attack was regulated by the speed and number of packets sent. Data collection and analysis modules are deployed on three low-level control devices: the device responsible for the human machine interface (HMI), the

device with the Supervisory Control And Data Acquisition (SCADA) system installed, the device responsible for the control of the Programmable Logic Controller (PLC). The parameters of each of the devices change a little differently due to the fact that they perform different functions. Consider the results of calculations for the SCADA device, which are shown in Fig. 3. The SCADA system passively collects data. In this case, the operator can connect to the module through the web interface for monitoring. As can be seen from the figures, each attack affects each parameter of the SCADA system. With the exception of the RAM load indicator, this parameter did not change during the experimental study, so the graph for it is not shown. Next, consider the impact of attacks on the PLC system. The calculation results are shown in Fig. 4. Figures 3 and 4 show that even for a normal situation, there are single peaks for processor load and transmitted traffic. This is due to the fact that since the system controls automated production, according to the algorithm, control commands are transmitted automatically at certain periods.

Thus, single excesses can occur, and such a situation will be considered normal. The main condition is that the sequence of peaks does not exceed three. In the case of an attack, we see entire sequences of exceeding values.

4 Discussion

For the PLC, such peaks are observed more often, because it directly controls microcontrollers. In addition, the peaks are systemic and regular, which indicates the monotonicity of the process. Tracking the frequency of peaks can later become an indication of normal operation and one of the signs of normal behavior.

In addition, you can link CPU peaks to traffic load. It is with this that the resulting excess loads are associated. During the TCP flood attack, all three parameters were affected. This is because the protocol itself is quite resource-intensive and the process of maintaining the set values most affects the device. For the PLC, during the implementation of the TCP flood attack, unacceptable consequences were observed when the automated production did not work correctly, and the work could stop. At the same time, the node stopped transmitting data, it could only write them to the internal database. At the same time, with each attack, an excess of CPU indicators and a level of traffic

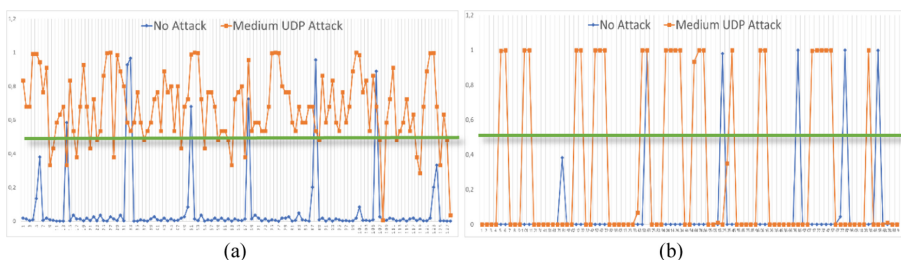


Fig. 4. The result of calculating the probability of values going beyond the confidence interval (a) for the CPU utilization level (b) for the network traffic utilization level, where the parameter excess level is shown vertically, and the time interval is horizontal. The green horizontal line indicates the limit value.

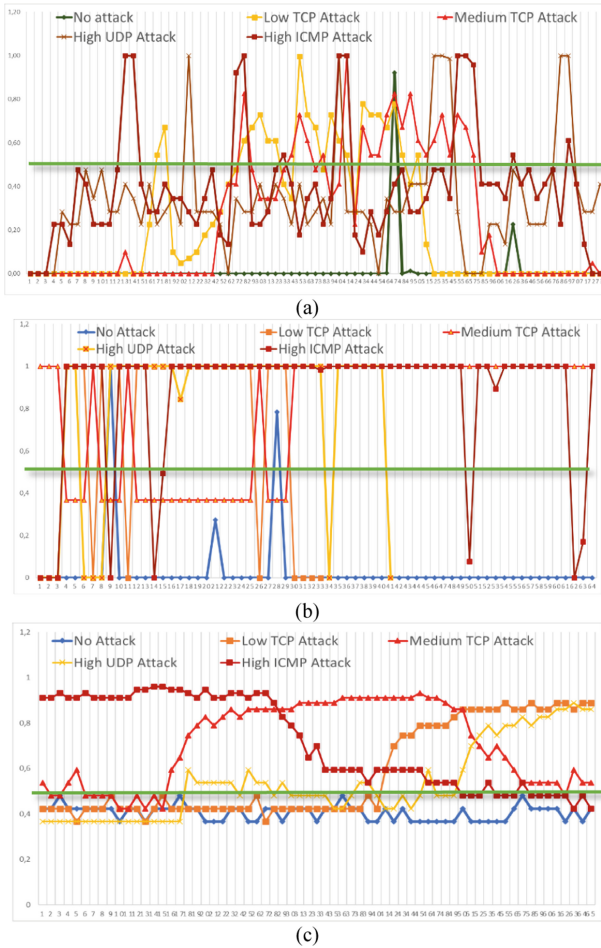


Fig. 3. The result of calculating the probability of values going beyond the confidence interval (a) for the CPU utilization level (b) for the network traffic utilization level (c) for the CPU temperature, where the parameter excess level is shown vertically, and the time interval is horizontal. The green horizontal line indicates the limit value.

congestion were observed. At the same time, the monitoring system could not receive correct data, so the graphs are not shown. However, by maintaining an internal database, incident detection becomes possible.

At the same time, during the UDP attack, the PLC flood reacted only to the CPU load and network traffic, other parameters did not change significantly. It should also be noted that the graphs show the total number of traffic as a summary parameter. When analyzing packets according to the protocols for the UDP flood attack, you can immediately see a significant excess of UDP packets compared to the normal state of operation. Thus, the detection becomes more accurate. In addition, if you track received and sent packets

separately, then the overshoot rate will increase for received packets compared to sent ones. These provisions will be explored in the future.

The ICMP flood attack of high intensity and the TCP attack of medium intensity had the greatest impact on the parameters of the SCADA system. This is due to the fact that SCADA does not exchange UDP traffic and does not have open ports that can be influenced. In this case, the PLC exchanges UDP traffic with microcontrollers and therefore the UDP flood attack has a significant impact. In one case or another, each of these attacks is detected by the analysis system successfully.

5 Conclusion

Thus, in this paper, the concept of CPS was considered from the point of view of functioning and the possibility of analyzing states. The relationship between the processes and levels of the CPS with the change in cyber-physical parameters is proved. Based on this evidence, a method for analyzing changes in cyber-physical parameters has been developed as a basis for detecting malfunctions in the system. The technique is based on computationally simple algorithms and allows to analyze the basic parameters that are typical for most CPS. These factors make it highly scalable and applicable to various types of CPS, despite the fragmentation and a large number of architectures. A distributed application architecture was developed for monitoring and analyzing trust in the CPS. The architecture, due to the distribution of calculations, allows continuous analysis of trust in the system locally on the control devices of low-level control. The calculation results show the possibility of detecting the consequences of the influences of denial-of-service attacks or CPS. In this case, three basic parameters are sufficient for detection. Thus, one of the features of the system is reflexivity in detecting anomalies, that is, we force devices to independently analyze their behavior and make a decision about the presence of anomalies.

Acknowledgments. The research was supported by the Council for Grants of the President of the Russian Federation at the expense of the scholarship of the President of the Russian Federation for young scientists and graduate students (Competition SP-2022) No. SP-858.2022.5 on the topic “Technology for ensuring cybersecurity of automated systems from active information attacks based on the principle of reflection”.

References

1. Choi, S., Woo, J., Kim, J., Lee, J.Y.: Digital twin-based integrated monitoring system: korean application cases. *Sensors* **22**, 5450 (2022). <https://doi.org/10.3390/s22145450>
2. Yang, B., Xin, L., Long, Z.: An improved residual-based detection method for stealthy anomalies on mobile robots. *Machines* **10**, 446 (2022). <https://doi.org/10.3390/machines10060446>
3. Kotenko, I.V.: Primenenie tekhnologii upravleniya informaciej i sobytijami bezopasnosti dlya zashchity informacii v kriticheski vazhnyh infrastrukturah. *Trudy SPIIRAN Vyp 1*, 2–7 (2012)

4. Lavrova, D.S., Zaitseva, E.A., Zegzhda, D.P.: Approach to presenting network infrastructure of cyberphysical systems to minimize the cyberattack neutralization time. *Autom. Control. Comput. Sci.* **53**(5), 387–392 (2019). <https://doi.org/10.3103/S0146411619050067>
5. Stevens, M.: Security Information and Event Management (SIEM). In *Proceedings of the Nebraska CERT Conference*, Omaha, NE, USA, 9–11 August 2005. <http://www.certconf.org/presentations/2005/files/WC4.pdf>
6. Knapp, E.D., Langill, J.T.: Chapter 12–Security Monitoring of Industrial Control Systems. In: Eric, D., Knapp, J.T. (eds.) *Industrial Network Security*, 2nd ed., pp. 351–386. Syngress, New York (2015)
7. Lavrova, D.S.: Podhod k razrabotke SIEM-sistemy dlya Interneta veshchej. *Probl. Inf. Bezopasnosti. Komp'yuternye Sist.* **2**, 51–59 (2016)
8. Siddiqui, S., Khan, M.S., Ferens, K., Kinsner, W.: Fractal based cognitive neural network to detect obfuscated and indistinguishable internet threats. In: *Proceedings of the 2017 IEEE 16th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC)*, Oxford, UK, 26–28 July 2017; pp. 297–308 (2017)
9. Wang, C., Wang, D., Xu, G., He, D.: Efficient privacy-preserving user authentication scheme with forward secrecy for industry 4.0. *Sci. China Inf. Sci.* **65**(1), 1–15 (2021). <https://doi.org/10.1007/s11432-020-2975-6>
10. Jiang, Y., Yin, S., Kaynak, O.: Data-driven monitoring and safety control of industrial cyber-physical systems: basics and beyond. *IEEE Access* **6**, 47374–47384 (2018)
11. Poltavtseva, M., Shelupanov, A., Bragin, D., Zegzhda, D., Alexandrova, E.: Key concepts of systemological approach to CPS adaptive information security monitoring. *Symmetry* **13**, 2425 (2021). <https://doi.org/10.3390/sym13122425>
12. Duhayyim, M.A., et al.: Evolutionary-based deep stacked autoencoder for intrusion detection in a cloud-based cyber-physical system. *Appl. Sci.* **12**, 6875 (2022). <https://doi.org/10.3390/app12146875/>
13. Thakur, S., Chakraborty, A., De, R., Kumar, N., Sarkar, R.: Intrusion detection in cyber-physical systems using a generic and domain specific deep autoencoder model. *Comput. Electr. Eng.* **91**, 107044 (2021)
14. Sauer, F., Niedermaier, M., Kiebling, S., et al.: LICSTER – a low-cost ICS security testbed for education and research. In: *6th International Symposium for ICS & SCADA Cyber Security Research* (2019). <https://doi.org/10.14236/ewic/icscsr19.1>
15. Gamec, J., Basan, E., Basan, A., Nekrasov, A., Fidge, C., Sushkin, N.: An adaptive protection system for sensor networks based on analysis of neighboring nodes. *Sensors* **21**, 6116 (2021). <https://doi.org/10.3390/s21186116>
16. Basan, E., Basan, A., Nekrasov, A.: Method for detecting abnormal activity in a group of mobile robots. *Sensors* **19**, 4007 (2019). <https://doi.org/10.3390/s19184007/>
17. Basan, E., Basan, A., Makarevich, O.: Detection of anomalies in the robotic system based on the calculation of kullback-leibler divergence. In: *2019 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, 2019, pp. 337–340 (2019). <https://doi.org/10.1109/CyberC.2019.00064>



Guaranteed Safe States in Speed Space for Ships Collision Avoidance Problem

A. S. Zhuk^(✉)

Admiral Ushakov Maritime State University, Novorossiysk, Russian Federation
alszhuk@yandex.ru

Abstract. The article has considered approach to ships collision avoidance problem for infinite time horizon in unpredictable navigation environment on the basis of guaranteed safe states in speed space. The possible collisions areas are determined on the basis of the reachable sets of target vessels in the speed space. By avoiding getting into the possible collisions area of each target vessel, the controlled vessel prevents any possible collision, while not having explicit information about the expected trajectories of the target vessels. The safety of navigation is achieved by a combination of reachable sets as functions of time for target vessels, taking into account their dynamic capabilities and the representation of sets in speed space. The control option, which is outside the reachable set of the target vessel in the speed space, is guaranteed to prevent a collision even under conditions of unpredictable motion of the target vessel. On the basis of the target vessels reachable sets in speed space the guaranteed safe states of controlled vessel are defined. Recommendations for practical use have been given. Performed researches contribute to the improvement of the ship's handling methods.

Keywords: Reachable set · Collision avoidance · Potential collision area · Speed space · Ship handling

1 Introduction

One of the most important tasks of ship control is undoubtedly collision avoidance. The features of different cases of approaching ships imply different assumptions, and hence different collision avoidance algorithms. Modern automatic radar plotting (ARPA) systems make extensive use of target motion calculations to select a collision avoidance maneuver based on the assumption that the target ship's motion parameters are known and predictable over time. However, in real maneuvering conditions, the predicted movement of the target ship can usually be considered reliable only for a short period of time.

When the functions of changing the coordinates of the target vessel are unknown, reachable areas can be used to determine areas of possible collision.

Here we will call the reachability area the totality of all possible positions of the target ship at a given time. By preventing contact entirely with the reachable area of the target vessel, the controlled vessel prevents any possible collision with the target vessel.

The area of possible collision will be called the area of the physical water space in which a collision with a given target vessel is possible. In the future, the areas of possible collision, contact with which the controlled vessel must avoid, will be considered in the speed space, and not in the physical water space [20, 21].

2 Formulation of the Problem

The task of ship motion control is to determine the conditions that guarantee collision avoidance even without accurate data on the future movement of target vessels as a function of time. This is achieved by constructing areas of possible collisions based on the reachability sets of target ships in the speed space and using the approach based on obstacles in the speed space to construct, in a form convenient for practical application, the sets of speed vectors of the controlled vessel, for which the movement of the controlled vessel occurs outside the areas of possible collisions at any time. By avoiding getting into the area of possible collisions of each target ship, the controlled ship prevents any possible collision, while not having explicit information about the expected trajectories of the target vessels.

The goal of control is to safely maneuver the vessel in a navigational environment with many potentially dangerous target vessels. Each target vessel has a constant linear speed and can develop an angular speed up to a certain maximum value. The equations of dynamics (1) of the target ships are written in the form of the Dubins model [1–7].

$$\dot{\vec{x}} = V[\sin K \cos K]^T; \quad |\dot{K}| \leq \omega_{max}; \quad \dot{V} = 0, \quad (1)$$

where \vec{x} – vessel coordinates vector in a rectangular coordinate system related to the meridian and the parallel; V – vessel's linear speed; K – vessel's heading in semi-circle system; ω_{max} – vessels's maximum angular speed.

In some cases, for example, when planning a route on a map, it is convenient to use a record of the form

$$R_{min} = \frac{V}{\omega_{max}}, \quad (2)$$

where R_{min} – minimum turning radius.

A controlled vessel can be described by various models of dynamics. We consider all ships to be point objects.

A navigational safety zone is a circle of a given radius R_{NSZ} . In this case, the dangerous approach situation occurs when the distance between vessels is less than the specified minimum safe distance equal to R_{NSZ} . The task of control includes the prevention of close quarter situations by measuring only the coordinates and headings of the target vessels, while not having any other additional information. In this case, the controlled vessel must comply with the preliminary passage plan as far as possible, but without the risk of collisions.

3 Method for Determining the Sets of Possible Ship Collisions in the Speed Space Based on Reachable Sets

To calculate the sets of possible collisions in the speed space, in addition to the parameters V and ω_{max} the coordinates and headings of each target vessel are required, as well as the minimum allowable distance of approach to the target ships - the radius of the navigation safety zone R_{NSZ} .

An important advantage of the proposed approach is the possibility of ensuring the safety of navigation on an unlimited planning horizon and in conditions of unpredictable movement of target vessels.

Knowing how the target ships will move, it is possible to calculate a safe maneuver on an unlimited planning horizon, which is implemented in modern ARPA. However, in real navigation conditions, it is not known for certain how all target ships will move. In these cases, ARPA does not guarantee the safety of navigation on an unlimited planning horizon against target vessels moving in an unpredictable manner. The idea of using reachability sets in the speed space makes it possible to determine the conditions for avoiding collisions, which play a major role in the algorithms for programming the motion of a controlled vessel [6, 8–17].

Based on the values of V and ω_{max} of the target vessel, the areas of possible collisions $APC_i(t)$ are determined as a function of time. The areas of possible collisions in the speed space $CSS_i(t)$ at each moment of time t are determined through the areas of possible collisions in the plane of the physical water space $APC_i(t)$ in the form

$$CSS_i(t) = \frac{APC_i(t)}{t} \quad \forall t > 0. \quad (3)$$

The controlled vessel, located at the initial moment of time at the origin of coordinates, whose speed vector is inside the $CSS_i(t)$, will be in the area of possible collisions $APC_i(t)$; while the controlled vessel, whose speed vector is outside the $CSS_i(t)$, will pass clear from the $APC_i(t)$.

Boundary $APC_i(t)$ at a given time t contains up to five possible parts. Four of them are set in parametric form depending on the heading K , which is set in a semi-circular system.

Let us determine the boundaries of the area of possible collisions of the target vessel in the coordinate system associated with the target vessel. At the initial moment of time, the heading of the target vessel is $K_{T0} = 0^\circ$, in the case of a non-zero initial heading of the target vessel, we use the off-bow angle of the target vessel instead of the heading.

The equations of the far boundary of the $APC_i(t)$ taking into account the R_{NSZ} based on [1–4, 6, 18, 19] have the form:

$$\begin{aligned} &\text{If } \max(-\omega_{max}t, -\pi) \leq K \leq 0, \\ &\quad \text{then } X_1(K, t) = -R_{min}(1 - \cos K) + (Vt + KR_{min} + R_{NSZ}) \sin K; \quad (4) \\ &\quad Y_1(K, t) = -R_{min} \sin K + (Vt + KR_{min} + R_{NSZ}) \cos K. \\ &\text{If } 0 \leq K \leq \min(\omega_{max}t, \pi), \\ &\quad \text{then } X_2(K, t) = R_{min}(1 - \cos K) + (Vt - KR_{min} + R_{NSZ}) \sin K; \quad (5) \\ &\quad Y_2(K, t) = R_{min} \sin K + (Vt - KR_{min} + R_{NSZ}) \cos K, \end{aligned}$$

where X_1, Y_1, X_2, Y_2 – coordinates of the port and starboard sections of the far boundary of the area of possible collisions, respectively.

Two sections of the $APC_i(t)$ boundary are arcs of circles of radius R_{NSZ} , which interfit with the port (4) and starboard (5) parts of the far APC boundary. If $|\omega t| \geq \pi$,

then the arcs of the circles degenerate and do not form part of the $APC_i(t)$. If $|\omega t| < \pi$, then the arc equations in parametric form have the form.

$$\begin{aligned}
 &\text{If } -\pi \leq K \leq -\omega_{max}t, \text{ then} \\
 &X_3(k, t) = -R_{min}(1 - \cos(\omega_{max}t)) + R_{NSZ} \sin K; \tag{6} \\
 &Y_3(K, t) = R_{min} \sin(\omega_{max}t) + R_{NSZ} \cos K; \\
 &\text{If } \omega_{max}t \leq K \leq \pi, \text{ then} \\
 &X_4(K, t) = R_{min}(1 - \cos(\omega_{max}t)) + R_{NSZ} \sin K \\
 &Y_4(K, t) = R_{min} \sin(\omega_{max}t) + R_{NSZ} \cos K, \tag{7}
 \end{aligned}$$

where X_3, Y_3, X_4, Y_4 – coordinates of the port and starboard arcs of circles that form sections of the APC boundaries, respectively.

Note that for $t > \pi / \omega$, the arcs of circles (6), (7) do not form the boundaries of the APC.

Sections (4)-(7) are mated and form the far boundary of the region of possible collisions $APC(K,t)$, where $K \in [-\pi, \pi]$.

Finally, the horizontal section connecting the lower points of the sections $[X_1 Y_1]$ and $[X_2 Y_2]$, or $[X_3 Y_3]$ and $[X_4 Y_4]$, forms the near boundary of the region of possible collisions $APC(t)$.

Let us determine the coordinates of the target vessel in a fixed coordinate system oriented north along the meridian and related to the initial position of the controlled vessel. The target ship at the initial time is at a point with route coordinates x_0, y_0 with a heading of K_{T0} , it is necessary to perform a coordinate transformation according to the known formulas for moving and rotating the coordinate axes:

$$x_i = x_0 + X_i \cos K_{T0} + Y_i \sin K_{T0}; \quad y_i = y_0 - X_i \sin K_{T0} + Y_i \cos K_{T0}; \tag{8}$$

where x_i, y_i – coordinates of the target vessel in a fixed coordinate system oriented north along the meridian and related to the initial position of the controlled vessel.

4 Simulation Results

In Fig. 1, the areas of possible collisions $APC_i(t)$ of the target vessel, which at the initial time $t_0 = 0$ is located in the coordinates $x_0 = 5.0$ nm, $y_0 = 5.0$ nm with an initial heading $K_{T0} = 135^\circ$ and moves with a constant speed of 10.0 knots. Minimum turning radius of the target vessel $R_{min} = 1.0$ nm. The radius of the navigation safety zone $R_{NSZ} = 1.0$ nm.

The boundary of the region of possible collisions in the speed space $CSS_i(t)$ is determined by the equations:

$$\dot{x}_i(K, t) = \frac{x_i(K, t)}{t}; \quad \dot{y}_i(K, t) = \frac{y_i(K, t)}{t}; \tag{9}$$

where \dot{x}_i, \dot{y}_i – coordinates of the boundary of the region of possible collisions in the speed space $CSS_i(t)$, respectively.

Any speed vector outside the $CSS_i(t)$ forms the movement of a controlled vessel with a constant heading and speed, safe at time t . To ensure safety on a given time interval, it is necessary to find the union of all instantaneous $CSS_i(t)$ on this interval, which we will call the set of possible collisions in the speed space (SCS). Thus, the SCS includes all vectors of the controlled vessel for which there is a probability of collision within the planning horizon.

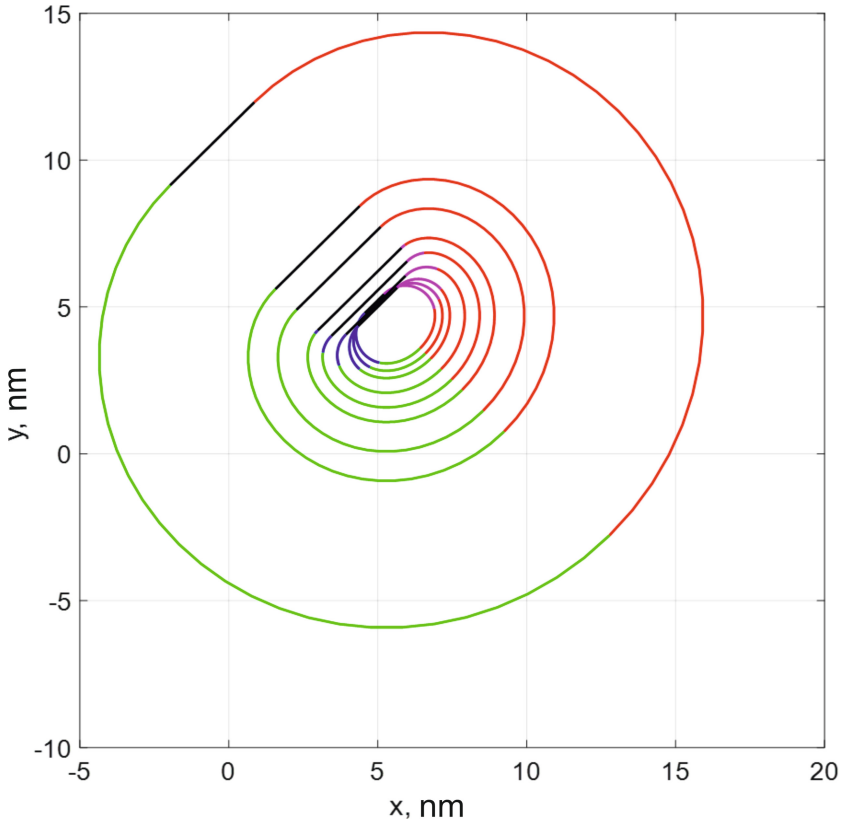


Fig. 1. Possible collisions areas in physical water space at moments $t = 6$ min, $t = 7.5$ min, $t = 9$ min, $t = 12$ min, $t = 15$ min, $t = 18$ min, $t = 24$ min, $t = 30$ min, $t = 60$ min

For the i -th target vessel

$$SCS_i(t_0, t_f) = \bigcup CSS_i(t) = \bigcup \frac{APC_i(t)}{t}, \forall t \in [t_0, t_f], \tag{10}$$

where t_f – planning horizon boundary.

$SCS_i(t_0, t_f)$ is a non-linear obstacle in the speed space, determined taking into account the entire set of reachability of the target vessel instead of a single calculated trajectory. The speed vectors outside the set $SCS_i(t_0, t_f)$ form the trajectories of the controlled

vessel with a constant course and speed, which are guaranteed to be safe in relation to all dynamically possible trajectories of the target vessel during the planning horizon.

In other words, the SCS can be considered as a set of all possible nonlinear single obstacles in the speed space $OS_i(t_0, t_f)$, associated with each dynamically feasible trajectory of the target ship, determined during the same planning horizon.

Both definitions can be considered equivalent, since both allow us to determine the boundaries of the set in the speed space corresponding to a given distance to the target vessel at a given time. In Eqs. (10), the SCS is defined in terms of the APC, while the SCS is a superset for all possible obstacles in the speed space $OS_i(t_0, t_f)$ of the target vessel:

$$OS_i(t_0, t_f) \subset SCS(t_0, t_f); OS(t_0, t_f) = \bigcup V(t). \quad (11)$$

In practice, both Eqs. (10) and Eqs. (11) can be used without sacrificing accuracy.

The SCS based on Eqs. (10) and (11) are identical when the SCS boundary does not contain a section, based on the nearest boundary of the region of possible collisions APC(t).

The use of reachability sets as functions of time is a convenient approach for determining the sets of possible collisions in speed space (SCS). On the other hand, modeling of all possible trajectories of the target vessel in the speed space on an infinite planning horizon to construct the boundaries of the SCS may require large computational costs.

The far boundary of the region of possible collisions in the speed space $CSS(K, t)$, where $K \in [-\pi, \pi]$; $t \in [t_0, t_f]$, is determined on the basis of Eqs. (4) - (9).

$$\dot{x}_i(K, t) = \frac{x_0 + X_i \cos K_{T0} + Y_i \sin K_{T0}}{t}; \dot{y}_i(K, t) = \frac{y_0 - X_i \sin K_{T0} + Y_i \cos K_{T0}}{t}; \quad (12)$$

The near boundary of the CSS is a straight line connecting the extreme points of the sections $[\dot{x}_1 \dot{y}_1]$ and $[\dot{x}_2 \dot{y}_2]$, or $[\dot{x}_3 \dot{y}_3]$ and $[\dot{x}_4 \dot{y}_4]$.

Figure 2 shows the set of possible collisions for the same target vessel as in Fig. 1, but already in the speed space. The envelope curve of the sets of possible collisions in the speed space $SCS_i(t_0, t_f)$ is the boundary of the obstacle in the speed space $OS_i(t_0, t_f)$, caused by the potentially possible movement of the target ship. The speed vector of the controlled vessel for guaranteed collision avoidance must be outside $OS_i(t_0, t_f)$. For example, for the case under consideration, the movement of a controlled vessel on a heading of 0° at a speed of 15 knots is guaranteed not to lead to a collision with the target ship, no matter how the target vessel maneuvers.

In situations with a large margin of time and significant unpredictability of the movement of target vessels, accurate and one-time programming of the entire maneuver may not be feasible. Instead, real-time programming is able to create and adapt controls as new information becomes available. The multifactor nature of the task often does not allow guaranteeing movement within the framework of preliminary passage plan. Instead, an iterative method of vessel motion generation [5, 6, 12–16], can be used, which should provide collision avoidance and at the same time minimize deviations from the preliminary passage plan.

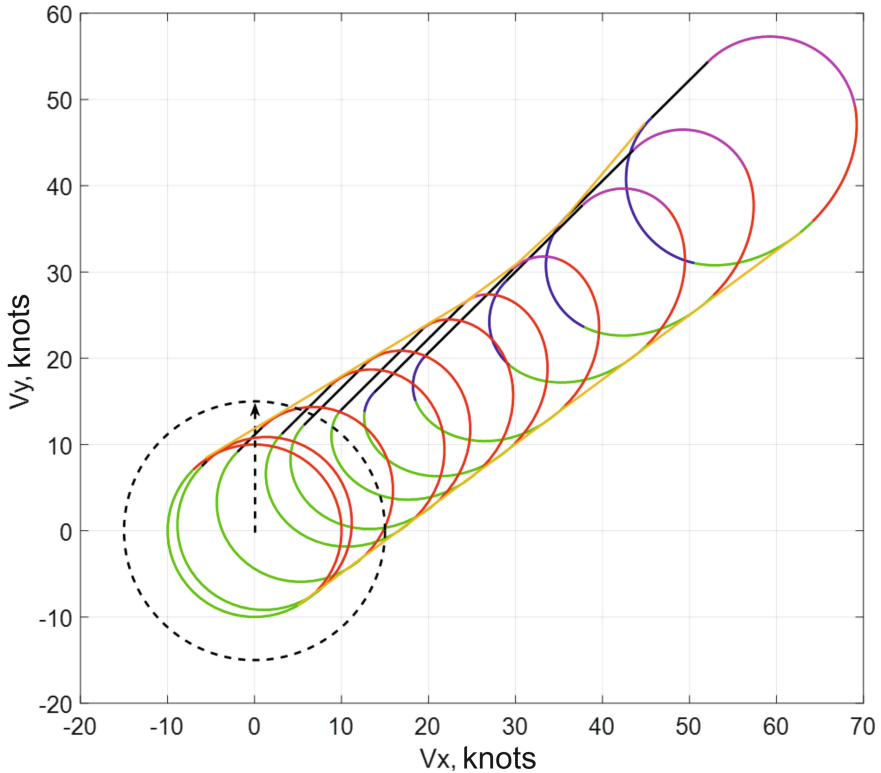


Fig. 2. Possible collisions areas in speed space at moments $t = 6$ мин, $t = 7.5$ min, $t = 9$ min, $t = 12$ min, $t = 15$ min, $t = 18$ min, $t = 24$ min, $t = 30$ min, $t = 60$ min, $t = 300$ min, $t = 60000$ min

5 Conclusion

In general, the safety of navigation is achieved by defining a space of safe states and restrictions on control, allowing the vessel to be only within this space. Based on [5, 6, 9] safe states are those states that do not violate the constraints associated with collision avoidance, and at the same time provide a transition to another safe state. For example, a decrease in speed to a complete stop with subsequent anchoring can be considered the achievement of a permanent safe state that propagates indefinitely in time, so any state from which the anchor can be dropped can be considered safe. However, in a real unpredictable navigational environment, it is much more difficult to determine sufficient conditions for a safe state, since even for a vessel at anchor, there is a possibility of collision with vessels underway, or the vessel may be surrounded by so many other vessels moving in such a way that a collision is inevitable.

The proposed method makes it possible to determine the sets of possible collisions based on the reachability sets of target vessels in the speed space in real time. Thus, the safety of navigation is guaranteed in the navigational environment of numerous vessels with dynamic limitations, but moving unpredictably, on an infinite planning horizon. The safety of navigation is achieved by combining the reachability sets as functions of time

for the target vessels, taking into account their dynamic capabilities and representing the sets in the speed space.

The approach considered above to the problem of guaranteed collision avoidance of vessels with respect to target vessels with dynamic constraints, but having unpredictable motion, ensures navigation safety over large planning horizons. The reachability set is considered as a set of all states corresponding to dynamically feasible predictable trajectories that can be realized from the current state. Determining the expected movement of a target vessel based on current parameters extends to determining the entire reachable set of a target vessel represented in speed space. Such sets remain limited even with the extension of the planning horizon to infinity. The control option, which is outside the reachable set of the target vessel in the speed space, is guaranteed to prevent a collision even in conditions of unpredictable movement of the target vessel. Thus, under the conditions of dynamic constraints and with known initial states, using an iterative approach, the safety of navigation on an unlimited planning horizon is guaranteed.

References

1. Zhuk, A.S.: The model of three-dimensional reachable set of ship motion. In: *Ekspluatatsiya morskogo transporta* **2**(83), 51–57 (2017). (in Russian)
2. Zhuk, A.S.: The three-dimensional ship's reachable and backward reachable sets with restricted control. *Ekspluatatsiya morskogo transporta*, № 2 (87), pp. 32–38 (2018). (in Russian)
3. Kumkov, S.I., Patsko, V.S., Pyatko, S.G., Fedotov, A.A.: Construction of a solvability set in the problem of aircraft handling under wind disturbance. *Trudy instituta matematiki i mekhaniki UrO RAN*. 2005. Tom 11, No. 1. S, pp. 149–159. (in Russian)
4. Fedotov, A., Patsko, V., Turova, V.: Reachable Sets for Simple Models of Car Motion. *Recent Advances in Mobile Robotics*, pp. 147–172 (2011)
5. LaValle, S.M.: *Planning Algorithms*. Cambridge University Press, Cambridge (2006)
6. Wu, A.: *Guaranteed Avoidance of Unpredictable, Dynamically Constrained Obstacles using Velocity Obstacle Sets*. Massachusetts: Massachusetts Institute of Technology, 116p. (2011)
7. Lin, Y., Saripalli, S.: Path planning using 3D Dubins Curve for Unmanned Aerial Vehicles. In: *2014 International Conference on Unmanned Aircraft Systems, ICUAS 2014 - Conference Proceedings*, pp. 296–304 (2014)
8. Zhuk, A.S.: Ship motion model based on the program of waypoint following // *Transport: nauka, tekhnika, upravlenie*. *Nauchnyi informatsionnyi sbornik* **6**, 32–36 (2020). (in Russian)
9. Korenev, G.V.: *Purpose and adaptability of motion*. M.: Nauka, 528 p. (1974). (in Russian)
10. Chernousko, F.L.: *State estimation for dynamic systems*. M.: Nauka, 319 p. (1988). (in Russian)
11. Hua, C., Yangang, L., Lei, C., Guojin, T.: Reachable set modeling and engagement analysis of exoatmospheric interceptor. *Chinese J. Aeronautics* **27**(6), 1513–1526 (2014)
12. Shima, T., Rasmussen, S.: *UAV Cooperative Decision and Control*, p. 164p. Society for Industrial and Applied Mathematics, Philadelphia (2009)
13. Allen, R.E., Clark, A.A., Starek, J.A., Pavone, M. A.: *Machine Learning Approach for Real-Time Reachability Analysis*. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE (2014)
14. Lin, Y., Saripalli, S.: Collision avoidance for UAVs using reachable sets. In: *2015 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE (2015)

15. Saravanakumar, A., Kaviyarasu, A.: Ashly Jasmine R. Sampling based path planning algorithm for UAV collision avoidance. *Sādhanā* **46**, 112 (2021)
16. Holmes, P., Kousik, S., Zhang, B., Raz, D., Barbalata, C., Johnson-Roberson, M., Vasudevan, R.: Reachable Sets for Safe, Real-Time Manipulator Trajectory Design. *Robotics: Science and Systems* (2020)
17. Mitchell, I.M., Bayen, A.M., Tomlin, C.J. A time-dependent Hamilton-Jacobi formulation of reachable sets for continuous dynamic games. *IEEE Trans, Automatic Control* **50**, 947–957 (2005)
18. Kheckert, E.V., Dantsevich, I.M., Lyutikova, M.N., Khaleeva, E.P.: The technology of geophysical vessel control for the study of the world ocean. In: *IOP Conference Series: Earth and Environmental Science*, vol. 872, No. 1, p. 012001 (October 2021). IOP Publishing
19. Dantsevich, I., Lyutikova, M., Fedorenko, V.: Numerical Method for Correcting Command Signals for Combined Control of a Multiengined Complex. In: *International Conference on Mathematics and its Applications in new Computer Systems*, pp. 117–131. Springer, Cham (2022)
20. Vasyutina, A.A., Popov, V.V., Kondratyev, A.I., Boran-Keshishyan, A.L.: Improvement of the vessel traffic control system for accident-free electronic navigation in the port area. *J. Phys. Conf. Ser.* **2061**(1), 012105 (2021)
21. Astrein, V.V., Kondratyev, S.I., Boran-Keshishyan, A.L.: Multicriteria assessment of optimal forecasting models in decision support systems to ensure the navigation safety. *J. Phys. Conf. Ser.* **2061**(1), 012108 (2021)



An Ensemble of UNet Frameworks for Lung Nodule Segmentation

Nandita Gautam¹, Abhishek Basu², Dmitry Kaplun³(✉), and Ram Sarkar¹

¹ Department of Computer Science and Engineering, Jadavpur University,
Kolkata, India

² Department of Computer Science and Engineering, National Institute of
Technology, Durgapur, Durgapur, India

³ Department of Automation and Control Processes, Saint Petersburg
Electrotechnical University “LETI”, Saint Petersburg, Russian Federation
dikaplun@etu.ru

Abstract. One of the most common types of cancer in the world is lung cancer, which is a cause of increasing mortality. It is most often discovered in the middle and later stages as it does not have obvious symptoms due to which its treatment is often missed. Studies show that most lung cancers are in the form of lung nodules, which can be categorized as benign or malignant. Thus, accurate early identification of malignant lung nodules that might later become cancerous is essential for the prevention of lung cancer. Computed Tomography (CT) images can be useful for identifying and segmenting these nodules. In this paper, we propose an ensemble model, called BUS-UNet++, to segment lung nodules using CT images. We use a combination of the ConvLSTM up-sampling architecture from BUS-UNet and skip connections in aggregation blocks from the UNet++ model to build the BUS-UNet++ ensemble. The dataset for this study is collected from the Lung Image Database Consortium Image Database Resource Initiative (LIDC-IDRI), where we have selected the modality of these images as CT images. These are further preprocessed to obtain nodule masks and nodule images, which constitute the Region of Interest (RoI). The accuracy achieved with the Adam Optimizer is 0.9731, the Intersection over Union (IoU) of about 0.8439, and the dice score coefficient (DSC) of about 0.958 are obtained by the proposed system. This ensemble model outperforms several state-of-the-art models used for the same purpose.

Keywords: Lung nodule · Image segmentation · U-Net++ · Ensemble learning · LIDC-IDRI

1 Introduction

There have been several deaths worldwide due to lung cancer disease, and research has been growing in this area ever since. Early detection of cancerous nodules as well as tumor stage identification are very crucial in the treatment of lung cancer. Researchers have developed several computer-aided diagnosis

(CAD) systems have been developed by researchers with the aim of reducing errors and assisting in human observation. These systems are mostly backed up by machine learning and deep learning algorithms. Although there is a limit to the image dataset availability of cancer patients, with the help of advanced deep learning based techniques, the classification and segmentation of the lung nodules have seen major progress in the past decade.

In this paper, we have proposed one such technique for lung nodule segmentation. Our proposed model, called BUS-UNet++, is an ensemble of BUS-UNet and UNet++ models. For evaluation purposes, we have considered a publicly available standard dataset, called the LIDC-IDRI dataset. One sample image from the said dataset is shown in Fig. 1

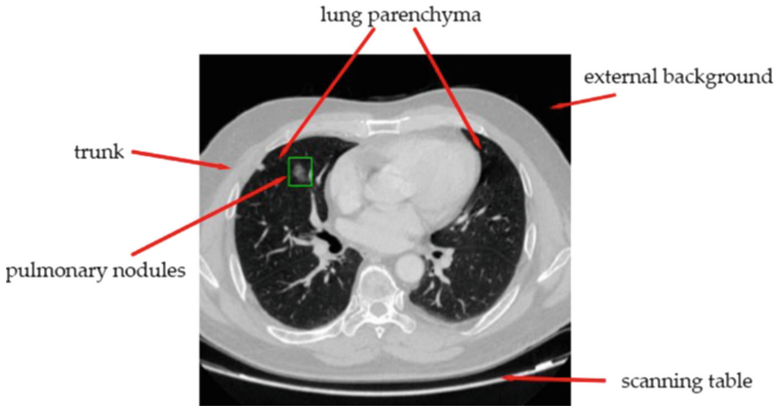


Fig. 1. A sample lung nodule image from the LIDC-IDRI dataset

UNets are highly popular due to their excellent performance in the field of biomedical image segmentation. Table 1 gives a summarized view of the state-of-the-art methods proposed in the field of lung nodule segmentation on the LIDC dataset. Yeganesh et al. [2] used a Bi-directional ConvLSTM U-Net having densely connected convolutions (BCDU-Net) for the segmentation of lung nodules. They also added a ResNet block to the architecture to achieve better performance. The algorithm gave a dice score coefficient of about 93.83% on the LIDC-IDRI dataset. Weihua et al. [3] used a perfectly symmetric structure deep convolution neural network (CNN), called U-Net-DCNN. They introduced skip-connections in the network to improve feature representation and achieved an overall dice score coefficient of 91.97%.

In [5], the authors proposed a 2DSegU-Net architecture, which is a hybrid of 2DU-Net and SegNet architecture to improve automated lung nodule detection from CT scans from the LIDC dataset. However, their proposed hybrid model with dice score coefficient 84.3% showed only a slight improvement over the 2D-UNet with dice score coefficient 83.0%. Usman et al. [7] obtained average

Intersection over Union (IoU) of 0.7485, which was competitive with other methods. They used adversarial augmentation method to improve robustness of the detection framework, followed by training with a 2D Residual U-Net. Apart from the UNet architecture, Aresta et al. [10] proposed a 3D iW-Net model which is an automatic segmentation based deep learning model on the CT images, where they achieved a sensitivity of 95.4%. Hongyang et al. [13] suggested a novel technique using multipatch based learning wherein they used a filter for enhancing the patches cut out from the lung images. They initially performed long contour mending followed by parenchyma segmentation and vessel elimination, after which they trained the CNN for nodule detection. They performed the max-pooling operation followed by a center-pooling operation. They achieved a good performance with an overall sensitivity of 94%. Tong et al. [16] implemented the 2D U-Net architecture based on CNNs. They rebuild the dataset and further utilized layers of convolutional, pooling and upsampling operations. They re-designed skip pathways in the architecture to more accurately segment for lesions in the lung image that appear at multiple scales. The performance of their proposed architecture was slightly improved with a dice score coefficient of 73.6%. Wang et al. [15] proposed the 3D Region based Convolutional Neural Network (RCNN) to segment 3D volume from a sparse annotation in which they basically modified the 2D MaskRCNN detection algorithm. They tested it over 547 CT images from LIDC dataset and achieved a dice score coefficient of 64%. In this work, we propose a combination of UNet++, which is a variant of UNet, (see Fig. 2) and BUS-UNet to segment the lung nodules. Overall pipeline of the proposed methodology is shown in Fig. 3 and the model architecture is shown in Fig. 4.

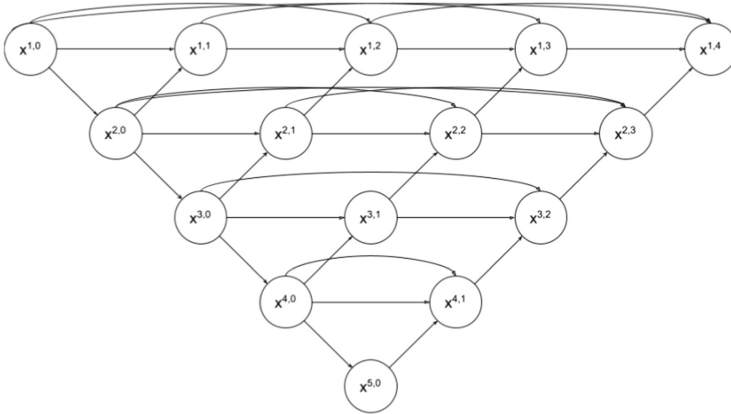


Fig. 2. A diagram showing the U-Net++ architecture

Table 1. Existing methods on lung nodule segmentation on LIDC dataset.

Authors	No. of nodules	Model	DSC (%)	Sensitivity (%)
Mamtha et al. [1]	1018 images	Integratingwater cycle algorithm and Bat algorithm	82.08	-
Yeganeh et al. [2]	1714 images	Res BCDU-Net	93.83	-
Weihua et al. [3]	1186 images	UNet DCNN	91.97	-
Guilherme et al. [4]	888 images	Automatic Detection with Gaze Information	-	69
Rocha et al. [5]	2653 nodules	2DU-NET 2DSegU-Net	83.0 84.3	89.8 85.8
Sunyi et al. [6]	1018 images	MIP-CNN	-	95.4
Usman et al. [7]	893 nodules	2D Residual U-Net	87.55± 10.58	91.62± 8.47
Amorim et al. [8]	1018 images	2D Modified U-Net	83	-
Shi et al. [9]	700 nodules	2DVGG 16+ SVM	-	90.00
Aresta et al. [10]	1012 images	3D iW-Net	75	-
Qin et al. [11]	1182 nodules	3D CGAN+3D CNN	84.33	85.11
Liu et al. [12]	3556 images	2D Mask R-CNN	-	-
Hongyang et al. [13]	1018 images	Multigrouppatch-based learning using Frangi Filter	-	94
Wu et al. [14]	1404 images	3D multi-task andInterpretableCNN	74.05	-
Wang et al. [15]	547 images	3D nodule R-CNN	64	-
Tong et al. [16]	1245 nodules	2D improved U-Net	73.6	-

2 Materials and Methods

2.1 UNet++

UNet++ [17] is a powerful architecture for medical image segmentation, which is a deeply-supervised encoder-decoder network, where the encoder and decoder subnetworks are connected through a series of nested, dense skip pathways. These aggregating blocks are aimed at aggregating features at varying semantic scales. Other than these essential skip connections, there are likewise skip connections from the encoder to aggregating blocks and other skip connections from aggregating blocks to either other features blocks or decoder layers. The aggregating blocks connect highlights from various scales and afterward apply convolutions to them prior to passing them further. The quantity of aggregating blocks is proportional to the square of the network depth. UNet++ also has a prunable architecture with deep supervision.

UNet++ is made up of UNets of varying depths, with their decoders densely connected at the same resolution using redesigned skip pathways. The architectural changes made in UNet++ enable the following benefits: First, because it embeds U-Nets of varying depths in its architecture, UNet++ is not susceptible to network depth selection. All of these UNets share an encoder in part, and their decoders are linked. All of the constituent UNets are trained simultaneously while benefiting from a shared image representation by training UNet++ with deep supervision. This design not only improves overall segmentation performance but also allows for model pruning during inference.

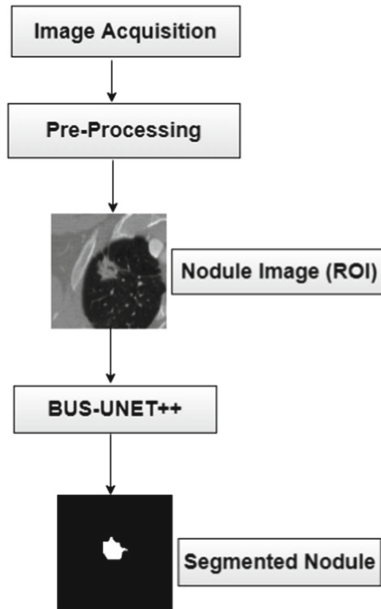


Fig. 3. Overall pipeline of the proposed methodology

Secondly, UNet++ is not hampered by overly restrictive skip connections that allow only the same-scale feature maps from the encoder and decoder to be fused.

2.2 BUS-UNet

The original proposed model, Big-U Small-U Net (BUS-UNet) [19], has 108 layers formed by chaining two BCDUNets, the first of which is deeper than the second. The Big-U, in particular, is deeper than the original BCDU-Net [2], while the Small-U is the same size as the original.

BCDU-Net was proposed as an extension of UNet, and it outperformed state-of-the-art segmentation alternatives. In its architecture, the contracting path consists of four steps, each with two convolutional 3×3 filters followed by a 2×2 maxpooling and a rectified linear unit (ReLU) activation function. The number of feature maps is doubled at each step. Image representations are extracted progressively along the contracting path, increasing the dimension of the representations layer by layer. Densely connected convolutions were proposed as a solution to the U-Nets' problem of learning redundant features in successive convolutions. They begin each step of the decoding path with an up-sampling function over the previous layer's output. Unlike in the original U-Net, where the corresponding feature maps in the contracting path are cropped and pasted onto the decoding path, the feature maps are processed with bidirectional convolutional LSTMs (BConvLSTM). Following each up-sampling procedure, the outputs are batch normalized, which improves the network stability by standardizing the network layer's inputs by subtracting the batch mean and dividing the result by the batch standard deviation. Batch normalization aids in the neural network training speed. A BConvLSTM layer receives the outputs after batch normalization. The BConvLSTM layer processes input data into forward and backward paths, from which a decision is made for the current input by addressing data dependencies in both directions. It should be noted that only the data dependencies for the forward path are processed in the original ConvLSTM.

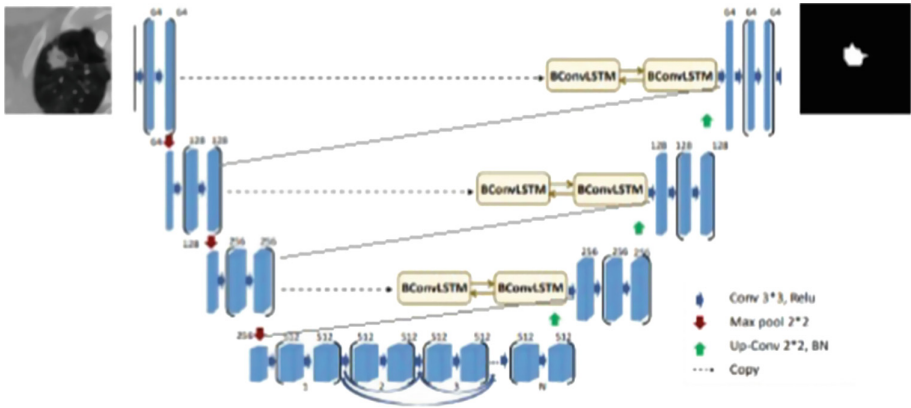


Fig. 4. Architecture of the proposed model, called BUS-UNet++.

2.3 Data Pre-processing

The training of the model essentially requires the medical image data. There is an imbalance in the positive and negative sample categories in the dataset that might have an impact on the performance of the neural network. This is

mainly due to the fact that the network is unable to reach its optimal value of the weights due to class imbalance. Therefore, we utilize the Augmented Generative Adversarial Network (AugGAN) [20] for the data enhancement. It solves the problem of limited positive samples to some extent in the dataset. Also, it reduces the possibility of model overfitting. Some samples images generated by the AugGAN model are shown in Fig. 5.



Fig. 5. Lung nodules as generated by the AugGAN model

2.4 Region of Interest (ROI)

We have extracted the ROI slice with a dimension of 512×512 from the scan in the LIDC-IDRI database. We have used the centroid information for this purpose. Further, we have cropped the slice into a 64×64 sized image with the same centroid. In this way, the ROI images are obtained as shown in Fig. 6 and Fig. 7.

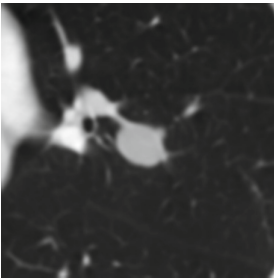


Fig. 6. ROI image [LIDC-IDRI-0014]



Fig. 7. Mask image [LIDC-IDRI-0014]

2.5 Proposed Model: BUS-UNet++

We have proposed the BUS-UNet++ network in this work, which has the BUS-UNet and UNet++ architectures in its background. We have implemented the BUS-UNet++ as an extension of these two architectures for medical image segmentation. The BUS-UNet++ consists of 68 layers. We have added four steps in the contracting path of the architecture, each with two convolutional with 3×3 filters followed by a 2×2 maxpooling and a ReLU activation function. At each

step, we have doubled the number of feature maps. In the contracting path, image representations are extracted progressively, increasing the dimension of the representations layer by layer. In BUS-UNet++, we have introduced direct skip connections between respective layers of encoder and decoder parts while also maintaining skip connections from the encoder to aggregating blocks. The skip connections from aggregating blocks to decoder layers and other aggregating blocks have been extended from the UNet++ model. We have used the aggregating blocks to concatenate features from different scales and then we have applied the convolution operation on them before passing them to the next layer. In each aggregation block, we have used the max-pooling operation, convolution layers, and dropout layers (to prevent overfitting). During each up-sampling operation, we have applied batch-normalization so as to standardize the inputs to the network layer. After this step, in order to process the input through the forward and backward paths, we have passed it to the bidirectional ConvLSTM layer.

3 Experimental Results

3.1 Dataset Description

In this study, the LIDC-IDRI dataset was used to evaluate the proposed model. This dataset has a total of 1186 lung nodules in a set of 888 CT images. For the experimentation, we have randomly divided the data into three parts, wherein 70% is used as training data, 20% is used as testing data, and 10% of the data is used for validation data. In order to judge a nodule in the LIDC-IDRI dataset, the annotations are done by four radiologists, where a nodule is determined by its radius. Nodules with a radius of greater than 3 mm are identified as nodules belonging to the RoI, whereas nodules with a radius of less than 3 mm are annotated as non-nodules.

3.2 Evaluation Metrics

This section describes the metrics which are used to evaluate the proposed method implemented in this paper. The first one is the dice score coefficient (DSC) index, which refers to the degree of fit between the original target and the segmented target.

The DSC value is higher based on how much the two objects fit and simultaneously lower the loss function value, and it indicates that the segmentation model of the lung nodules is more accurate. DSC is calculated using Eq. (1):

$$DSC = \frac{2TP}{2TP + FP + FN} \quad (1)$$

Here, true positive (TP) represents the area where the lung nodule exists as well as its correctly segmented area, false positive (FP) represents the area where

the lung nodule exists but is not correctly segmented, and false negative (FN) represents the area where the lung nodule does not exist and is not segmented. When the loss function is infinitely close to 0, the invalid dice score coefficient reaches close to 1. In an ideal scenario, the model segmentation result matches the real result at this point. The relationship between DSC and loss coefficient is defined in Eq. (2):

$$loss = 1 - DSC \quad (2)$$

IoU: It is another standard metric for semantic image segmentation that we have used in this study to evaluate our model performance.

We have obtained an IoU of 0.84 and a DSC of 0.958 with our proposed method as shown in Table 2 and Table 3.

Table 2. Lung nodule segmentation results of BUS-UNet++ with varied optimizers.

Optimizer	Accuracy	DSC	IoU
Adam	0.9771	0.9580	0.8439
SGD	0.9529	0.8594	0.7521
AdaGrad	0.9470	0.6771	0.7289
AdaDelta	0.9180	0.5591	0.6011

Table 3. Performance comparison of the proposed model with state-of-the-art methods on the LIDC-IDRI dataset

Work Ref.	Method	DSC (%)
Badrinarayanan et al. [21]	SegNet	84.21
Ronneberger et al. [22]	U-Net	82.90
Zhou et al. [17]	UNet++	83.32
Huang et al. [18]	UNet-3+	89.72
Weihua et al. [3]	UNet DCNN	91.97
Yeganeh et al. [2]	Res BCDU-Net	93.83
Khoong et al. [19]	BUS-UNet	94.25
Proposed	BUS-UNet++	95.80

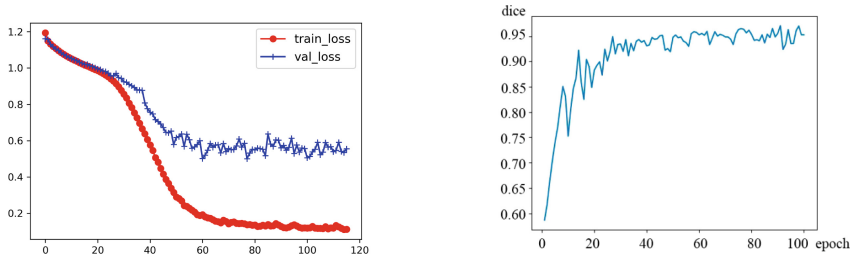


Fig. 8. Training and validation loss values (left) and the network DSC value (right) of the proposed model.

4 Discussion

This paper analyzes the performance of the various standard models along with the proposed model through a set of experiments. All the evaluation metrics mentioned are calculated and it is observed that the proposed model gives the best result with Adam optimizer, with an accuracy rate of 0.9771, as the loss function stabilizes more quickly as compared to other optimizers tested for the same purpose. The BUS-UNet++ helps to improve the segmentation accuracy of the lung nodules and has achieved good results with a dice score coefficient of 0.9580 as shown in Fig. 8. The model also outperforms several state-of-the-art networks such as UNet DCNN and BUS-UNet that have generated a dice score coefficient of 0.91 and 0.94 respectively. However, there are a significant number of false positives which can be reduced by improving the feature extraction pipeline before training the network.

5 Conclusion

Lung cancer is one of the most common types of cancer in today's world. Research reveals that accurate early identification of malignant lung nodules is essential for the prevention of this disease. In this paper, we have proposed a U-Net based model, called BUS-UNet++, for segmenting the lung nodules in the LIDC-IDRI dataset. The BUS-UNet++ network is a combination of UNet++ and BUS-UNet architectures. The CT images for this purpose are extracted from the LIDC public repository and further preprocessed to obtain the nodule image along with its mask image. A set of experiments has been performed with several U-Net variants such as U-Net++, UNet-3++ and BUS-UNet to generate the segmentation maps. The performance metrics such as accuracy, DSC and IoU for BUS-UNet++ have shown that it outperforms the other three variants used in this study. Although this paper has focused only on the segmentation of lung nodules, it can be further extended by performing a classification of the segmented nodules and classifying them as benign or malignant nodules.

Acknowledgment. The authors are grateful for the infrastructural support provided by the Centre for Microprocessor Applications for Training, Education and Research (CMATER) Laboratory of the Computer Science and Engineering Department, Jadavpur University, Kolkata, India.

References

1. Shetty, M.V., Jayadevappa, D., Veena, G.N.: Water Cycle bat algorithm and dictionary-based deformable model for lung tumor segmentation. *Int. J. Biomed. Imaging* (2021). Article ID 3492099, 12 pages. <https://doi.org/10.1155/2021/3492099>
2. Jalali, Y., Fateh, M., Rezvani, M., Abolghasemi, V., Anisi, M.H.: ResBCDU-Net: a deep learning framework for lung CT image segmentation. *Sensors* **21**(1), 268 (2021). <https://doi.org/10.3390/s21010268>
3. Liu, W., Liu, X., Li, H., Li, M., Zhao, X., Zhu, Z.: Integrating lung parenchyma segmentation and nodule detection with deep multi-task learning. *IEEE J. Biomed. Health Inform.* **25**(8), 3073–3081 (2021). <https://doi.org/10.1109/JBHI.2021.3053023>
4. Aresta, G., et al.: Automatic lung nodule detection combined with gaze information improves radiologists' screening performance. *IEEE J. Biomed. Health Inform.* **24**(10), 2894–2901 (2020). <https://doi.org/10.1109/JBHI.2020.2976150>
5. Rocha, J., Cunha, A., Mendonça, A.M.: Conventional filtering versus u-net based models for pulmonary nodule segmentation in CT images. *J. Med. Syst.* **44**(4), 1–8 (2020). <https://doi.org/10.1007/s10916-020-1541-9>
6. Zheng, S., Guo, J., Cui, X., Veldhuis, R.N.J., Oudkerk, M., van Ooijen, P.M.A.: Automatic pulmonary nodule detection in CT scans using convolutional neural networks based on maximum intensity projection. *IEEE Trans. Med. Imaging* **39**(3), 797–805 (2020). <https://doi.org/10.1109/TMI.2019.2935553>
7. Usman, M., Lee, B.-D., Byon, S.-S., Kim, S.-H., Lee, B., Shin, Y.-G.: Volumetric lung nodule segmentation using adaptive ROI with multi-view residual learning. *Sci. Rep.* **10**(1), 1–15 (2020)
8. Amorim, P.H.J., de Moraes, T.F., da Silva, J.V.L., Pedrini, H.: Lung nodule segmentation based on convolutional neural networks using multi-orientation and patchwise mechanisms. In: Tavares, J.M.R.S., Natal Jorge, R.M. (eds.) *VipIMAGE 2019*. LNCVB, vol. 34, pp. 286–295. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32040-9_30
9. Shi, Z., et al.: A deep CNN based transfer learning method for false positive reduction. *Multimedia Tools Appl.* **78**(1), 1017–1033 (2018). <https://doi.org/10.1007/s11042-018-6082-6>
10. Aresta, G., Jacobs, C., Araújo, T., et al.: iW-Net: an automatic and minimalistic interactive lung nodule segmentation deep network. *Sci. Rep.* **9**, 11591 (2019). <https://doi.org/10.1038/s41598-019-48004-8>
11. Qin, Y., Zheng, H., Huang, X., Yang, J., Zhu, Y.M.: Pulmonary nodule segmentation with CT sample synthesis using adversarial networks. *Med Phys.* **46**(3), 1218–1229 (2019). Epub 2019 Jan 31. PMID: 30575046. <https://doi.org/10.1002/mp.13349>
12. Liu, M., Dong, J., Dong, X., Yu, H., Qi, L.: Segmentation of lung nodule in CT images based on mask R-CNN. In: 2018 9th International Conference on Awareness Science and Technology (iCAST), pp. 1–6 (2018). <https://doi.org/10.1109/ICAwST.2018.8517248>

13. Jiang, H., Ma, H., Qian, W., Gao, M., Li, Y.: An automatic detection system of lung nodule based on multigroup patch-based deep learning network. *IEEE J. Biomed. Health Inform.* **22**(4), 1227–1237 (2018). <https://doi.org/10.1109/JBHI.2017.2725903>
14. Wu, B., Zhou, Z., Wang, J., Wang, Y.: Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 1109–1113 (2018). <https://doi.org/10.1109/ISBI.2018.8363765>
15. Wang, W., Lu, Y., Wu, B., Chen, T., Chen, D.Z., Wu, J.: Deep active self-paced learning for accurate pulmonary nodule segmentation. In: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (eds.) *MICCAI 2018*. LNCS, vol. 11071, pp. 723–731. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00934-2_80
16. Tong, G., Li, Y., Chen, H., Zhang, Q., Jiang, H.: Improved U-NET network for pulmonary nodules segmentation. *Optik* (2018)
17. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Trans. Med. Imaging* **39**(6), 1856–1867 (2020). <https://doi.org/10.1109/TMI.2019.2959609>
18. Huang, H., et al.: UNet 3+: a full-scale connected unet for medical image segmentation. In: *ICASSP 2020–2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1055–1059. IEEE (2020)
19. Khoong, W.H.: BUS-UNet: An Ensemble U-Net Framework for Medical Image Segmentation
20. Huang, S.-W., Lin, C.-T., Chen, S.-P., Wu, Y.-Y., Hsu, P.-H., Lai, S.-H.: AugGAN: cross domain adaptation with GAN-based data augmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) *ECCV 2018*. LNCS, vol. 11213, pp. 731–744. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01240-3_44
21. 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)
22. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) *MICCAI 2015*. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28



No-Reference Metrics for Images Quality Estimation in a Face Recognition Task

Aleksander S. Voznesensky¹, Aleksandr M. Sinitca¹, Evgeniy D. Shalugin²,
Sergei A. Antonov², and Dmitrii I. Kaplun²(✉)

¹ Centre for Digital Telecommunication Technologies, St. Petersburg Electrotechnical University “LETI”, 197022 St. Petersburg, Russia
{asvozesenskiy,amsinitca}@etu.ru

² Department of Automation and Control Processes, St. Petersburg Electrotechnical University “LETI”, 197022 St. Petersburg, Russia
{edshalugin,saantonov,dikaplun}@etu.ru

Abstract. No-reference metrics (BRISQUE, NIQE, PIQE) and full-reference metrics (PSNR, SSIM) for face recognition quality estimation are presented in this paper. Different noise types are considered: gaussian and salt&pepper noise with SNR in range 0...40 dB. Regression models between full-reference and no-reference metrics are considered. The quality of the regression models is estimated via R^2 and RMSE.

Keywords: no-reference image quality estimation · full-reference image quality estimation · peak signal-to-noise ratio · structural similarity · R-squared

1 Introduction

One of the main parts of data processing is the quality assessment of the sample, but it is usually a non-trivial issue. And the face recognition task is not an exception. For example, Qiang Meng et al. [9] shown, that quality of an image has a strong correlation with recognition quality.

There are several ways to use image quality assessment for a Face Recognition (FR) task. One of the most popular use-cases is to use it as a preprocessing step of an FR model, so that there is no need to trigger the recognition process itself if the quality is too poor [5]. Another well-known usage is to make conditional enhancement based on the estimation of image quality [4, 13]. Also image quality estimation in FR task can be used for quality summarization, video frame selection, face detection filters, compression control, and others [12].

Another usage, which we propose in this paper, which has not been discussed and researched yet to the best of our knowledge is to use image quality assessment algorithm as a discrimination algorithm for images with poor quality, which are not suitable as references. In this use case, the problem appears to be not how to get a good quality image to use in a face recognition (FR) model, but how to get a reference image that could be good enough for different FR models to increase the accuracy and solve the problem of pure quality references.

Obviously, image quality is important for the inference stage as for the new reference adding stage because a low quality reference can cause many false positive events in real-world systems. On the basis of this assumption, we research methods for preliminary estimation of the quality of reference images, which will be done at the reference preparation stage. The main challenge is the absence of high-quality reference for each of a reference; thus a no-reference metric must be used.

This study is designed to improve the quality of face recognition in video surveillance systems. The main goal of the work is to estimate the dependence between the no-reference and full reference metrics to determine the quality of the images for the face recognition task. The study evaluates 3 no-reference metrics (BRISQUE, NIQE, PIQE) and two full-reference metrics (PSNR, SSIM).

The quality of face reference images can be divided into two parts. The first is features such as face orientation, occlusion, etc. The second is general image quality such as blur, noise, etc. We aimed at the second quality estimation.

2 Materials and Methods

2.1 Dataset

Labeled Faces in the Wild (LFW) aligned by deep funneling [7] was used as a basic dataset. It contains 13233 images from the web and has 1680 different identities with two or more images per person.

To study approaches to image quality assessment, synthetic augmented dataset was created based on LFW. All images were converted to grayscale. Then the following noises were added to the images:

1. Gaussian White Noise.
2. Salt & Pepper Noise.

11 noise levels were added to each image for each noise type. All noise levels have been generated according to the fixed array of SNR values: [0, 1, 3, 6, 8, 10, 12, 15, 20, 30, 40] dB to allow cross-analysis of noise according to the SNR. Thus, a dataset of 291126 noisy images was obtained.

The dataset was randomly divided into the training (80%) and testing (20%) subsets.

Gaussian White Noise. The mathematical model of Gaussian noise is described using a normal distribution, by changing the mathematical expectation, one can achieve a lighter or darker image, and by changing the variance, one can obtain multicolored noise normally distributed in the image.

$$p(x) = \sqrt{\frac{1}{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (1)$$

where x – gray value, σ – standard deviation and μ – mean.

Gaussian White Noise is ergodic random process with zero mean (μ) [8].

The main sources of Gaussian noise are caused by poor lighting and/or high temperature. This noise is also called electronic noise, as it occurs on sensors and amplifiers. The sources of this noise are thermal vibrations of atoms or thermal radiation of objects [1].

To generate Gaussian White Noise one should set the variance and the mean of a noise. Since $\mu = 0$, the variance is the only parameter.

To generate 11 levels of noise for each image required noise variance was calculated according to the formula:

$$\sigma_n^2 = \frac{\sigma_s^2}{10^{\frac{SNR}{10}}}, \tag{2}$$

where σ_s^2 – is a variance of an image (signal) and $SNR \in [0, 1, 3, 6, 8, 10, 12, 15, 20, 30, 40]$ dB.

Salt and Pepper Noise. Impulse Value Noise is also called data dropout noise, since it is mainly associated with errors in image transmission or storage [2].

This type of noise is manifested during the information transmission due to errors in bit transmission, errors in analog-to-digital conversion, non-working “pixels” on the camera sensor or image storage errors.

Impulse Value Noise accidentally changes some pixel values. For “salt” part of the noise, there is a probability of changing the pixel value to a , and for “pepper” noise there is a probability of changing the value to b , where e.g. a and b could be the max and minimum possible pixel values.

The following is the distribution density function for Salt & Pepper noise (for gray scale image):

$$P(x) = \begin{cases} P_s, x = s; \\ P_p, x = p; \\ 0 \text{ otherwise} \end{cases} . \tag{3}$$

In S&P noise, with probability P_s , the pixel value will become $x = s$ and with probability P_p will become $x = p$, there can be no other transformations, so in other cases the probability = 0. From the definition of the distribution density function, it is clear that $P_p = 1 - P_s$.

To generate Impulse Value Noise the one should set the amount and the proportion noise parameters. Amount indicates the ratio of the number pixels affected by noise to the general pixels number in image. Proportion indicates the probability of affected pixel to be S&P.

Because we don't generate Impulse Noise with variance parameter, but with amount and proportion, to generate 11 noise levels one should express variance of an Impulse Noise via amount and proportion:

$$\begin{aligned} \sigma_n^2 = & a * b * ((s - a * (b * s + p - b * p))^2) + \\ & a * (1 - b) * ((p - a * (b * s + p - b * p))^2) + \end{aligned}$$

$$(1 - a) * ((0 - a * (b * s + p - b * p))^2), \quad (4)$$

where a – amount, b – proportion, s – “salt” value of a pixel, p – “pepper” value of a pixel.

To generate 11 noise levels of possible variances, noise was calculated with a from 0.00001 to 0.99 with steps 0.00001 and $b = 0.5$.

Then, for each image and SNR, the required noise variance was calculated, and the nearest value was taken from the array of possible noise variances. The corresponding to required noise variance parameters of amount and proportion were determined and the Impulse Noise was generated.

2.2 Image Quality Assessment – IQA

Image quality (IQ) can be assessed using objective or subjective methods. Objective methods based on image quality assessments are performed by different algorithms that analyze the distortions and degradations introduced in an image. Subjective methods based on the way in which humans experience an image quality. Objective and subjective methods of quality assessment don’t necessarily correlate with each other [15].

Subjective methods for image quality assessment belong to the large area of psychophysics research, a field that studies the relationship between physical stimulus and human perceptions. A subjective IQA method will typically consist on applying mean opinion score (MOS) techniques, where a number of viewers rate their opinions based on their perceptions of image quality. These opinions are afterwards mapped onto numerical values 0–100.

These methods can be classified depending on the availability of the source and test images:

1. Single-stimulus: the viewer only has the test image and is not aware of the source image.
2. Double-stimulus: the viewer has both the source and test image.

Since visual perception can be affected by environmental and viewing conditions, the International Telecommunication Union produced a set of recommendations for standardized testing methods for subjective image quality assessment.

IQA algorithms take an arbitrary image as input and output a quality score as output [15]. There are three types of IQAs:

1. Full-Reference IQA: Here you have a ‘clean’ reference (non-distorted) image to measure the quality of your distorted image. This measure may be used in assessing the quality of an image compression algorithm where we have access to both the original image and it’s compressed version.
2. Reduced-Reference IQA: Here you don’t have a reference image, but an image having some selective information about it (e.g. watermarked image) to compare and measure the quality of distorted image.

3. **Objective Blind or No-Reference IQA:** The only input the algorithm gets is the image whose quality you want to measure. This is thus called, No-Reference or Objective-Blind.

An example of reference-based evaluations is the Peak Signal-to-Noise Ratio (PSNR) and the Structural Similarity Index (SSIM). No-reference image quality assessment (Blind/Referenceless Image Spatial Quality Evaluator (BRISQUE), Natural Image Quality Evaluator (NIQE), Perception based Image Quality Evaluator (PIQE)), does not require a base image to evaluate image quality, the only information that the algorithm receives is a distorted image whose quality is being assessed. These metrics are commonly used to analyze the performance of algorithms in different fields of computer vision like image compression, image transmission, and image processing.

In this paper, we will discuss No-Reference IQA Metrics. Before we go deeper into the theory, let's first understand two basic terms:

1. **Distorted Image.** A distorted image is a version of the original image that is distorted by blur, noise, watermarking, color transformations, geometric transformations and so on and so forth.
2. **Natural Image.** An image directly captured by a camera with no post processing is a natural image in our context.

Peak Signal-to-Noise Ratio – PSNR. PSNR is derived from the mean square error, and indicates the ratio of the maximum pixel intensity to the power of the distortion [6]. Like MSE, the PSNR metric is simple to calculate but might not align well with perceived quality.

$$PSNR = 10 \log_{10} \frac{MAXI^2}{MSE} \quad (5)$$

Here, $MAXI$ is the maximum possible pixel value of the image. When the pixels are represented using 8 bits per sample, this is 255. For color images with three RGB values per pixel, the definition of PSNR is the same except that the MSE is the sum over all squared value differences (now for each color, i.e. three times as many differences as in a monochrome image) divided by image size and by three. Alternately, for color images the image is converted to a different color space and PSNR is reported against each channel of that color space, e.g., HSV or Lab.

Structural Similarity – SSIM. Structural similarity (SSIM) index. The SSIM metric combines local image structure, luminance, and contrast into a single local quality score [15]. In this metric, structures are patterns of pixel intensities, especially among neighboring pixels, after normalizing for luminance and contrast. Because the human visual system is good at perceiving structure, the SSIM quality metric agrees more closely with the subjective quality score.

The SSIM index is calculated on various windows of an image. The measure between two windows x and y of common size $N \times N$ is:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + c_1)(2\mu_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \quad (6)$$

with:

μ_x – the average of x ; μ_y – the average of y ; σ_x^2 – the variance of x ; σ_y^2 – the variance of y ; σ_{xy} – the covariance of x and y ; $c_1 = (k_1L)^2$, $c_2 = (k_2L)^2$ – two variables to stabilize the division with weak denominator; L – the dynamic range of the pixel-values (typically this is $2^{\#bits \text{ per pixel}} - 1$; $k_1 = 0.01$ and $k_2 = 0.03$ by default).

The SSIM formula is based on three comparison measurements between samples of x and y : luminance (l), contrast (c) and structure (s). The individual comparison functions are:

$$l(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \quad (7)$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \quad (8)$$

$$s(x, y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \quad (9)$$

with, in addition to above definitions: $c_3 = c_2/2$

SSIM is then a weighted combination of those comparative measures:

$$SSIM(x, y) = l(x, y)^\alpha c(x, y)^\beta s(x, y)^\gamma \quad (10)$$

Setting the weights α , β , γ to 1, the formula can be reduced to the form shown above.

To evaluate the image quality, this formula is usually applied only on the luma, although it may also be applied on color (e.g., RGB) values or chromatic (e.g. Lab) values. The resultant SSIM index is a decimal value between 0 and 1, and value 1 is only reachable in the case of two identical sets of data and therefore indicates perfect structural similarity. A value of 0 indicates that there is no structural similarity. For an image, it is typically calculated using a sliding Gaussian window of size 11×11 or a block window of size 8×8 . The window can be displaced pixel by pixel on the image to create an SSIM quality map of the image. In the case of video quality assessment, the authors propose using only a subgroup of possible windows to reduce the complexity of the calculation.

Blind/Referenceless Image Spatial Quality Evaluator – BRISQUE. A BRISQUE model is trained on a database of images with known distortions, and BRISQUE is limited to evaluating the quality of images with the same type of distortion. BRISQUE predicts the score using a support vector regression (SVR) model trained on an image database with the corresponding differential mean opinion score (DMOS) values: 0 - best IQ (high SNR value); 100 - poorest IQ (low SNR value). The database contains images with known distortion, such as

compression artifacts, blurring, and noise, and contains pristine versions of the distorted images. The image to be scored must have at least one of the distortions for which the model was trained [10].

Natural Image Quality Evaluator – NIQE. Although a NIQE model is trained on a database of pristine images, NIQE can measure the quality of images with arbitrary distortion. NIQE is opinion-unaware and does not use subjective quality scores. The trade-off is that the NIQE score of an image might not correlate as well as the BRISQUE score with human perception of quality. NIQE measures the distance between the NSS-based features calculated from image A to the features obtained from an image database used to train the model. The features are modeled as multidimensional Gaussian distributions [11].

Perception Based Image Quality Evaluator – PIQE. The PIQE algorithm is opinion-unaware and unsupervised, which means that it does not require a trained model. PIQE can measure the quality of images with arbitrary distortion and in most cases performs similar to NIQE. PIQE estimates blockwise distortion and measures the local variance of perceptibly distorted blocks to compute the quality score. PIQE calculates the no-reference quality score for an image by using blockwise distortion estimation [14].

2.3 Statistical Analysis

Regression models between full-reference (PSNR, SSIM) and no-reference metrics (BRISQUE, NIQE, PIQE) are estimated using R^2 and Root Mean Square Error (RMSE) [3].

R^2 measures how much variability in the dependent variable can be explained by the model. It is the square of the correlation coefficient (R), and that is why it is called R^2 .

$$R^2 = 1 - \frac{SS_{regression}}{SS_{total}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (11)$$

R^2 is calculated by the sum of squares of the prediction error divided by the total sum of the square that replaces the calculated prediction with the mean value. R^2 value is between 0 and 1 and a bigger value indicates a better fit between the prediction and the actual value.

R^2 is a good measure to determine how well the model fits the dependent variables. However, it does not take into account the overfitting problem. If your regression model has many independent variables, because the model is too complicated, it may fit very well to the training data, but performs poorly for testing data. Therefore, adjusted R^2 is introduced because it will penalize additional independent variables added to the model and adjust the metric to prevent overfitting issues [3].

While R^2 is a relative measure of how well the model fits dependent variables, Mean Square Error (MSE) is an absolute measure of the goodness for the fit.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \tag{12}$$

MSE is calculated by the sum of the square of the prediction error which is the real output minus the predicted output, and then divide by the number of data points. It gives you an absolute number on how much your predicted results deviate from the actual number. You cannot interpret many insights from one single result, but it gives you a real number to compare against other model results and helps you select the best regression model.

The root mean square error (RMSE) is the square root of MSE. It is used more commonly than MSE because firstly the MSE value can sometimes be too big to compare easily. Second, the MSE is calculated by the square of error, and thus the square root brings it back to the same level of prediction error and makes it easier to interpret [3].

3 Results

3.1 Regression Models Quality: SSIM vs. PSNR

Figure 1 illustrates the same regression models (model type, model order, close quality metrics R^2 and RMSE) SSIM vs PSNR for Gaussian white noise (GWN) and salt&pepper noise (SPN).

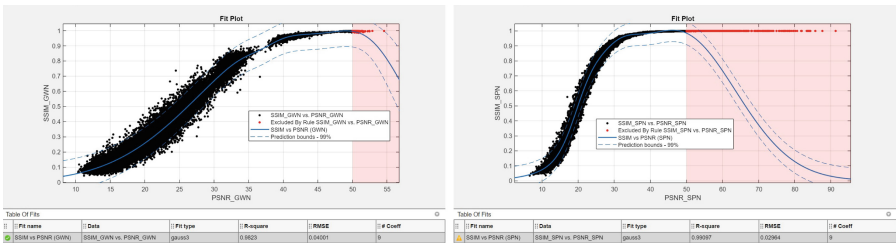


Fig. 1. Regression model SSIM vs. PSNR for GWN (Left) with $R^2 = 0.98$; RMSE = 0.04 and SPN (Right) with $R^2 = 0.99$; RMSE = 0.03

So, we can conclude, that PSNR and SSIM are highly dependent (obviously the model is not linear, the dependency is similar to a sigmoid).

3.2 Regression Model Quality: No-Reference Metrics vs. Full-Reference Metrics

Figure 2 illustrates the same regression models (model type, model order, close quality metrics R^2 and RMSE) BRISQUE vs SNR for GWN and SPN.

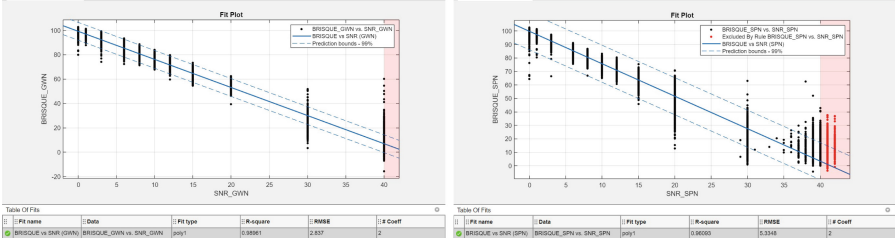


Fig. 2. BRISQUE vs SNR (GWN): $R^2 = 0.99$; RMSE=2.84 (SPN): $R^2 = 0.96$; RMSE=5.33

Note that points in “red zone” (SNR > 40 dB) are not taken into account when building regression models. So, we can see that BRISQUE and SNR are highly dependent and model is strong linear.

Figure 3 illustrates the same (model type, model order, close quality metrics R^2 and RMSE) regression models BRISQUE vs SSIM for GWN and SPN.

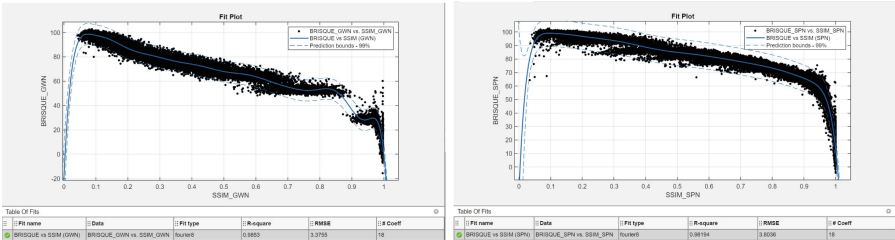


Fig. 3. BRISQUE vs SSIM (GWN): $R^2 = 0.99$; RMSE=3.38 (SPN): $R^2 = 0.98$; RMSE=3.80

So, we can see that BRISQUE and SSIM are highly dependent via complex Fourier model.

Figures 4 illustrates the same (model type and model order) regression models NIQE vs SNR for GWN and SPN.

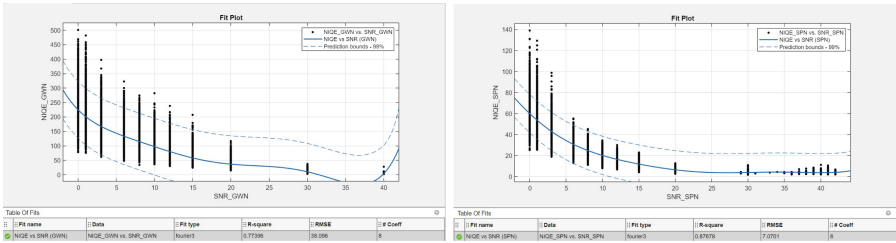


Fig. 4. NIQE vs SNR (GWN): $R^2 = 0.77$; RMSE = 38.01 (SPN): $R^2 = 0.88$; RMSE = 7.07

In this case quality metrics R^2 and RMSE are worse and not close, regression model is similar to exponential.

Figure 5 illustrates the same (model type and model order) regression models NIQE vs SSIM for GWN and SPN.

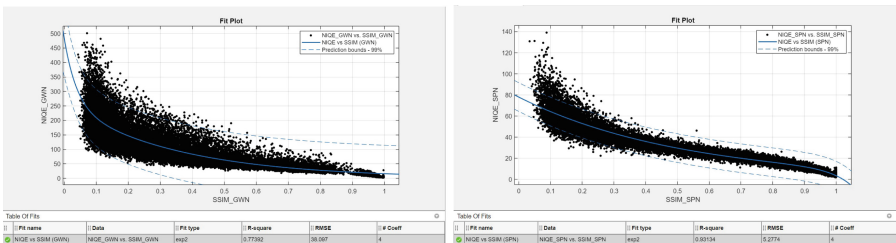


Fig. 5. NIQE vs SSIM (GWN): $R^2 = 0.77$; RMSE = 38.01 (SPN): $R^2 = 0.93$; RMSE = 5.28

In this case quality metrics R^2 and RMSE are worse and not close, regression model is similar to exponential.

Figure 6 illustrates the regression models PIQE vs. SNR for GWN and SPN.

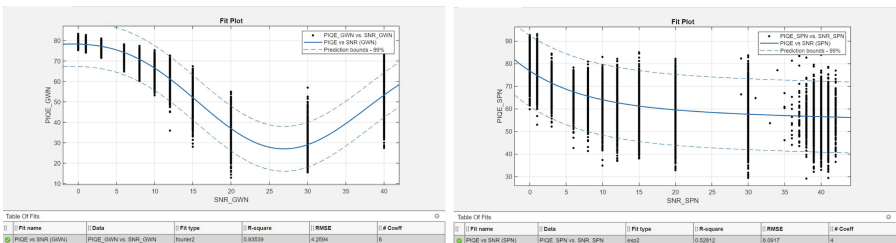


Fig. 6. PIQE vs SNR (GWN): $R^2 = 0.94$; RMSE = 4.26 (SPN): $R^2 = 0.53$; RMSE = 6.09

Here, quality metrics R^2 and RMSE are worse and not close, the regression model is almost random.

Figure 7 illustrates the same (model type and model order) regression models PIQE vs. SSIM for GWN and SPN.

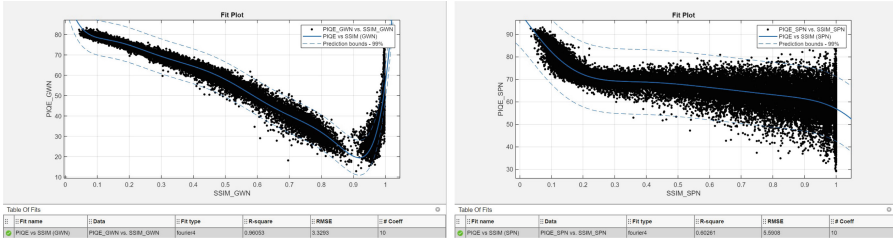


Fig. 7. PIQE vs SSIM (GWN): $R^2 = 0.96$; RMSE=3.33 (SPN): $R^2 = 0.60$; RMSE = 5.59

Here, quality metrics R^2 and RMSE are worse and not close, the regression model is a complex Fourier model.

4 Discussions

In this paper, no reference metrics (BRISQUE, NIQE, PIQE) and full-reference metrics (PSNR, SSIM) are evaluated for face recognition quality assessment. Different noise types are considered: gaussian and salt&pepper noise with SNR in the range 0...40 dB. Regression models between full-reference and no-reference metrics are considered. The quality of the regression models is estimated using R^2 and RMSE.

The results of all the experiments conducted are summarized in Table 1.

Table 1. Regression models quality: no-reference metrics vs full-reference metrics

NR-metric	FR-metric	GWN: R^2 ; RMSE	SPN: R^2 ; RMSE
BRISQUE: opinion-aware; supervised	SNR	0.99; 2.84	0.96; 5.33
BRISQUE: opinion-aware; supervised	SSIM	0.99; 3.38	0.98; 3.80
NIQE: opinion-unaware; supervised	SNR	0.77; 38.01	0.88; 7.07
NIQE: opinion-unaware; supervised	SSIM	0.77; 38.01	0.93; 5.28
PIQE: opinion-unaware; unsupervised	SNR	0.94; 4.26	0.53; 6.09
PIQE: opinion-unaware; unsupervised	SSIM	0.96; 3.33	0.60; 5.59

Regression functions between FR- and NR-metrics have to be non-increasing functions. For FR-metrics: A higher score indicates better perceptual quality

(PSNR, SSIM). For NR-metrics: a lower score indicates better perceptual quality (BRISQUE, NIQE, PIQE). Using BRISQUE and NIQE we obtain non-increasing regression functions, which correspond to the theory. But BRISQUE is much better than NIQE. Sometimes, using PIQE we get better results in R^2 and RMSE metrics, but despite this fact, PIQE is unusable, because regression functions are not non-increasing.

5 Conclusion

From Table 1 and Figs. 2, 3, 4, 5, 6 and 7 we can conclude that BRISQUE is the best solution for NR-IQA. NIQE is significantly inferior to BRISQUE. PIQE is unusable. The experimental results correlate with the theory.

In terms of the task, future work is required. To begin with, it is important to assess how the results obtained during the work will correlate with the results of the accuracy indicators for different FR models, including the evaluation of various image enhancement and denoising algorithms and their impact on the accuracy. Further research in terms of different noise types such as anisotropic noise, periodic noise, quantization noise, etc. is required. It is also important to evaluate the conclusions obtained when using color images (RGB, HSV, Lab), since grayscale images are rarely used in the FR problem. Other approaches for NR-IQA like neural-network-based solutions should be explored, developed, and evaluated. Other image quality factors, other than noise level, such as facial expression, pose, and illumination, should be examined. Finally, pipeline based on described algorithms to improve the quality of FR systems (image discrimination, reducing type I and type II errors while face detection) should be created.

References

1. Boyat, A., Joshi, B.K.: Image denoising using wavelet transform and median filtering. In: 2013 NIRMA University International Conference on Engineering (NUICONE), pp. 1–6 (2013)
2. Boyat, A.K., Joshi, B.K.: A review paper: noise models in digital image processing. arXiv preprint [arXiv:1505.03489](https://arxiv.org/abs/1505.03489) (2015)
3. Chicco, D., Warrens, M.J., Jurman, G.: The coefficient of determination r-squared is more informative than smape, mae, mape, mse and rmse in regression analysis evaluation. *Peer J. Comput. Sci.* **7** (2021)
4. Grm, K., Scheirer, W.J., Štruc, V.: Face hallucination using cascaded super-resolution and identity priors. *IEEE Trans. Image Process.* **29**, 2150–2165 (2019)
5. Grother, P., Hom, A., Ngan, M., Hanaoka, K.: Ongoing face recognition vendor test (frvt) part 5: Face image quality assessment (4th draft). National Institute of Standards and Technology. Technical report 1 (5) (2021)
6. Horé, A., and Ziou, D. (2010). Image quality metrics: Psnr vs. ssim. In: 2010 20th International Conference on Pattern Recognition, pp. 2366–2369 (2010)
7. Huang, G. B., Mattar, M., Lee, H., Learned-Miller, E.: Learning to align from scratch. In: *Nips* (2012)
8. Marmarelis, V.Z.: *Nonlinear Dynamic Modeling of Physiological Systems*, vol. 10. Wiley, Hoboken (2004)

9. Meng, Q., Zhao, S., Huang, Z., Zhou, F.: Magface: a universal representation for face recognition and quality assessment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14225–14234 (2021)
10. Mittal, A., Moorthy, A.K., Bovik, A.C.: Blind/referenceless image spatial quality evaluator. In: 2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR), pp. 723–727 (2011)
11. Mittal, A., Soundararajan, R., Bovik, A.C.: Making a “completely blind” image quality analyzer. *IEEE Signal Process. Lett.* **20**, 209–212 (2013)
12. Schlett, T., Rathgeb, C., Henniger, O., Galbally, J., Fierrez, J., Busch, C.: Face image quality assessment: a literature survey. In: *ACM Computing Surveys (CSUR)* (2020)
13. Song, Y., et al.: Joint face hallucination and deblurring via structure generation and detail enhancement. *Int. J. Comput. Vis.* **127**(6), 785–800 (2019)
14. Venkatanath, N., Praneeth, D., Bh., M.C., Channappayya, S.S., Medasani, S.S.: Blind image quality evaluation using perception based features. in: 2015 Twenty First National Conference on Communications (NCC), pp. 1–6 (2015)
15. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error measurement to structural similarity (2004)



Using Soft Decisions for Error Correction

Tatyana Pavlenko^(✉)  and Oleg Malafey

North-Caucasus Federal University, Stavropol, Russia
misstanya.pavlenko@yandex.ru

Abstract. The article considers a communication channel with erasures, which allows making soft decisions in order to achieve the maximum gain in the fidelity of information reception. A comparative analysis of error correction algorithms using the erasure signal for the most reliable symbols of the same name in a multiple repeated block of information encoded with a redundant code is carried out.

Keywords: Erasure signal · repetition codes · loss of information · noise immunity · single symbol distortion probability

1 Introduction

The issue of ensuring the reliability of information transmission in networks is of great importance. If an error appears in the text during the transmission of a regular telegram or a crackling noise is heard during a telephone conversation, then, as a rule, errors and distortions are detected without problems in meaning. But when transmitting data, one error per thousand transmitted signals can greatly affect the quality of information [7].

Data transmission networks in strategic facility management systems require the quality of information transmission to be no lower than 10^{-9} . In practice, this indicator is approximately equal $10^{-3} - 10^{-4}$, which is unacceptable [2].

There are many methods that ensure the reliability of information transmission, differing in the means used for their implementation, in the time spent on their use at the transmitting and receiving points, in the additional time spent on transmitting a fixed amount of data, in the degree of ensuring the reliability of information transmission. The form of practical implementation of the methods consists of two parts: software and hardware. The ratio between the two can be very different, up to the almost complete absence of one part.

Among numerous error protection methods, three groups of methods are distinguished: group methods, error-correcting coding, and error protection methods in feedback transmission systems.

In this paper, we will consider a communication channel with erasures that allows making soft decisions, as well as a comparative analysis of error correction algorithms using the erasure signal for the most reliable symbols of the same name in a multiple repeated block of information encoded with a redundant code.

2 Correction of Erasures in Codes with Repetition

Problems of the complexity of correction of erasures of large weights were actively studied in 1963–1980. For example, in [5, 6], algorithms and schemes for correcting heavy weight erasures and detecting errors are studied for linear and some cyclic codes. To date, the issue of complexity is not so critical, although it remains important.

Correction of erasures in codes with repetition makes it possible to significantly increase the noise immunity by fairly simple means [3, 9].

The ability of the code to detect and correct errors is due to the presence of redundant elements in the codeword $r = n - k$, where k – the single elements that are contained in the codeword are the n – single elements that are included in the codeword of the redundant code ($n > k$) [1].

On Fig. 1 is a graph illustrating the statistical characteristics of a binary symmetric erasure channel (BSEC), which are determined by three probabilities:

1. The probability of element transformation in the absence of erasure p ;
2. Element erasure probability s ;
3. The probability of correctly receiving an element in the absence of erasures q .

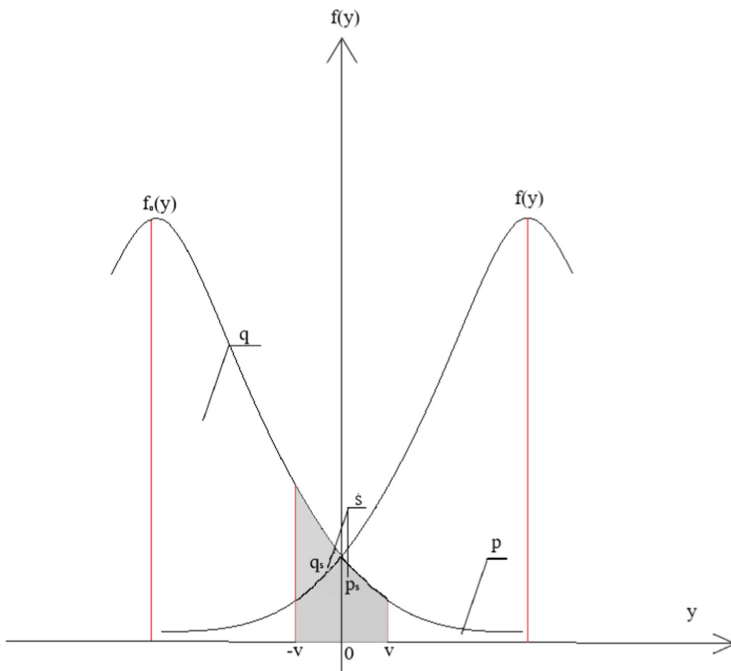


Fig. 1. Symmetric erasure interval

In this case, the condition is always satisfied

$$p + q + s = 1. \tag{1}$$

At the output of binary receivers with a symmetrical erasure interval, random variables are identified with element 0 only when

$$y_i < -v,$$

and with element 1 when

$$y_i > v,$$

where v — the erasure threshold and the distribution $f_0(y)$ is a mirror image of the distribution $f_1(y)$ with respect to zero.

If it turns out that

$$-v < y_i < v,$$

then an erasure symbol is additionally selected Θ , which fixes the fact of the unreliability of this element.

It is obvious that the probability of element transformation p is determined by the formula

$$p = \int_{-\infty}^{-v} f_1(y)dy = \int_v^{\infty} f_0(y)dy, \quad (2)$$

and the probability of erasure is given by the formula

$$s = \int_{-v}^v f_0(y)dy = \int_{-v}^v f_1(y)dy. \quad (3)$$

From Fig. 1 and expressions (2) and (3) it is obvious that

$$s = p_s + q_s;$$

$$p_0 = p + p_s; \quad (4)$$

$$q_0 = q + q_s;$$

where p_s — is the probability of correct erasure of the element, the q_s — probability of false erasure, p_0 — the probability of element distortion, the q_0 — probability of correct reception of the element.

Erasures, which Θ , mark reliable elements of the received codeword, can be used to improve the noise immunity of reception in the same way as a posteriori error probabilities are used P_j in optimal and quasi-optimal methods. There are three methods for processing redundant codes using erasures [6].

In accordance with the first method, the code combination is erased if at least one element is erased or if an error is detected by the code. This method belongs to the adaptive reception methods, which indirectly take into account the quality of the used

communication channel. It allows you to provide a high value of fidelity, which is achieved by increasing the loss of information [4].

The second method differs from the first only in that the redundant code is used to correct errors. Its efficiency is not high [8].

The third method is used to correct erasures and detect errors in non-erased positions. Erasure of the code combination is carried out either in the case when the number of erased elements is greater than $d - 1$ (where the d – code distance), or when the code detects errors on non-erased positions. This method and some of its modifications are close enough to the optimal decoding methods [12].

If there is a double repetition of a message encoded with redundant (n, k) – code, then if an error is found in the first repetition, it is advisable to remember it (block \vec{Y}_1). When receiving the second repetition (block), the corresponding \vec{Y}_2 erasures are fixed $\vec{\Theta}$ and the result of modulo two addition ($\vec{\Omega}$) of the block elements of the same name (\vec{Y}_1 and \vec{Y}_2) is determined. The result of logical multiplication $\vec{\Theta}$ and $\vec{\Omega}$

$$\vec{e} = \vec{\Theta} \vec{\Omega} \tag{5}$$

with a high probability indicates distorted elements of the second repetition, which are inverted in accordance with \vec{e} . Adjusted Combination

$$\vec{X} = \vec{Y}_2 \boxplus \vec{e} \tag{6}$$

is subjected to code verification and, if there are no errors, is issued for further processing.

Figure 2 shows a timing diagram illustrating the procedure for correcting a single error in the second iteration.

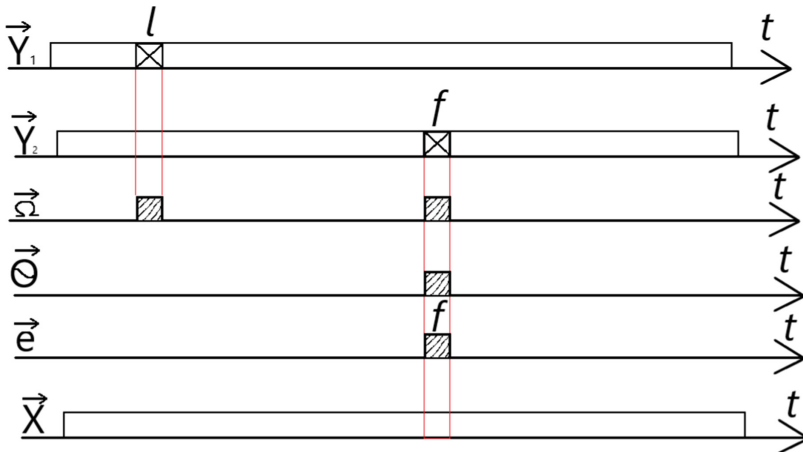


Fig. 2. Timing diagram illustrating the correction of a single error

Incorrect correction of elements in the second repetition will occur in those cases when false erasure will correspond to the same-named non-matching elements. If the

total number of such corrections, together with distortions at non-erased positions of the second repetition, exceeds the detecting ability of the code, then an undetectable error will occur.

The probability of undetected errors can be calculated by the formula

$$P_{\text{HO}}(s, n) \approx \frac{1}{2^r} \sum_{i=d}^n C_n^i p_0^i (1-p_0)^{n-i} + \frac{1}{2^r} \sum_{i=1}^{d-1} \sum_{j=d-i}^n C_n^i p_0^i q_s^i (1-p_0)^{n-i} C_{n-i}^j p^j (1-p)^{n-(i+j)}. \quad (7)$$

We obtain an approximate expression if we select the constituent terms that have the greatest weight:

$$P_{\text{HO}}(s, n) \approx \frac{1}{2^r} C_n^d p_0^d + \frac{1}{2^r} C_{n-1}^{d-1} n p_0 q_s p^{d-1}. \quad (8)$$

The probability of detected errors, which characterizes the loss of information when receiving the second repetition, can be calculated by the formula

$$P_{\text{OO}}(s, n) = \sum_{i=1}^{d-1} \sum_{j=0}^{d-1-i} C_n^i C_{n-i}^j p_0^i q_s^j p^j (1-p_0)^{n-i} (1-p)^{n-(i+j)}, \quad (9)$$

or approximately according to the formula

$$P_{\text{OO}}(s, n) \approx n p_0 q_s. \quad (10)$$

The technical implementation of this method is considered in [11].

3 Combination of Indirect and Code Methods of Error Detection

Extensive experimental studies of the efficiency of corrective codes in real radio channels have shown that in the error correction mode at, the $n \leq 511$ codes increase the reliability by no more than one order of magnitude, while the losses due to code redundancy reach 50–80%. Calculations show that with an increase in the length of the combination to 64–128 single symbols, the loss in speed decreases to 6–10%, and the coefficient of fidelity increase increases by two orders of magnitude [3].

In data transmission systems, in order to achieve the maximum gain in reception fidelity, several code and several indirect error detection methods can be used simultaneously. Combined is such a principle of error detection, in which the coordination of error detection methods becomes of particular importance [4].

Consider the probability space of errors in reception. Errors have a different nature: $0 \rightarrow 1$ or $1 \rightarrow 0$, different multiplicity and random nature of the distribution within the received code combination. Let P — the probability of erroneous reception of single symbols. The use of one or another method of error detection ensures the identification of some part of them. Let us designate the space of elementary error events Ω as, and its elements as $\omega_1, \omega_2, \dots$. Therefore, $\Omega = \{\omega_i, i = 1, 2, 3, \dots N\}$, where N — is the number of elements in Ω . Let the errors detected by the first method be the number A_1 , which is a subset of Ω , i.e. $A_1 \in \Omega$. Then

$$P(A_1) = \sum_{\omega_i \in A_1} P(\omega_i), \tag{11}$$

and the probability of undetectable errors

$$P_{HO} = P - P(A_1). \tag{12}$$

If we use a combination of two error detection methods (indirect and code), the probability of joint error detection will be determined by the sum of events A_1 and A_2 , i.e. union of subsets $A_1 \cup A_2$ and intersection $A_1 \cap A_2 = \emptyset$, where

$$P(A_1 \cup A_2) = \sum_{\omega_i \in A_1 \cap A_2} P(\omega_i) = P(A_1) + P(A_2) - P(A_1 \cap A_2). \tag{13}$$

With the simultaneous use of several methods (indirect and code), the probability of detecting errors

$$P\left(\bigcup_{i=1}^{\alpha} A_i\right) = \sum_{i=1}^{\alpha} P(A_i) - \sum_{i<1}^{\alpha} P(A_i A_j) + \sum_{i<j<\gamma} P(A_i A_j A_\gamma) - \dots + (-1)^{\alpha-1} P\left(\bigcap_{i=1}^{\alpha} A_i\right). \tag{14}$$

Then the probability of undetectable errors

$$P_{HO} = P - P\left(\bigcap_{i=1}^{\alpha} A_i\right). \tag{15}$$

With such combined error control, detection methods should duplicate less and complement each other more, i.e. the probabilities of intersections of the set A_i should tend to zero [5].

By introducing a quality detector into the PDI system, along with code methods for protecting information, it is possible to reduce the cost of APD. In addition, taking into account the parameters of distortion packetization allows, for example, using a signal quality detector with optimal erasure zones, to reduce the probability of an undetected error by two to three orders of magnitude, to reduce the amount of transmission rate loss in systems with overdemand, to reduce code redundancy [6].

Correction of erasures in codes with repetition can significantly increase the noise immunity by fairly simple means by implementing the following method.

A N – multiple repetition of messages encoded (n, k) – by the code is carried out, where $\{m - 1 < N < 2m - 1, m = 3, 4, \dots\}$. If there is a double repetition of the message, if an error is found in the first repetition, (\vec{Y}_1) it is advisable to remember it. When receiving the second repetition, (\vec{Y}_2) the erasures corresponding to it are fixed $(\vec{\Theta})$, and the result of modulo two addition of the same symbols of the same name is determined \vec{Y}_1 and \vec{Y}_2 , $\vec{\psi} \vec{E} = \vec{\Theta} \vec{\psi}$ with a higher probability indicates distorted symbols of the second repetition, which are inverted in accordance with \vec{E} . The corrected combination $\vec{X} = \vec{Y}_2 \oplus \vec{E}$ is subjected to a code check and, in the absence of errors, is issued for further processing, and together with the first one \vec{Y}_1 , a code for the number of units in the same symbols of two repetitions is formulated.

$$R = \sum_{i=1}^2 Y_i. \quad (16)$$

Incorrect correction of elements in the second repetition occurs in those cases when the same name mismatched characters correspond to a false erasure. If the total number of such corrections, together with distortions at non-erased positions of the second repetition, exceeds the detecting ability of the code ($\sigma = d - 1$), then an undetectable error will occur. The probability of undetected errors is determined by the relation

$$P_{\text{HO}}(s, n) \cong \frac{1}{2^{n-n_s}} \left[\sum_{i=d}^n C_n^i P_o^i (1 - P)^{n-i} + \sum_{i=1}^{d-1} \sum_{j=d-1}^n C_n^i P_0^i q_s^j (1 - P_0)^{n-i} C_{n-i}^j P^j (1 - P_0)^{n-(i+j)} \right]. \quad (17)$$

If we select the terms that have the greatest weight, then

$$P_{\text{HO}}(s, n) \cong \frac{1}{2^{n-n_s}} \left[C_n^d P_o^d + C_{n-1}^{d-1} n P_0 q_s P^{d-1} \right]. \quad (18)$$

The probability of error detection, which characterizes the loss of information when receiving the second repetition, we find by the formula

$$P_{\text{OO}}(s, n) \cong \sum_{i=1}^{d-1} \sum_{j=0}^{d-1-i} C_n^i C_{n-i}^j P_0^i q_s^j (1 - P)^{n-i} (1 - P)^{n-(i+j)} \quad (19)$$

or approximately $P_{\text{OO}}(s, n) \cong n P_0 q_s$, where j – is the number of characters to be erased; i – number of errors on non-erased positions; P_0 – probability of distortion of a single character; P – the probability of character transformation in the absence of erasures; q_s – probability of false erasure; n – the number of characters in the received combination; d – code distance [8].

When the second repetition does not satisfy the fidelity condition, subsequent $N - 2$ repetitions are received with the correction of unreliable elements, the location of which

determines the erasure signal Θ if the code for the number of ones in the same symbols of the previous group of repetitions has a maximum or minimum value

$$R_{N-1} = \max\left(\sum_{i=1}^{N-1} \vec{Y}_i\right) \cup \min\left(\sum_{i=1}^{N-1} \vec{Y}_i\right). \tag{20}$$

With a limited number of memory elements ρ , each of which has n cells, the maximum code for the number of ones in the same symbols is limited by the number of repetitions $R_N = 2^\rho - 1$ and determines the possibility of erasure correction. Therefore, with the further reception of the next repetition, the correction effect does not take place. But with this restriction, the majority processing of received repetitions of the message is possible with the formation of the voting result by the majority, determined by the rule

$$\beta_{m(2m-1)} = \begin{cases} 1, & \text{if } \sum_{i=1}^{2m-1} \vec{Y}_i \geq m \\ 0, & \text{if } \sum_{i=1}^{2m-1} \vec{Y}_i < m, \text{ где } m = 2^\rho - 1 \end{cases}. \tag{21}$$

The use of a signal that takes into account the quality of the communication channel Θ for the correction of single errors in the symbols of the same name of the second and all subsequent ones, up to and $2^\rho - 1$ including repetitions of the message, makes it possible to increase noise immunity. This can be shown by the example of determining the probability of information loss by comparing the considered method and the method of adaptive majority decoding [3].

4 Discussion

If in a known device the probability of distortion of a single character in the final code combination is estimated by the value

$$P_{\text{э1}} = mP_0^m,$$

where m — is the odd number of repetitions exponent $N = 2m - 1, m = 3$;
 P_0 — probability of distortion of a single symbol of a code combination;
 then in the proposed device:

$$P_{\text{э2}} = (2m - 1)P_0^{m+1}P_T,$$

where P_T — is the probability of symbol transformation in the absence of Θ .
 In this case, the loss of information can be estimated by the expression:

$$P_{\Pi} = 1 - (1 - P_{\text{э}})^n n P_{\text{э}}.$$

Then if

$$P_0 = 10^{-2}(\text{low quality channel}),$$

$$P_T = 2 \cdot 10^{-2},$$

$$m = 3$$

$$\text{then } P_{\Pi 1} = n(2m - 1)P_0^{m+1},$$

$$\text{a } P_{\Pi 2} = n(2m - 1)P_0^{m+1}P_T.$$

Therefore, the loss of information is reduced in

$$\eta = \frac{P_{\Pi 1}}{P_{\Pi 2}} = \frac{1}{P_T} = \frac{1}{2 \cdot 10^{-2}} = 50.$$

Thus, when receiving five repetitions of the message, the device implements an extended set of decision rules. At the same time, taking into account the “erasure” signal Θ makes it possible to correct a certain proportion of errors in the repetition for the most reliable preceding symbols, which reduces the total number of errors in the received message. This facilitates error correction with an extended set of decision rules, thereby increasing noise immunity [12].

Consider the case of receiving seven repetitions and analyze the degree of information loss reduction.

The probability of distortion of a single character in the final codeword obtained as a result of the majority processing of $2m - 1$ repetitions is determined by expression $P_{\rho}(m) \cong C_{2m-1}^m P_0^m$, and in the combined method that combines the demodulation and decoding procedure, $P'_{\rho}(m) \cong C_{2(m-1)}^{m-1} P_0 P^{m-1}$. Therefore, the loss of information for each of the compared methods can be found using the expressions

$$P_n = nC_{2m-1}^m P_0^m$$

$$P'_n = nC_{2(m-1)}^{m-1} P_0 P^{m-1} \quad (22)$$

The ratio of these values will determine the degree of information loss reduction

$$\eta = \frac{P_n}{P'_n} = \frac{C_{2m-1}^m}{C_{2(m-1)}^{m-1}} P_0^{m-1} P^{1-m} \quad (23)$$

If we take specific values $P_0 = 10^{-2}$, $P = 10^{-3}$, $m = 2$, which corresponds to the presence of seven repetitions $N = 7$, $\rho = 3$, then.

$$\eta = 1,75 \cdot 10^3.$$

The nature of the change in the loss probability from the number of repetitions parameter for the two considered algorithms (Fig. 3) shows that the new algorithm has improved probabilistic-temporal characteristics and can be used in radio communication systems with poor quality communication channels [12].

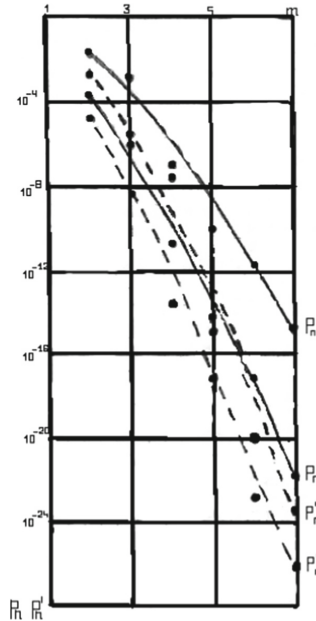


Fig. 3. Dependences of the loss probability P_{Π} on the number of repetitions m for various $P_0 = 10^{-1}$ and $P_0 = 10^{-2}$

5 Conclusion

Currently, the corrective ability of LDPC codes in the erasure channel is being actively investigated. There is a long list of works in this area. For example, in [2] the correction of erasures by composition of LDPC codes with a Hamming code is studied, while deriving and using (in a concatenated decoding scheme) exact estimates of the proportion of correctable erasures of a large weight for a binary Hamming code. We would also like to mention the work [4], where the fraction of erasures corrected by non-binary codes is considered. To estimate the proportion of correctable erasure combinations, the results related to the study of minimal words in linear codes [12] are useful.

References





1. Cheung, K.M.: The weight distribution and randomness of linear codes. Jet Propulsion Lab., California Inst. of Tech., Pasadena, CA. TDA Progress Report 42-97. USA. 1989, pp. 208-215. <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19890018521.pdf>. Accessed 22 Nov 2016
2. Qiao, F., Zhu, P., Pan, L., Liu, H., Zhang, Z., Xu, J.: New FN/BBHH Combined Erase Scheme for Scalable SONOS Device. In: 2014 International Symposium on Next Generation Electronics (ISNE), pp. 1-2 (2014). 10.1109 / ISNE.2014.6839330
3. Lue, H., et al.: Investigation of the charge loss mechanism of SONOS type devices using hot hole erasure and charge retention improvement techniques. In: 2006 IEEE International Reliability Physics Symposium, pp. 523-529 (2006). <https://doi.org/10.1109/RELPHY.2006.251273.J>

4. Chen, G. Xie , K. Luo, W. Cheng, P. Lu , and Y. Wang, “ Research characteristics stripes erasing and records in magnetic records With replacement conjugated composite carriers. In: 2018 IEEE International Magnetism Conference (INTERMAG), pp. 1–1 (2018). <https://doi.org/10.1109/INTMAG.2018.8508564>
5. Weight distribution. <http://www.ec.okayama-u.ac.jp/~infsys/kusaka/wd/index.html>. Accessed 27 Nov 2016
6. Akimov, P.S.: Signals and their processing in the information system. Akimov P. S., Senin A. I., Solenov V. I. - M. Radio and communication (1994). 256 p.
7. Werner, M.: Fundamentals of coding. Textbook for universities. Per. with him. M.: Technosphere (2005). 288 S
8. Davydov, A.A., Tombak, L.M.: On the number of words of minimum weight in block codes. *Prob. Inf. Transmiss.* **24**(1), 11–24 (1988)
9. Zyablov, V.V., Rybin, P.S.: Correction erasures codes With small density checks. *Probl. Inf. Transm.* **45**(3), 15–32 (2009)
10. Klyuchko, V.I., et al.: Device for correcting errors in a code combination. Klyuchko V.I. - Copyright certificate No. 634469, BI No. 43 (1978)
11. Malofey O.P. Error correction device with an extended set of decision rules and an erasure signal. Malofey O.P., Malofey A.O. et al. Patent No. 2208907 dated 07/20/2003
12. Pavlenko, T.A.: Ways to reduce the probability of information loss in infocommunication systems with poor quality communication channels. In: Pavlenko, T.A., Malofey, O.P. (eds.) *Natural sciences - the basis of the present and the foundation for the future. Materials of the VIIIth annual scientific and practical conference of the North Caucasian Federal University “University science for the region” - Stavropol: NCFU (2020). - 600–604 p.*

Actual Problems of Mathematical Education



Interactive Methods in the Study of the Discipline “Mathematics” for Non-mathematical Specialties

Irina Lavrinenko^{1,2} , Natalia Semenova^{1,2} , and Valentina Baboshina²  

¹ Department of Mathematical Modeling, Faculty of Mathematics and Information Technologies, North-Caucasus Federal University, Stavropol, Russia

² North-Caucasus Center for Mathematical Research, North-Caucasus Federal University, Stavropol, Russia

valentina03012000@gmail.com

Abstract. The paper considers some aspects of the use of interactive methods in the study of the discipline “Mathematics” for non-mathematical specialties at the North Caucasus Federal University. The discipline “Mathematics” is included in the block of the mandatory part of the undergraduate curriculum in the natural sciences, and interactive teaching methods are designed to help students in the study of a non-core discipline. The use of interactive technologies contributes to the implementation of the competence-based approach. The paper gives example of the use of interactive methods in the practical classes of training groups of training areas 03.19.01 - Biotechnology and 05.30.01 - Medical Biochemistry, which contribute to mastering the modern mathematical apparatus for further use in solving theoretical and applied tasks in professional activity. A questionnaire was conducted among students of these areas in order to determine the effectiveness of the interactive teaching methods used. The results of the interview were analyzed. It was revealed that the use of interactive methods in the learning process helps to increase motivation and develop independence, stimulates an independent search for new knowledge, improves student performance and makes it possible to apply mathematical knowledge, skills and abilities in practical research.

Keywords: Interactive Methods · Competence-based Approach · Increased Motivation · Non-mathematical Specialties · Teaching Methods

1 Introduction

In modern society, knowledge and the level of intellectual development of a person are becoming the most important strategic resources, raising the social status of education and making ever higher demands on its level and quality. High-quality training of a graduate provides not only the availability of fundamental knowledge in special areas but also the formation of a number of general professional competencies, such as the ability to search and analyze professional literature, conduct experimental research, and process and interpret empirically obtained data. The student must have a certain amount of

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023

A. Alikhanov et al. (Eds.): APAMCS 2022, LNNS 702, pp. 489–499, 2023.

https://doi.org/10.1007/978-3-031-34127-4_47

knowledge and be a creatively active person. These requirements are reflected in the Federal State Educational Standards of Higher Education for the discipline “Mathematics” [1].

It should be noted that students entering non-mathematical areas do not always have the proper level of school mathematical training for high-quality mastery of the university program in the discipline “Mathematics”, and not all students at the beginning of the educational process can see real prospects for the application of mathematical methods in their future professional career [2, 3]. During the forced transition to distance learning in educational institutions during the coronavirus pandemic, it was difficult to implement the individual trajectory of schoolchildren and effectively organize the interaction of students with each other and with the teacher [4]. These factors force teachers to consider and analyze various teaching methods in order to choose such methods that help increase students’ motivation to study higher mathematics, develop independence, initiative and creativity in solving problems of various kinds [5, 6].

Thus, the teacher of the university is faced with the question of the correct organization of the educational process. For the implementation of the tasks set, it seems to be a priority to use interactive methods to create conditions for more active interaction between students. Interactive teaching methods have relatively recently entered the arsenal of teaching methods used in higher education [7–9]. Figure 1 presents the basic principles of interactive teaching methods.

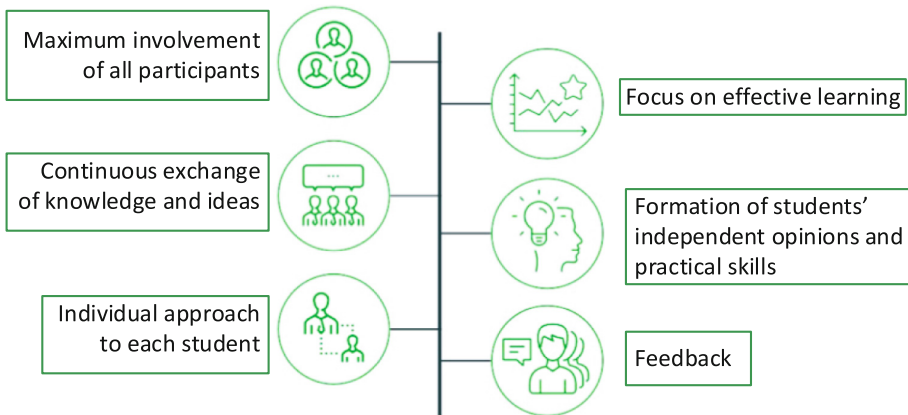


Fig. 1. Basic principles of interactive teaching methods

Modern pedagogical research has identified factors for the effectiveness of the use of interactive learning, among which are: the development of cognitive and mental activity; involvement of students in the process of cognition as active participants; increased motivation to study the discipline; creating a favorable, creative atmosphere in the classroom; development of communicative competences of students; developing the skills of an independent search for information and determining the level of its reliability, which is especially important when reducing the share of traditional classroom work and increasing the amount of independent work in curricula [10].

In a number of works [11–14], devoted to the use of interactive technologies, a wide range of interactive approaches have been developed, among which are methods applicable to the discipline “Mathematics” in practical classes: business and role-playing games, brainstorming, round tables and discussions, situational analysis, work in small groups, etc. The purpose of this work is to demonstrate the increase in the effectiveness of teaching students of non-mathematical specialties when the teacher uses interactive teaching methods.

2 Materials and Methods

2.1 Methods Used in Classes with Students

Let us consider the methods used in the classes on the discipline “Mathematics” for the preparation of students of the profile 19.03.01 - Biotechnology and 30.05.01 - Medical Biochemistry at the North Caucasus Federal University.

2.1.1 «Tic-Tac-Toe» Game

Assumes the use of an interactive whiteboard. It is held at the first practical lesson in order to increase the motivation of students to study higher mathematics and create a favorable, creative atmosphere in the future in the learning process. One group of students selects nine questions that confirm the need to study mathematics. Selected questions will be presented on the playing field. Another group is preparing a presentation. To conduct the event itself, two teams and support groups are formed from among the students who are not involved in the preparation. The playing teams take turns choosing a cell on the playing field, trying to place their three crosses or zeroes vertically, horizontally or diagonally, which happens only when the team answers the chosen question correctly. This is not always possible, and, as a rule, there is a complete analysis of all the questions by the students who make up the assignments. The teacher helps in the selection of tasks in the preparation process, recommending literature, sums up the game, arguing his conclusions.

2.1.2 «Math Casino» Game

The game is played before studying the topic “Fundamentals of Probability Theory”. The game is based on the task “Who took what?” from the book by Ya. I. Perelman “Live Mathematics” [15], where chips are offered instead of nuts. On the playing field (Fig. 2(a)) there are 24 chips and three items. The leader is the teacher. The conditions of the game are as follows: three students can play, the first is given one chip, the second is given two chips, and the third is given three. Three players are asked to choose items “a”, “b”, or “c”. At the moment of choice, the leader turns away. For example, “a” is a car, “b” is a country house, “c” is a computer. Everyone must take more chips, the owner of the item “a” - as much as he was given; subject “b” - twice the number of chips that he was awarded; object “c” - four times more than the number of chips that was handed to him. The rest of the chips remain untouched. When all this is done, the host announces who took what, quickly assessing the number of untouched chips. Students are invited

to explain this “trick”. If this does not happen, then the teacher motivates students to study the topic “Fundamentals of Probability Theory” by finding a solution. Figure 2(b) provides a clue to this mathematical “trick”.

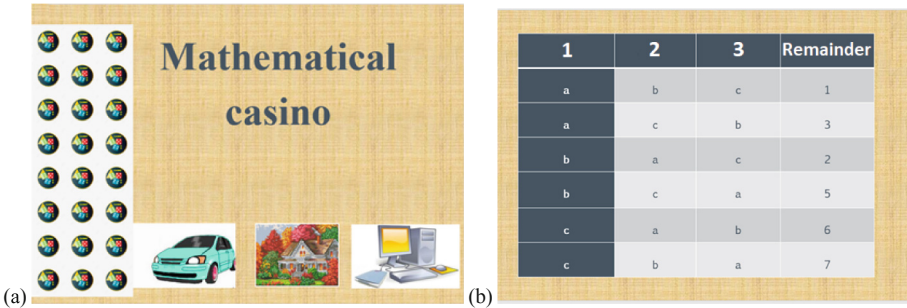


Fig. 2. Playing field on the interactive whiteboard screen

If the game is played in a face-to-face format, as objects “a”, “b”, and “c” you can use a key, a phone, a watch, but if you have an interactive whiteboard, you can use presentation slides and simulate chip selection by crossing them out with an electronic stylus included with the electronic board. In the context of the coronavirus pandemic, there is an experience of playing this game in a distance learning format on the BigBlueButton video conferencing and distance learning platform.

These games enable you to activate the learning process and transfer it from passive to active. They provide an opportunity to form the motivation to study the material at the initial stage. The game “Tic-Tac-Toe” let you to assess the degree of mastery of the material at the stage of completion of the course.

2.2 Discussion Methodology

On topics that allow group discussion of the issue, we conduct discussions that form a solution that suits all participants in the educational process. The method of organizing and conducting discussions includes three stages: preparatory, discussion and final.

2.2.1 Topic “Solving Systems of Linear Algebraic Equations”

Stage I: three groups of students are invited to prepare material on solving systems of equations by Cramer, Jordano-Gauss methods, in a matrix way.

Stage II: each group of students, in the process of presenting the material, reasonably proves the advantage of their method; students of other groups ask questions, clarify points that cause difficulties; the teacher, watching the discussion, evaluates the theoretical background, students’ activity, creativity and the validity of their conclusions.

Stage III: students are offered individual tasks for solving systems of equations by three methods. At the end of the lesson, the results are summarized, as a result of which the conclusion is formed that the Jordan-Gauss method is applied to any systems, other methods under certain conditions.

2.2.2 Topic “Integration of Rational Functions”

Stage I: three groups of students are invited to prepare material on three methods of decomposing a fraction into the sum of simple fractions: indefinite coefficients, private values, and combined; the teacher determines the fraction expansion method for each group.

Stage II: each group sets out the essence of the proposed method, argues its application for decomposing a fraction into a sum of the simplest ones; students ask questions during the discussion; the teacher evaluates the activity of students, the degree of assimilation of the material, reasoning.

Stage III: students in groups are asked to integrate a rational function, presenting it as the sum of an integer part and a proper fraction, expanding it into a sum of simple fractions in the most appropriate way.

When summing up, students come to the conclusion that when integrating fractions, it is necessary to approach rationally the use of a certain method of expanding a fraction into the sum of simple fractions of types I-III, which can simplify the solution.

2.3 Case Technology

Case technologies include: the method of situational analysis, situational tasks and exercises, analysis of specific situations (case study), the method of cases, etc. The method of cases is the study of a fictional or real situation to identify problems, effective solutions and the possibility of practical application of knowledge gained. To develop the skills of solving problems on voluminous topics, we use case-based individual homework. For example, in the first lesson on the topic “Fundamentals of Linear Algebra”, each student receives his own specific system of four linear equations with four unknowns, practicing the basic methods for solving problems on the topics of determinants, matrices and systems of linear equations, according to the following plan:

1. Compose the matrix A from the coefficients of the unknowns of the system of linear equations.
2. Find the matrix A determinant:
 - a) using properties and definitions,
 - b) decomposition by elements of some row or column,
 - c) after receiving zeros in any row or column.
3. Compare the results obtained in task 2 and indicate the most rational calculation method.
4. Solve a system of linear equations using Cramer’s formulas, calculating all the necessary determinants using a rational method.
5. Find the matrix inverse for A :
 - a) using the attached matrix,
 - b) elementary transformations.
6. Compare the results obtained in task 5, and indicate the advantages and disadvantages of the applied methods for finding the inverse matrix.
7. Write down the system of linear equations in the form of a matrix equation.
8. Solve a system of linear equations in matrix form.
9. Compare the results obtained in tasks 4 and 8.

10. Solve the system of linear equations by successive elimination of unknowns.
11. Compare the results obtained in tasks 4, 8 and 10.
12. Change one of the equations of the system so that it becomes inconsistent. Describe how it was done, justifying the loss of solutions.
13. Solve the system of linear equations (task 12) by successive elimination of unknowns.
14. Find a general and any particular solution of systems of linear equations, consisting of:
 - a) first three equations of the initial system;
 - b) first two equations of the initial system;
 - c) the first equation of the initial system.

The teacher can use such individual homework assignments completed by the student for interviews in the final lessons in the sections.

2.4 Small Group Work

Provides for the division of the group of students into small teams to discuss certain issues and develop solutions to the educational problem. This method enable you to involve all students in the work, trains the skills of cooperation and interpersonal communication.

2.4.1 Topic “Differentiation Technique”

When developing differentiation skills, the teacher emphasizes the need to know by heart the table of derivatives, with the help of which the differentiation of frequently occurring functions is reduced to mechanical procedures, but there are a number of functions for which a special differentiation technique is used. The teacher demonstrates by examples the differentiation of functions given implicitly and parametrically, as well as finding derivatives of exponential functions using logarithmic differentiation.

Students are divided into groups of 4; when forming small groups, it is necessary to take into account the observance of the principle of heterogeneity, since training in heterogeneous groups allows weak students to catch up to the level of average students and, at the same time, stimulates the learning process of medium and strong ones. Each group receives an option that provides the same type of tasks for each student, but of varying degrees of difficulty, depending on the level of preparation. While performing tasks, students have the opportunity to consult stronger ones to help those who experience difficulty. Groups report on their work. The teacher sums up, making comments and arguing his conclusions.

2.4.2 Topic “Direct Integration”

Initial integration skills are associated with the so-called direct integration, covering the use of the table of integrals, the properties of integrals and some elements of transformations that bring the integral to the form of any tabular integral. Direct calculation of integrals using a table is relatively rare. Therefore, for the calculation of integrals, a number of techniques have been developed that allow reducing this integral to tabular ones. Mastering these techniques allows students to find integrals more successfully.

The teacher talks about bringing the integral to a tabular one using the “bringing under the sign of the differential” technique. A number of examples are worked out on the board, groups are given handouts with tasks. In the variant, there are several propaedeutic tasks for restoring the differential, as well as a number of integrals reduced to tabular ones using the method being studied (Table 1).

When studying the next topic “Methods of integration” at the stage of forming the skill of integration by substitution, we note that quite often this technique can be replaced by “bringing under the differential sign”, which simplifies the solution. And here, it should be noted that, in contrast to the general methods of differentiation, which are applied almost mechanically, integration requires great skill - in each individual case, one must be able to find a suitable technique and apply it in the most advantageous way (Table 2).

Table 1. An example of a task for a group of students on the topic “Direct Integration”.

	1 st member	2 nd member	3 rd member	4 th member
Exercise № 1	$d(\quad) = 2x dx$	$d(\quad) = x^3 dx$	$d(\quad) = \cos x dx$	$d(\quad) = \sin x dx$
Exercise № 2	$d(\quad) = \frac{dx}{1+x^2}$	$d(\quad) = \frac{dx}{x}$	$d(\quad) = \frac{dx}{\cos^2 x}$	$d(\quad) = \frac{dx}{\sin^2 x}$
Exercise № 3	$\int \sin 3x dx$	$\int e^{-2x} dx$	$\int \frac{dx}{x-10}$	$\int (x^2 + 1)^4 x dx$
Exercise № 4	$\int \sqrt{x^2 + 1} x dx$	$\int e^{\cos x} \sin x dx$	$\int \frac{x^2 dx}{\sqrt[3]{1+x^3}}$	$\int \frac{\sqrt{1+\ln x}}{x} dx$

Table 2. An example of a task for a group of students on the topic “Linear Differential Equations of the First Order”.

	Exercise № 1	Exercise № 2
1 st member	$y' - 4x^3 y = 0$	$y' - \frac{y}{x} + 1 = 0$
2 nd member	$xy' = 3 - y$	$y' - \frac{3y}{x} = 2x^4$
3 rd member	$y' - 4x^3 y = 0$	$y' + \frac{4y}{x} = x$
4 th member	$y' - y = 0$	$y' - \frac{3y}{x} = x$

Table 3. Questionnaire results, percentage of positive answers

Question	19.03.01 – Biotechnology	30.05.01 – Medical biochemistry
1	80%	78%
2	77%	70%
3	73%	74%
4	70%	78%
5	77%	78%
6	77%	70%
7	80%	78%

2.4.3 Topic “Linear Differential Equations of the First Order”

During the frontal survey, the teacher repeats the algorithm for solving first-order linear differential equations using the Bernoulli substitution method. It clearly demonstrates that when solving an equation with separable variables, which are obtained in the process of transformations in order to obtain a more convenient form of writing the solution, one can take not C , but $\ln C$, which does not violate the generality of reasoning, and also when integrating expressions $\int \frac{du}{u}$, it is permissible to omit the sign of the modulus, which as a result of arbitrariness C will not change the result of integration. The teacher gives handouts with assignments, in which, as propaedeutic assignments, there are equations with separable variables.

Students are divided into groups; each group performs its own version. Group’s report on the work done by displaying their results on the screen of the interactive whiteboard. The principle of working in small groups is described in the first topic. The teacher sums up, making comments and arguing his conclusions.

When describing the teaching methods used, it should be noted that in almost every of them, it is possible to use interactive whiteboards. When conducting practical or lecture classes, it is permissible to combine all the advantages of a classical presentation with the possibilities of high technologies. As classes conducted using interactive whiteboards become more interesting and richer, the level of assimilation of the material increases.

3 Results

To evaluate the effectiveness of interactive teaching methods, a questionnaire was conducted among students of the directions 19.03.01 - Biotechnology and 30.05.01 - Medical biochemistry. In the groups of direction 19.03.01 classes were conducted using not only traditional approaches but also with the systematic use of active and interactive methods, and in the group of direction 30.05.01 the use of interactive methods was episodic. It should be noted that the level of school mathematical training in these areas was somewhat different, for students of the direction 05.30.01 - Medical biochemistry, in general, it was lower, this is evidenced even by the fact that when entering a university, the results

of the Unified State Examination in mathematics were optional for them. The students were asked the following questions.

1. Do interactive teaching methods make mathematics classes fun and interesting?

- yes;
- no.

2. Does the use of interactive learning improve your level of knowledge in the subject?

- yes;
- no.

3. Do interactive methods contribute to the active involvement of students in the communicative (communication) process?

- yes;
- no.

4. Are interactive teaching methods aimed at the active interaction of all participants in the educational process?

- yes;
- no.

5. Do interactive teaching methods contribute to introspection (reflection) of their knowledge?

- yes;
- no.

6. Do interactive teaching methods influence the formation of positive learning motivation?

- yes;
- no.

7. Is the use of interactive teaching methods a promising direction in the educational process?

- yes;
- no.

Let's present the results of the questionnaire. For each of the questions, the number of positive and negative answers was counted. The ratio of “yes” answers, expressed as a percentage, is presented in Table 3. The diagram in Fig. 3 presents the results clearly.

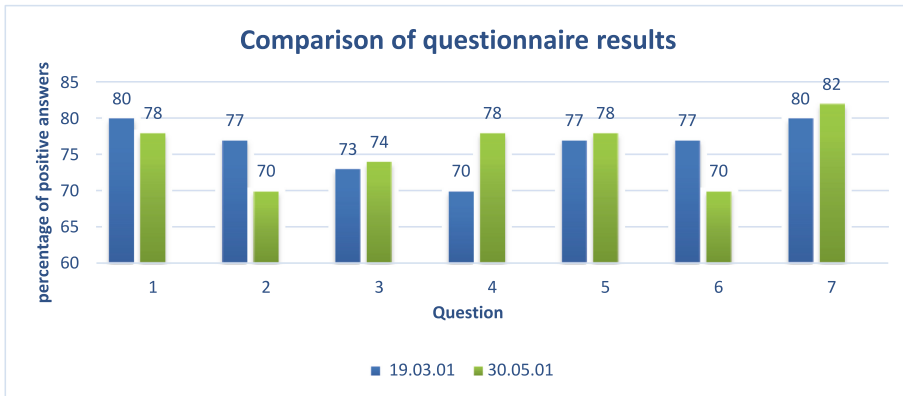


Fig. 3. Comparison of questionnaire results

4 Discussion

A high percentage of “yes” answers to question 1, almost the same in the two studied groups (80% and 78%), showed that interactive teaching methods make it possible to turn mathematics classes into an exciting process even for students of non-mathematical specialties and with different levels of preparation. The use of interactive methods, according to students, contributes to an increase in the level of preparation in the subject, but a lower percentage of “yes” answers (question 2 - 77% and 70%) among students of the second group, from our point of view, is due to the episodic use of interactive technologies. The students of the studied groups take almost the same position regarding the influence of interactive methods for increasing communicative activity (question 3 - 73% and 74%). The answers to question 4 showed that the participants in the educational process actively interact in the educational process between themselves and the teacher, and a higher percentage in the second group (70% and 78%) indicates that the students of the second study group assessed that mutual assistance, the interaction of strong and weak students in small groups, positively affects the performance of all students.

Interactive teaching methods contribute to an objective assessment of one’s own knowledge in the process of active joint activity, this is noted by students of both groups (question 5 - 77% and 78%), and this, in turn, stimulates the process of an independent search for new knowledge, thereby contributing to the development of the motivational sphere learning activities (question 6 - 77% and 70%). According to students, the use of interactive methods is a promising direction in the educational process (question 7 - 80% and 82%).

5 Conclusion

The results of the questionnaire, as well as the monitoring of knowledge under current control, show that the use of interactive methods in the process of mathematical training of students in non-mathematical areas increases the motivation to study the subject, develops the cognitive activity of students at all stages of acquiring knowledge, and

contributes to the formation of skills for applying and analyzing problems in professional activities. The use of interactive technologies creates a favorable, creative atmosphere in the classroom, develops the communicative competencies of students, which, of course, improves the quality of training of future specialists.






Acknowledgments. The work is supported by North-Caucasus Center for Mathematical Research under agreement № 075–02–2022–892 with the Ministry of Science and Higher Education of the Russian Federation.

References

1. Order of the Ministry of Science and Higher Education of the Russian Federation dated August 10, no. 736: On approval of the federal state educational standard of higher education - bachelor's degree in the field of study 03.19.01 Biotechnology (2021)
2. Weurlander, M., Cronhjort, M., Filipsson, L.: Engineering students' experiences of interactive teaching in calculus. *Higher Educ. Res. Develop.* **36**(4), 852–865 (2017)
3. Giorgdze, M., Dgebuadze, M.: Interactive teaching methods: challenges and perspectives. *Int. E-J. Adv. Educ. (IJAEDU)* **3**(9), 544–548 (2017)
4. Abykanova, B., Sadirbekova, D., Sardarova, Z., Khairzhanova, A.K., Mustagaliyeva, G.S.: Interactive teaching methods as pedagogical innovation. *Biosci. Biotechnol. Res. Commun.* **14**(05), 171–175 (2021)
5. Abykanova, B., et al.: The use of modern information technologies in the educational process. *Ad. Alta J. Interdiscip. Res.* **10**(1), 37–40 (2020)
6. Gushchin, Y.V.: Interactive teaching methods in higher education. *Psychol. J. Int. Univ. Nat. Soc. Man Dubna* **2**, 1–18 (2012)
7. Yakovleva, N.O., Yakovlev, E.V.: Interactive teaching methods in contemporary higher education. *Pacific Sci. Rev.* **16**(2), 75–80 (2014)
8. Norin, V.A., Norina, N.V., Pukharenko, Y.V.: Interactive methods of teaching at Russian engineering universities. *Educ. Inf. Technol.* **23**, 2801–2820 (2018)
9. Vinogradova, M.V., Yakobyuk, L.I., Zenina, N.V.: Interactive teaching as an effective method of pedagogical interaction. *Espacios* **39**(30), 15–17 (2018)
10. Zikirova, N., Abdullayeva, N., Nishanova, O., Djalilov, B., Nishanbayeva, E.: Interactive Strategies and Methods of Education. *Int. J. Recent Technol. Eng. (IJRTE)*, **8**(4), 2277–3878 (2019)
11. Senthamarai, S.: Interactive teaching strategies. *J. Appl. Adv. Res.* **3**(S1), 36 (2018)
12. Norin, V.A., Norina, N.V., Pukharenko, Y.V.: Interactive methods of teaching at Russian engineering universities. *Educ. Inf. Technol.* **23**, 2801–2820 (2018)
13. Masran, S.H., Marian, M.F., Yunus, F.A., Rahim, M.B., Baser, J.A.: Effectiveness of using an interactive media in teaching and learning: a case study. In: 2017 IEEE 9th International Conference on Engineering Education (ICEED) (2017)
14. Lim, W.N.: Improving student engagement in higher education through mobile-based interactive teaching model using Socrative. In: 2017 IEEE Global Engineering Education Conference (EDUCON) (2017)
15. Perelman, I.: *Jivaya matematika [Live mathematics]. Mathematical Stories and Puzzles* (2020)



Applied Mathematics and Informatics Bachelor's and Master's Educational Programs Continuity During Updating Educational Standards

Irina Zhuravleva¹ , Ludmila Andrukhiv^{1,2} , Elena Yartseva¹ ,
and Valentina Baboshina²  

¹ Department of Mathematical Modeling, Faculty of Mathematics and Information Technologies, North-Caucasus Federal University, Stavropol, Russia

² North-Caucasus Center for Mathematical Research, North-Caucasus Federal University, Stavropol, Russia

valentina03012000@gmail.com

Abstract. The education system in the modern world is built from the steps - junior school, high school, college, university (bachelor's, master's, postgraduate). Teachers and psychologists are concerned about the continuity of these levels of education for the education of employed professionals. The paper is devoted to the analysis of the short-term continuity of the educational programs of undergraduate and graduate programs in applied mathematics and computer science in the context of updating the survey. The implementation has been studied of the mechanisms of succession of professional and research components of the training of students of the directions 01.03.02 - "Computational mathematics and mathematical modeling" and 01.04.02 - "Mathematical modeling" in the educational practice of the North Caucasus Federal University. Curricula, the general vectors of training are considered, examples of reporting in areas (final theses and exams, papers in scientific journals) are given. A mechanism is proposed for the formation of a set of professional competencies for bachelor's and master's degree graduates in the specialty "Mathematical Modeling" correlated with the requirements of professional standards. The most productive forms of continuity of the research component of educational programs are identified in the field of applied mathematics and computer science at NCFU.

Keywords: Continuity · Applied Mathematics and Informatics · Professional Standards · Labor Functions · Professional Competencies

1 Introduction

Improving the system of continuous higher education involves the emergence of new educational models [1]. The problem of the continuity of purpose educational programs is becoming increasingly important, which is aimed at overcoming the gap between different levels of education in modern conditions of the implementation of a two-level

system of higher education (bachelor and master) [2]. Many scientists are still engaged in the research on various aspects of continuity in education. The influence is being studied of distance learning on students [3]. The continuity of school and university programs [4], as well as continuous education as an individual educational strategy [5], is considered. Examples are given of the influence of gap year on young people [6] entering a university. Samples are shown of the formation by teachers of the continuity of educational programs [7]. Among the most significant components of continuity, in the field of vocational education are the continuity of learning objectives and the content of the main educational programs; continuity of pedagogical technologies used in the educational process; continuity of knowledge, abilities, skills and competencies formed as a result of mastering educational programs of different levels [8]. By continuity in higher education, we refer to the process of ensuring a continuous relationship between the levels (bachelor's degree—master's degree). It contributes to the formation of universal, general professional and professional competencies of students and improving the professional and personal development of graduates.

Training in the area of Applied Mathematics and Informatics (AMI) is focused mainly on mathematical modeling. Mathematical modeling—an ideal scientific symbolic formal modeling, in which the description of an object is carried out in the language of mathematics, and the study of the model is carried out using certain mathematical methods [9]. North Caucasus Federal University (NCFU) is preparing in the following areas in the field of applied mathematics and informatics: 01.03.02 (bachelor's degree), area of training—“Computational mathematics and mathematical modeling”, and 01.04.02 (master's degree), area of training - “Mathematical modeling”. Currently, in these areas of training, there are not even projects of exemplary educational programs (EP), which can be recommended by the Federal Educational and Methodological Association. A review of modern psychological and pedagogical literature on continuity in higher education and analysis of regulatory documents for the development of EPs in higher education allows us to note that by now theoretical knowledge has been accumulated on the issues of continuity of two levels: bachelor and master. However, the practical implementation of the continuity of the EP in the field of applied mathematics and informatics (AMI) is not sufficiently represented. The main aim of the work is to analyze the complexity of mechanisms for implementing the continuity of EP based on the disclosure of the essence of the professional and research component of student training.

2 Applied Mathematics and Informatics Bachelor's and Master's Educational Programs Continuity

The research was conducted on the basis of the Faculty of Mathematics and Computer Science named after Professor N.I. Chervyakov of NCFU. Theoretical and empirical methods were used, such as analysis, systematization and generalization of psychological and pedagogical literature and regulations, observation. EPs were developed and analyzed in the areas of training 01.03.02 Applied Mathematics and Informatics (bachelor), area of training - “Computational Mathematics and Mathematical Modeling”, and 01.04.02 Applied Mathematics and Informatics (Master), area of training - “Mathematical Modeling» on the basis of theoretical understanding and generalization of literary

sources and regulatory documents. In the current updated version of the Federal State Educational Standards of higher education (FSES HE 3 + +) of the 3rd generation [10, 11] it is normatively fixed for educational organizations that they can develop EP on their own and form requirements for the results of its development in the form of universal, general professional and professional competencies of graduates. Competence is understood as a feature consisting of knowledge, understanding and actions [12]. Competencies are the result of mastering the main educational program and consist of a set of knowledge, skills, experience, personal and professional qualities of a graduate.

Universal competencies (UC) characterize the over-professional abilities of the individual, which ensure the successful activity of a person in a wide variety of both professional and social spheres. The list of UCs is one in terms of levels of education for all areas and specialties. UCs are of a supra-subject nature; therefore, their formation occurs regardless of the specific academic discipline of the educational program throughout the entire period of study, using various forms of organizing the educational process. Scientific position is substantiated in a collective monograph on the problems of measuring and evaluating UC, which is that universal competencies (soft skills) are formed not instead of, but together with professional competencies (hard skills) and constitute a unified context of socially significant educational results for the digital economy [13]. Let us consider the composition of groups of general professional competencies (GPC) of AMI graduates. Continuity can be traced here in the name of the category, as well as in the content aspect. At the same time, for bachelor's programs, GPCs are presented more widely in quantitative terms, and for master's programs—in qualitative terms. So, if in GPC-2 a bachelor is “able to use and adapt existing mathematical methods and programming systems to develop and implement algorithms for solving applied problems” [10], then a master in this case “is able to improve and implement new mathematical methods for solving applied problems” [11].

Professional competencies (PC) are described independently by the university. They are the most closely related to the current requirements of the labor market [14]. PCs are formed in accordance with the types of activities to which the educational program is oriented, and on the basis of professional standards (PS) corresponding to the professional activities (PA) of graduates. The development of competencies is preceded by an analysis of experience and consultations with leading employers and/or associations of employers of those industries in which graduates are in demand. The choice of professional standards is carried out by the university on the basis of an application to the FSES using the PS register posted on the specialized website of the Ministry of Labor and Social Protection of the Russian Federation “Professional Standards” [15]. If the PS has been changed or has become invalid during the period of the FSES or new PSs have appeared that more closely correspond to the area of training of the EP, it becomes necessary to choose independently another PS, possibly absent in the annex to the FSES. The selection of a suitable PS from the registry is based on the main areas and main goals of each type of PD.

The non-profit Association of Computer and Information Technology Enterprises is created to represent the interests of the Russian IT industry in the domestic and foreign markets and it recommends universities actively use methodological developments and educational modules of Russian IT companies when developing educational programs

in the field of IT, as well as use certain tools for monitoring the quality of training in IT areas, taking into account the current practice in the development of professional standards [15]. The practical implementation of the continuity of the EP is complicated by the fact that there are no exemplary basic educational programs in the field of AMI. Certain mechanisms have been adopted in the educational practice of NCFU to ensure the continuity of the professional component of student training, taking into account this circumstance.

3 Updating Educational Standards

A university determines independently the areas of professional activity when forming an educational program and selects the types of tasks of professional activity on the basis of the FSES, determines professional tasks within the framework of the selected type and determines the objects of activity and the focus of the profile. This choice determines the selection of professional standards associated with the program. At the same time, the choice of specific PS for the same training area of training at different levels of training differs markedly. For example, the PC of bachelors of the direction 01.03.02 in NCFU is formed on the basis of the PS: 01.003 Teacher of additional education for children and adults, 06.001 Programmer, 06.016 Project manager in the field of information technology and 06.022 System analyst. PCs installed by the master's program of direction 01.04.02 can be developed on the basis of completely different PS, for example: 40.011 Specialist in research and development and 06.028 System programmer. In accordance with the Order of the Ministry of Labor of Russia dated April 12, 2013 No. 148n, only those labor functions are selected from the PS, the implementation of which requires an education level corresponding to the Federal State Educational Standard. Thus, only those parts of the PS are selected that can be used in the program in accordance with the level of qualification and requirements for education and training: labor functions, labor actions, knowledge and skills that can be formed in the educational process. The formulation of successive professional competencies was carried out on the basis of the tasks of the professional activity of graduates, as well as the generalized labor functions of the selected PS.

Here are the sources of the formation of PC-1 related to research activities, as well as indicators of its achievement, formulated on the basis of correlation with the selected professional standards. For the bachelor 01.03.02, area of training "Computational Mathematics and Mathematical Modeling", the tasks of professional activity are: the use in research and applied activities of modern mathematical apparatus, modern programming languages and software, modern information and communication technologies, achievements of science and technology in applied mathematics and informatics, operating systems and network technologies; collection, processing, and interpretation of experimental data necessary for project activities in the field of information technology.

Corresponding labor functions: formalization and algorithmization of tasks; collection and processing of design research results.

PC-1 formulation: The ability to mathematically correctly formulate natural science problems and investigate them using computational methods.

PC-1 achievement indicators are formulated on the basis of correlation with professional standards 001 - “Programmer”, 016 - “IT Project Manager”, 022 - “System Analyst”:

1. PC-1.1 Knows the methods of mathematical formulation of natural science problems is to understand the limits of applicability and requirements for the computational algorithm.
2. PC-1.2 Able to mathematically correctly set natural science problems in various fields, to study them by computational methods.
3. PC-1.3 Owns the techniques of building mathematical models, is able to choose the methodology for solving the problem and use modern tools for the implementation of computational methods.

Let us describe the sources of the formation of PC-1 for the magistracy of the direction 01.04.02, area of training “Mathematical Modeling”. The tasks of professional activity are: the research of new scientific results, scientific literature or research projects in the field of applied mathematics and informatics in accordance with the subject of ongoing research; compilation of scientific reviews, abstracts and bibliography, preparation of scientific and scientific-technical publications on the subject of ongoing research; formation of research programs in new directions; application of fundamental knowledge obtained in the field of mathematical and (or) natural sciences; creation, analysis and implementation of new computer models in modern natural science, technology, economics and management; construction of mathematical models and their processing by analytical methods; development of algorithms, methods, software, tools on the subject of ongoing research projects; study of systems by methods of mathematical forecasting and system analysis; development and application of modern high-performance computing technologies, the use of modern supercomputers in ongoing research.

Relevant labor functions: implementation of scientific management of research on individual tasks.

PC-1 wording: The ability to conduct scientific research and obtain new scientific and applied results independently and as part of a research team based on existing methods in a particular area of professional activity.

Indicators of achievement of PC-1 are formulated on the basis of correlation with professional standards 40.011 “Specialist in research and development” and 06.028 “System programmer”:

1. PC-1.1 Demonstrates fundamental knowledge in the field of mathematics and natural sciences.
2. PC-1.2 Builds mathematical models and investigates them by analytical and numerical methods.
3. PC-1.3 Develops algorithms, methods, software and tools on the subject of ongoing research projects.

It can be seen that one PC is formulated from several sources.

When choosing professional standards that correspond to the area of training and the professional tasks for which students will prepare, the university may encounter a situation where the area of training does not correspond to any of the PS from the annex

to the FSES or the registry. It is proposed to use different PSs that contain similar labor functions in such situations. In the order of the Ministry of Labor of Russia dated April 12, 2013 N 148n “On the approval of qualification levels for the development of draft professional standards”, nine qualification levels were identified. From the sixth level, higher education is required. In real practice, the developments of EP HE can be used by PS with suitable labor functions that do not correspond to terms of qualifications. For example, PS 06.001 “Programmer” does not imply the need for higher education, as it describes third and fourth skill levels corresponding to secondary vocational education. Thus, there is a lack of a clear correlation between the FSES HE and PS, and the developers of educational programs have to take this into account.

Let’s compare the labor functions of bachelor and master graduates in terms of the continuity of authority and responsibility. It is easy to see that the sixth level, corresponding to the preparation of a bachelor, implies certain powers and responsibilities in the implementation of the labor functions: independent activity, involving the definition of tasks for one’s own work and/or subordinates to achieve the goal; ensuring the interaction of employees and related departments and responsibility for the result of work at the level of a department or organization. For the seventh level, corresponding to the preparation of the master, it is typical to expand the powers to determine the strategy, manage processes and activities, including innovative ones, with decision-making at the level of large organizations or departments, as well as responsibility for the performance of large organizations or departments.

4 Discussion

The leading type of professional activity for the academic area “Applied Mathematics and Informatics” at all levels of training is research. Let’s consider the implementation of the mechanisms of continuity of the research component of student training, which has developed in the educational practice of NCFU. Obtaining the primary skills of research work (RW) is part of the mandatory part of the educational and industrial practice of the EP of the bachelor and master programs. Research practice programs coincide in terms of the logic of content and the sequence of stages and tasks. Educational practice is carried out at the graduating department of mathematical modeling, in university departments with the necessary human and scientific and technical potential. RW is part of work practice, which is carried out over several semesters and ensures the formation and development of the student’s professional competencies necessary for conducting independent or group RW.

A report on research work may include the preparation and defense of a term paper, writing a final qualifying work (FQR) as a graduate exam, preparing and presenting a report at a scientific and scientific-practical conference at a regional or international level. The result of research for a graduate is not only the development of the necessary competencies but also a published scientific work in the form of an FQR. At the master’s level, the FQR is presented as independent scientific research in the form of a master’s thesis. The difference between the bachelor’s and master’s degree programs lies in the level and depth of the problem under study, as well as in the requirements for the obtained scientific results. Therefore, the continuity of research activities at the undergraduate

and graduate levels primarily concerns the increase in the complexity of the process of working on the FQR.

The approach to personnel support of undergraduate and graduate programs coincides. At least 5% of teachers must be attracted from third-party organizations according to the area of training and at least 60% of the total number of those involved in teaching must have an academic degree. Among other common forms of continuity of the research component of training, one can single out scientific Olympiads and competitions, participation in master classes, a school for a young scientist, scientific conferences, participation in joint scientific collections, project activities that require special scientific research. A review of the materials of the official websites of a number of leading Russian universities, as well as the websites of subject Olympiads, conferences, scientific collections and student scientific forums, showed that not only young scientists and masters but also undergraduate students take part in various forms of RW. Olympiads of various profiles are a link between educational and extracurricular scientific activities at students of all levels. For example, bachelors and masters are admitted to the annual international All-Russian Student Olympiad in mathematics. In addition to the classical form of holding Olympiads, today there are also innovative design and research competitions, design competitions and tournaments, case Olympiads and hackathons.

Interestingly, the preparation of students for the Olympiad can be considered as a mechanism for implementing the continuity of education through the technology of master classes. Such master classes are useful for both undergraduate and graduate students, who can consolidate their professional orientation, get a steady interest in core subjects and, ultimately, can contribute to the formation of a professional subject position. In our opinion, the most productive form of the NCFU practice is the involvement of students in research, development and innovation work and projects. Students are involved in the work of the section "Problems of modular arithmetic and neurocomputer technologies in application to info communication systems" within the framework of the annual university conference "University Science for the Region". Conferences of various levels attract not only undergraduates but also senior undergraduate students. Students regularly participate in scientific seminars and master classes conducted by graduate students, attend face-to-face lectures designed to deepen students' knowledge of the topic of the seminar. The purpose of such seminars can be exclusively research activities, often not directly related to practice in a particular area of professional activity.

The participation of students in the work of the educational and scientific center for computational mathematics of parallel programming on supercomputers, in the activities of the research laboratory "Mathematical Modeling" and North Caucasus Center for Mathematical Research, has broader goals. Undergraduates, graduate students and university professors participate in the work of such communities. Their activities are related to scientific research in the field of mathematics and related areas, focused on solving practical professional problems.

NCFU students annually participate in the implementation of the "Startup as a Diploma" program, and also become winners in the "Umnik" contest. "Umnik" is a large-scale program that allows you to find and support talented young people in Russia. The best participants receive grants for the development of research projects. In the context of studying the mechanisms of succession, these programs are interesting because they are

focused not only on research activities in general but also on developing the ability and readiness of students to carry out the practical aspects of their research in the conditions of the modern labor market.

5 Conclusion

The paper describes the mechanism for the formation of a set of successive professional competencies for graduates of undergraduate and graduate programs in “Applied Mathematics and Informatics”, formulated by the authors in accordance with the types of activities that the educational program is focused on, as well as taking into account the requirements for graduates in the market work, wishes of the leading employers in the region, domestic and foreign experience. The most interesting form of implementing the mechanisms for the continuity of the research component of training in the field of applied mathematics and informatics in the educational practice of NCFU is the involvement of students in research, development and innovation work and projects, which increases the competitiveness of graduates in the modern labor market. The research of the current practice of preparing students in Applied Mathematics and Computer Science in the educational practice of NCFU allows us to conclude that in the context of updating educational standards, the master’s program in the mode of continuity solves the problem of deepening the professional component of the bachelor’s degree program, while at the same time orienting undergraduates to the development of research skills.

Acknowledgments. The work is supported by North-Caucasus Center for Mathematical Research under agreement № 075–02–2022–892 with the Ministry of Science and Higher Education of the Russian Federation.







References

1. Schraube, J.E., Osterkamp, U.: Standpoint of the subject: selected writings of Klaus Holzkamp. 1st edn. Palgrave Macmillan (2013)
2. Cristea, G.: The Psychological Pedagogy Paradigm *Procedia. Social and Behavioral Sciences* **76**, 233–236 (2013)
3. Wu, L., Liu, Q., Zhou, W., Mao, G., Huang, J., Huang, H.: A Semantic web-based recommendation framework of educational resources in e-learning. *Technol. Knowl. Learn.* **25**(4), 811–833 (2018). <https://doi.org/10.1007/s10758-018-9395-7>
4. Nampota, D., Thompson, J.: Curriculum continuity and school to university transition: Science and technology programmes in Malawi. *J. Comparative Educ.* **38**(2), 1–16 (2008)
5. Zeer, E.F., Zavodchikov, D.P., Zinnatova, M.V., Lebedeva, E.V.: Individual’naya obrazovatel’naya traektoriya kak ustanovka sub"ekta v sisteme nepreryvnogo obrazovaniya [Individual educational trajectory as intention of subject in continuing education system]. *Nauchnyi Dialog* **1**, 266–279 (2008)
6. Day, T.: Academic continuity: staying true to teaching values and objectives in the face of course interruptions. *Teach. Learn. Inquiry: ISSOTL J.* **3**(1), 75–89 (2015)
7. Regehr, C., Nelson, S., Hildyard, A.: Academic continuity planning in higher education. *J. Bus Contin. Emer. Plan* **11**(1), 73–84 (2017)

8. Copenheaver, C.A., Peterson, J.A., DeBose, K.G.: Discipline continuity across undergraduate and graduate degrees. *Nat. Sci. Educ.* **42**, 131–136 (2013)
9. Mityushev, V., Nawalaniec, W., Rylko, N.: Introduction to mathematical modeling and computer simulations. Taylor & Francis (2018)
10. Federal State Educational Standard of Higher Education in Russia - Bachelor's degree in the area of study 01.03.02 Applied Mathematics and Informatics
11. Federal State Educational Standard of Higher Education in Russia - Master's program in the area of study 01.03.02 Applied Mathematics and Informatics
12. Kohler, J.: Europäische Qualifikationsrahmen und seine Bedeutung für die Einzelstaatlichen Studiensysteme [European Qualifications Framework for Lifelong Learning]. *Qualität in Studium und Lehre*, pp. 1–26. Stuttgart (2008)
13. Obedkova, L.P., Efremov, A.A., Sekerin, V.D., Gorokhova, A.E., Slepov, V.A.: Formation of competencies in higher education by bachelors and masters. *Utopía y Praxis Latinoamericana* **25**(5), 215–220 (2020)
14. Kuzenkov, O.A., Zakharova, I.V.: Mathematical programs modernization based on Russian and international standards. *Mod. Inf. Technol. IT-Educ.* **14**(1), 233–244 (2018)
15. Registry of professional standards Homepage. <https://profstandart.rosmintrud.ru/pdf>. Accessed 17 June 2022



Designing Computer Simulators in Support of a Propaedeutic Course on Neural Network Technologies

Andrej Ionisyan , Irina Zhuravleva  , Irina Lavrinenko ,
Aleksej Shaposhnikov , and Violetta Liutova 

Department of Mathematical Modeling, Faculty of Mathematics and Information Technologies,
North-Caucasus Federal University, Stavropol, Russia
izhuravleva@ncfu.ru

Abstract. The article is devoted to the development of computer simulators of artificial neural networks for conducting practical classes and organizing control in the “Neurocomputer technologies” subject for schoolchildren of 8-11th grades, who are studying in the system of additional education. The development and application technology of software modules is described for propaedeutics of mastering the mathematical apparatus, which is necessary for further study of the theory of artificial neural networks in the relevant courses of higher professional education. The program provides an opportunity to study at the level of medium or increased complexity for highly motivated students to study natural science disciplines. In this case, special attention will be paid to the practice of working with mathematical models of artificial neural networks and their implementation on a computer. Further professional education of “highly motivated” gifted children is highly likely to be associated with scientific research, which, regardless of the field of human knowledge, is currently based on the use of artificial neural network, neuromathematics and neuropackages. Since the main tasks solved with the help of neuro-mathematics are related to the search for hidden patterns, classification, forecasting, dimension reduction, the scope of their application covers all scientific areas.

Keywords: software · modeling · neural networks · additional education

1 Introduction

Neural networks are a class of models based on a biological analogy with the human brain. They exist for solving a variety of data analysis tasks after passing the training stage on the available data. An artificial neural network (ANN) is a software or hardware embodiment of such a model, built on the principle of organization and functioning of biological neural networks [1]. Currently, the ANN technology is being actively developed and used in the world. It is designed to facilitate the solution of poorly formalized tasks of automating the classification process, forecasting, recognition process, decision-making process; control tasks, encoding and decoding information; approximation of

dependencies, etc. Neural network programming specialists are in demand in the labor market.

In Russia, this technology is not considered in computer science lessons in secondary schools of general educational, lyceums and gymnasiums due to its complexity. However, the propaedeutic acquaintance with neurocomputer technologies in the system of additional education of children can be held on the material available and familiar to schoolchildren of 8-11th grades. The result of such a propaedeutic course may be the appearance of a general idea of neurocomputer technologies among school graduates. It also may be a systematic vision of the subject and an understanding of logical connections in the field of artificial intelligence working with large data.

The authors' program of the propaedeutic course "Neurocomputer technologies" is a program for teaching high school students mathematical informatics through full-time. During the program developing, the mandatory minimum students' knowledge on relevant topics and the requirements for the level of students training in mathematics and computer science were taken into account. The course is aimed at familiarizing students of 8-10th grades with the sections of mathematics, which are important parts of the programmer's professional knowledge. But they are not sufficiently covered in the secondary school curriculum; mastering the modern apparatus of elementary mathematics theory by students; presenting modern approaches to the study of ANN and solving typical problems that arise in the development of neural networks.

A distinctive feature of the program of the propaedeutic course is the creation of conditions for the development and support of gifted children, including assistance to them in vocational orientation and continuing education. In addition, this program provides skills among students for the formation of self-education, since practical work presupposes the mastery of students' ability to work with reference literature and use the Internet resources. The skills developed in this course related to the theory of elementary mathematics, algorithms and reasoning systems are used in other components of mathematical and information education, and education in general.

The program is designed for 36 h, 12 h of which are allocated for the study of theory, 12 h are for practical work and 12 h are for the control. The course material is available to students with different initial levels of mathematical training, however, it is assumed that students have basic arithmetic knowledge and have a desire to learn the course materials.

2 Problem Statement

The most effective technologies for the analytical study of practical problems which do not have a universally recognized solution algorithm are the use of neural networks. Neural network technologies include work on training a neural network on specially selected examples. The main function of training a neural network that reproduces the work of the brain and associative thinking is recognition, the ability to determine similarities and differences. At the stage of ANN training, the basic relationships between input parameters are formed, which are formed into invisible tables (images), subsequently used in doing tasks on the network.

Software that simulates the operation of a neural network is called a neurosimulator or a neuropackage. There is a huge variety of neural packages, the possibility of using neural

networks is also included in almost all known statistical packages. Modern software used for INS modeling involves the use of a complex interface and a large number of different procedures that require a clear understanding of algorithms for working with mathematical packages [2].

For the propaedeutics of studying neurocomputer technologies, affordable and easy-to-learn software is needed that meets the requirements of supporting classical neural network models, the availability of data analysis tools, as well as the requirements of accessibility, reliability, visibility and the availability of a Russian-language interface.

The criteria for comparing neuropackages can be the ease of use, the visibility of the information presented, the ability to use different structures, the speed of work and the availability of documentation. The choice is determined by the qualifications and requirements of the user. Most neuropackages assume a certain sequence of actions e.g. creating a network (user selection of parameters or approval of default settings), then training the network and giving the user a solution.

Analysis of existing solutions allowing to design, train and use neural networks for practical purposes, showed that they can be divided into several groups. They are universal and specialized neural network packages, add-ons for application computing packages, as well as presented software simulators.

Universal neural network packages, which are stable in operation, have specialized functionality and interface. They are commercial or research, which are often free. However, universal packages have too many functions and a complex English-language interface that is difficult for students to master. In addition, commercial packages often cost a lot of money.

Among the universal products, NeuroSolutions [3] can be distinguished, which has powerful functionality that allows you to build neural networks of any complexity. It supports a large number of information input and data processing mechanisms. It is a good network designer. In addition, it is stable and supports expansion through user modules. The disadvantages of Neuro Solutions include high system requirements, high cost and lack of a Russian-language interface and help, which complicates the development of this product.

A powerful graphical editor and neural network simulator that supports neural networks of various architectures of any size is a MemBrain [4]. Like NeuroSolutions, it supports a large number of information input and data processing mechanisms. It is stable and supports expansion through script support. MemBrain is free for educational purposes, but has a complex and non-Russified interface, which makes it difficult to master this product.

Stuttgart Neural Network Simulator (SNNS) is a neural network simulator developed at the University of Stuttgart [5]. SNNS supports classical neural network paradigms and learning algorithms. Its functionality can be expanded through user modules, the source codes of this program are available for study and modification.

One of the first domestic neural network simulators is NeuroPro 0.25 (developed by V. G. Tsaregorodtsev, Institute of Computational Modeling SB RAS), the last freely distributed version of which was released in 1998 [6]. It is able to build multilayer neural networks with a sigmoid activation function. It provides several algorithms for training networks, various methods of automatic optimization of the neural network.

The program has a Russian-language interface and help. The use of NeuroPro in the educational process is hindered by a low level of visibility, a user-friendly interface, the lack of convenient means of data entry and the lack of the possibility of expanding the functionality of the program.

Deductor 5.2 is an analytical platform developed by BaseGroup Labs [7]. The support of neural network models in it plays the role of one of many data analysis tools. This product has a free version for educational purposes, a Russian-language interface and help. In Deductor 5.2, a neural network model of a multilayer perceptron and two learning algorithms are implemented, including an error back propagation algorithm. Deductor 5.2 most fully meets the above requirements. The only drawback is overloaded functionality. It is well suited for conducting research or in-depth study of the theory of data processing and knowledge extraction.

Among the add-ons for supporting neural network models of application computing packages, the most convenient tool for developing complex neural systems is Matlab with neural network tools attached to it. A set of neural network extensions for the Matlab `Matlab_Neural_Network` application computing package provides a convenient environment for the synthesis of neural network techniques with other data processing methods, such as wavelet analysis, statistics, financial analysis, and others. The applications developed in the Matlab system can then be retransmitted to C++. `Statistica_Neural_Networks` is a set of neural network extensions for the Statistica applied statistics package and `Excel_Neural_Package`. It is a set of libraries and scripts for Excel spreadsheets that implements some neural network data processing capabilities. The disadvantages of these settings include the high cost of the environment itself and the settings, an interface that is inconvenient for students to master.

Specialized neural network packages are designed to solve only a certain class of problems or only a specific task. Moreover, the methods of solving these problems are often determined by the developer himself, which makes their use in the educational process unacceptable. As an example, let's compare the capabilities of the most available specialized packages [2].

BrainMaker is designed to build neural networks of back propagation. The package includes a program for the preparation and analysis of source data (`netmaker.exe`), a program for training and launching neural networks (`brainmak.exe`), as well as a set of utilities of wide application. Deductor Academic, which is an analytical platform, is the basis for creating complete applied solutions in the field of data analysis. The technologies implemented in Deductor make it possible to go through all the stages of building an analytical system on the basis of a single architecture, from creating a data warehouse to automatically selecting models and visualizing the results obtained.

We can also mention the following packages. They are IBM SPSS Neural Network (it is a neural network based on a multilayer perceptron (MLP) with automatic and manual architecture selection), Neuroshell Trader (it is a program for creating neural networks for market analysis) and an "Eye" (the program is designed for processing aerospace information) [8]. There are also a large number of proprietary simulators distributed free of charge, which have one or more disadvantages. For instance, an inconvenient interface, limited functionality, they are often unstable in operation and do not have the possibility of modifying the embedded algorithms. The elective course, which aims to

get acquainted with the principles of creating a mathematical model of the ANN, does not set the task of mastering complex software tools. Since for the first immersion in the subject area of ANN, it is more important for students to understand the basic principles of ANN, and based on the fact that there are many tasks and interactive objects in ANN modeling that require programming skills. It became necessary to select or develop a tool that provides ready-made templates for ANN management.

To test and consolidate in practice the basic knowledge of the mathematical basics necessary for further development of the course “Neurocomputer Technologies”, software is required that allows students to test the level of mastery of basic arithmetic operations on vectors necessary for the study and creation of artificial neural networks. For instance, vectors’ addition or subtraction, scalar production vectors, calculation of the length of a vector, the outer product of vectors, the cosine of the angle between vectors, the square of the distance between vectors and convolution of vectors.

Thus, in order to conduct practical classes and organize independent work of students, as well as for self-control in the process of mastering the course, accessible and easy-to-learn software is required that allows you to clearly demonstrate the peculiarities of the behavior of the neural network simulator. Due to the lack of neural simulators that would fully meet the requirements of supporting classical neural network models, the availability of data analysis tools, as well as the requirements of accessibility, reliability, visibility and having a Russian-language interface, it was decided to create our own software tools for working with neural networks, available for mastering by students of 8-11th grades.

3 Development of the Methodology

When developing the program of the course “Neurocomputer Technologies”, the authors proceeded from the fact that propaedeutic training, which is the basis of continuing education in mathematics and computer science, can occur at different levels of complexity.

The first is the level of popularization, which is an introduction to the subject and differs in an elementary form of presentation. At this level, the course program is aimed at eliminating the gap in the mathematical and information literacy of students of secondary schools, lyceums and gymnasiums. The course content includes the most important questions of the theory of artificial neural networks. The course materials use the school mathematics course, as well as deepen and expand it, which creates conditions for understanding the most important concepts and terms of the theory of neural networks.

Further professional education of “highly motivated” gifted children is highly likely to be associated with scientific research, which, regardless of the field of human knowledge, is currently based on the use of ANN, neuromathematics and neuropackages. Since the main tasks solved with the help of neuro-mathematics are related to the search for hidden patterns, classification, forecasting, dimension reduction, the scope of their application covers all scientific areas.

The main purpose of mastering the discipline “Neurocomputer Technologies” is to study the main types of ANN currently being created in scientific laboratories, as well as generalization of the main ideas, approaches and methods of working with them. The

main objective of the course is to master the students of the theory of elementary mathematics modern apparatus, the application of modern approaches to the study of INS, the solution of typical problems arising in the training of neural networks. To date, algorithms for constructing and training artificial neural networks have been implemented as modules in various mathematical packages, for example, in MatLab, Statistica, Scilab, etc. These algorithms can be implemented independently in any available programming language, such as Python or C ++, with a certain desire, qualification and sufficient time, which is possible within the framework of a propaedeutic course.

In accordance with the main objective of the course, the following didactic units were included in the content of the discipline.

- The concept of a mathematical model.
- The concept of an arithmetic vector. Scalar product of arithmetic vectors. Convolution of arithmetic vectors.
- Nonlinear functions: exponent, tangent and logarithm. Complex (composite) function.
- Recursive functions.
- The structure of the human brain, the structure of a biological neuron, the structure of an artificial neuron (McCulloch model), neuron activation functions, artificial neuron training (Rosenblatt algorithm), the simplest single-layer network (perceptron), restrictions on the tasks solved by the perceptron.
- Two-layer neural networks, three-layer neural networks, training of multilayer neural networks, the method of backward error propagation, stochastic methods of training multilayer neural networks.
- Methods of teaching ANN without a teacher, the Winner-takes-all principle, Kohonen network, Grossberg layer.
- The use of counter-distribution networks.
- Neural networks with feedback, associative memory, Hopfield network, Hamming network, Hopfield network information capacity, associative memory applications.
- Methods of accelerating the work of ANN and algorithms of their training, the use of convolutional neural networks for the recognition of graphic images.

In practical classes, students will have the opportunity to learn how to create a mathematical model of ANN and teach the main types of ANN.

The advanced level involves mastering the skills of creating mathematical models of the main types of ANN, as well as the skills of teaching the main types of ANN. The tasks of the workshop assume that students have basic programming skills, since all tasks are performed on a computer using one of the programming languages or using specially developed pedagogical software tools for modeling individual components and types of artificial neural networks. When performing the work, it is supposed to use a set of author's software "Neurocomputer technologies (training course)".

Mastering the course involves the following practical work:

1. Introduction to mathematical and simulation modeling. The purpose of the lesson: to consolidate in practice the basic knowledge of the mathematical foundations of the course. Issues under consideration are finding the sum, difference, product and quotient of numbers, extracting the square root of a number, finding the sum and

difference of vectors a and b , calculating the scalar product of vectors a and b , calculating the outer product of vectors a and b , finding the length of vector a and finding convolution of vectors a and b . In support of the lesson, the “Testing module of basic operations on arithmetic vectors” NT_vectors was developed: https://github.com/anserion/NT_vectors. The module tests the mastery of basic arithmetic operations on vectors necessary for studying and creating artificial neural networks, for instance, addition/subtraction of vectors, scalar product of vectors, vector length, outer product of vectors, cosine of the angle between vectors, square of the distance between vectors, convolution of vectors.

2. Perceptrons. The purpose of the lesson is to consolidate in practice the basic knowledge of the simplest neural networks specifically perceptrons and methods of their training. Issues under consideration are the structure of the human brain, the structure of a biological neuron, the structure of an artificial neuron (McCulloch model), the activation function of a neuron, artificial neuron training (Rosenblatt algorithm), the simplest single-layer network (perceptron) and restrictions on the tasks solved by the perceptron. In support, a computer program was developed that simulates the process of recognizing graphic images according to Rosenblatt’s theory of perception (1957) “Rosenblatt Perceptron Training Model” NT_perceptron: https://github.com/anserion/NT_perceptron. The sensory field of the program has a size of 5×5 cells, the operation of 10 A-elements and two R-elements is simulated. To train the perceptron, an error correction algorithm with quantization is implemented (additionally, manual input of R-element coefficients is allowed). The verification module assumes that the “nearest distance” criterion has a higher recognition quality than the perceptron design and gives the percentage of correct responses of the perceptron relative to the “nearest distance” model.
3. Multilayer neural networks. The purpose of the lesson is to consolidate in practice the basic knowledge of multilayer neural networks and methods of their training. Issues under consideration are two-layer neural networks, three-layer neural networks, training of multilayer neural networks, the method of error back propagation and stochastic methods of training multi-layer neural networks. Three software modules were developed to support the lesson.
 - 1) “The academic model of a three-layer neural network with extended visualization” NT_triple_L3_5x5: https://github.com/anserion/NT_triple_L3_5x5. The project is a computer program that simulates the process of recognizing graphic images using a three-layer neural network. The sensory field of the program has a size of 5×5 cells, the work of 5 neurons of the first layer, 6 neurons of the second layer and 25 neurons of the third layer is simulated. The outputs of the neurons of the third layer are organized in the form of a 5×5 square image of cells, which makes it possible to study more clearly the behavior of three-layer neural networks. To train a neural network, an error back propagation algorithm is implemented.
 - 2) “The academic model of a three-layer neural network” NT_triple: https://github.com/anserion/NT_triple. The project presents a computer program that simulates the process of recognizing graphic images using a three-layer neural network. The sensory field of the program has a size of 5×5 cells, the work of 5 neurons of the first layer, 5 neurons of the second layer and 4 neurons of the third layer is simulated. To train a neural network, an algorithm for backward error propagation is implemented.

- 3) “Training model of bidirectional associative memory neural network (The academy-ic model of bidirectional associative memory neural network)” NT_bidirmem: https://github.com/anserion/NT_bidirmem. The project presents a computer program that simulates the process of operation of the bi-directional associative memory (BAM) of restoring and associating graphic images using an artificial neural network. The sensory field of the program has a size of 5x5 cells, the operation of a continuous two-layer neural network consisting of 50 neurons is simulated.
4. Counter-distribution networks. The purpose of the lesson is to consolidate in practice the basic knowledge on the architecture of counter-distribution networks and methods of their training. Considered issues are methods of teaching artificial neural networks without a teacher, the Winner-takes-all principle, the Kohonen network. The Grossberg layer and the application of counter-propagation networks. In support of the “The academic model of a counterpropagation neural network” NT_counter was developed: https://github.com/anserion/NT_counter. The project presents a computer program that modulates the process of recognizing graphic images using a neural network of counter-propagation. The sensory field of the program has a size of 5x5 cells, the work of 10 Kohonen layer neurons, 25 Grossberg layer neurons is simulated.
5. Recurrent neural networks. The purpose of the lesson is to consolidate in practice the basic knowledge on the architecture of recurrent neural networks and methods of their training. Issues under consideration are the concept of neural networks with feedback, associative memory, Hopfield network, Hamming network, Hopfield network information capacity, applications of associative memory. Two software modules have been developed in support. They are:
 - 1) “The academic model of a simple RBF-neural network” NT_RBF: https://github.com/anserion/NT_rbf. The project presents a computer program that simulates the process of recognizing graphic images using a simple RBF neural network. The sensory field of the program has a size of 5x5 cells, the work of 4 RBF neurons in the hidden layer and 4 neurons in the output layer is simulated. Neural network training has been replaced by direct calculation of the coefficients of the hidden and output layers.
 - 2) “The academic model of Hopfield’s neural network” NT_hopfield: https://github.com/anserion/NT_hopfield. The project presents a computer program simulating the process of restoring graphic images using the Hopfield neural network (associative memory). The program’s sensory field has a size of 5x5 cells, the operation of a continuous Hopfield neural network consisting of 25 neurons is simulated.
6. Convolutional neural networks. The purpose of the lesson is to consolidate in practice the basic knowledge on the architecture of convolutional neural networks and methods of their training. The issues under consideration are methods of accelerating the work of artificial neural networks and algorithms for their training, the use of convolutional neural networks for the recognition of graphic images. In support, a “Training model of a convolutional neural network was developed. (The academic model of convolution neural network)” NT_convolution: https://github.com/anserion/NT_convolution. The project presents a computer program that simulates the operation of a convolutional neural network for recognizing graphic images. The sensory field of the program has a size of 25x25 cells, the operation of a convolutional neural network consisting of two convolutional, two subdiscretizing and three output “secondary”

layers of neurons is simulated. The first convolutional layer contains 4 cores with a size of 5×5 , the second layer is a subsampling convolution to a 5×5 matrix, the third layer is a convolutional layer with 8 cores with a size of 3×3 , the fourth layer is a subsampling convolution to single signals. Layers from the fifth to the seventh are a fully connected three-layer neural network of direct distribution (classifier). The fifth layer contains 10 neurons connected to the outputs of the fourth layer. The sixth layer contains 10 neurons connected to the outputs of the fifth layer. The seventh layer contains 25 neurons connected to the outputs of the sixth layer. Training of all layers, including convolutional ones, is carried out by the method of error back propagation. The neural network successfully recognizes 4 noisy standards, 5×5 in size, after about 20–30 thousand training cycles (self-learning is performed on noisy standards selected in random order).

The recommended material and technical support includes a computer class with the number of serviceable computers according to the number of students in the group plus the teacher's workplace and productivity sufficient for mastering the subject. All computers of the class (laboratories, classrooms) are supposed to have access to the Internet sources of subject at a speed sufficient for comfortable work.

The authors proceeded from the fact that the main criterion for mastering the course is the successful completion of tasks by students in practical classes. All the tasks of the workshop were designed in such a way that both at the final and intermediate stages it was possible to check the correctness of the execution. To control the correctness of practical work, special test cases were used, for which the result is known for theoretical reasons. The data of these examples were offered by the teacher or, in some cases, generated by the students themselves.

4 Discussion

To conduct practical classes and organize independent work of students, as well as for self-control during the course development, the authors developed an affordable and easy-to-learn software that allows to clearly demonstrate the peculiarities of the behavior of an ANN simulator of neural networks.

As long-term practice has shown, this course can be mastered by students with different initial levels of mathematical training, who possess basic arithmetic knowledge and have a desire to learn the course materials. The authors proceeded from the fact that the main criterion for mastering the course is the successful completion of tasks by students in practical classes. It is too early to carry out a systematic assessment of the effectiveness of the pedagogical software tools developed by the authors. However, observation of the behavior of trainees showed that the use of computer simulators in practical classes arouses keen interest and resonates with different categories of trainees. Assuredly, fully mastering the methods of neural network synthesis and their practical application to solve various practical problems involves preliminary mastery of the basics of such subjects as mathematical analysis, linear algebra, optimization methods, learning theory, data analysis, programming. However, the use of training simulators at this stage of training allows you to immerse yourself in the conceptual apparatus the theory of

elementary mathematics, as well as the principles of solving typical problems that arise when training neural networks.

The results of the assessment of students' knowledge showed that in the process of performing practical work on the study of neural networks on the material available to schoolchildren of 8-11th grades. They develop skills related to the theory of elementary mathematics, algorithms and reasoning systems used in various components of mathematical and information education.

It is not possible to estimate the material costs, since the authors of the course did not initially envisage the use of expensive software. Nevertheless, the use of educational software modules of computer simulators of the ANN at this stage of training is preferable, since in this case a large amount of rough work is already done. And the student can focus on the methods of solving the task, interpreting and presenting the results of modeling and formulating conclusions based on the data analysis. Thus, with the help of specially created computer simulators of the work of the ANN, students can overcome the problems associated with access to professional thematic packages and study the work of the ANN in any convenient place and at any time. It can be concluded that the creation of pedagogical software tools to support the development of this course contributes to the propaedeutics of the formation of competencies necessary for studying the course on neural networks.

5 Conclusion

The methodology of conducting classes has been tested in the State Educational Institution "Center for Creative Development and Humanitarian Education for Gifted Children "Poisk" (Stavropol, Russia) in the course "Neurocomputer Technologies". The author's computer simulators of artificial neural networks are used in the training, which allow to simplify the learning process. At the same time, they provide students with a real experience of using neurocomputer technologies.

The interface of the developed modules is made in Russian language, intuitive and contains visual hints. In addition, each simulator is accompanied by a separate instruction and options for performing the work. This provides working with the simulator intuitive, which is especially important in the case of self-completion of the task on a home computer.

The study showed that the use of imitating the ANN software at the propaedeutic stage allows avoiding a large number of typical errors in the process of understanding the basic algorithms of ANN training and their application. All the listed models are available for use under free licenses without any restrictions.

Currently, the work is under the registration of this software. The next stage of improving the developed simulators will be the export of content in HTML5, which can work on any web browser and device, including a desktop computer, tablet and smartphone.

References

1. Abiodun, O.I., Jantan, A., Omolara, A.E., Dada, K.V., Mohamed, N.A., Arshad, H.: State-of-the-art in artificial neural network applications: a survey. *Heliyon* **4**(11), e00938 (2018)
2. Kirichenko, A.A.: Neuropackages are a modern intellectual tool of a researcher [Electronic resource]. <https://www.hse.ru/data/2013/08/26/1290192359/>. Accessed 22 Mar 2022 (in Russian)
3. NeuroSolutions. <http://www.neurosolutions.com/neurosolutions/>. Accessed 22 Jan 2022
4. MemBrain Neural Network Editor and Simulator. https://membrain-nn.de/main_en.htm. Accessed 10 Oct 2021
5. Stuttgart Neural Network Simulator. <https://www.ra.cs.uni-tuebingen.de/SNNS/welcome.html>. Accessed 10 Oct 2021
6. NeuroPro - nejronnye seti, metody analiza dannyh: ot issledovaniy do razrabotok i vnedrenij/[Neuroprocessor networks, data analysis methods: from research to development and implementation]. <https://www.mql5.com/en/articles/830>. Accessed 5 Oct 2021 (in Russian)
7. BaseGroup Labs. <http://www.basegroup.ru/>. Accessed 10 Dec 2020 (in Russian)
8. IBM SPSS Neural Network. <https://www.ibm.com/products/spss-neural-networks>. Accessed 10 Oct 2021

Author Index

A

Abdulkadirov, R. 26
Abdulsalyamova, A. S. 373
Abdulsalyamova, Albina 343
Almamedov, Parviz 364
Andruhiv, L. 26
Andrukhiv, Ludmila 500
Antonov, Sergei A. 462
Antonov, V. O. 205
Antonov, V. V. 270
Apekov, Aslan 3, 179

B

Babenko, Mikhail 258, 277
Baboshina, Valentina 489, 500
Baragunova, L. A. 131, 137
Baranov, Nikolay 410
Basalov, Yu. A. 157
Basan, Elena 430
Basu, Abhishek 450
Basyuk, Anatoly 430
Bergerman, Maxim 343
Beshtokov, Murat 106, 118
Beshtokova, Zaryana 15
Bezuglova, Ekaterina 300, 321
Bondarenko, Daria Nikolaevna 401
Boyarskaya, Emiliya Evgenevna 401

C

Chakraborty, Nabarun 217
Choudhury, Atlanta 243

D

Dobrovol'skii, N. M. 157
Dobrovol'skii, N. N. 81, 157
Dobrovol'skii, N. M. 81
Dokhov, Rezuan 146
Dzhiyanov, T. O. 45

E

Ernazarov, Mirzohid 54

F

Fedorenko, Vladimir 167

G

Gautam, Nandita 450
Gladkov, Andrey 321

I

Ionisyan, Andrej 509
Ionisyan, Andrey S. 335

K

Kalita, Diana 364
Kaplun, Dmitrii I. 462
Kaplun, Dmitry 450
Khamukova, Liana 3, 179
Khaydarov, O. Sh. 87
Kholiyarov, E. Ch. 87
Kholiyarov, Erkin 54
Kholliiev, Fakhriddin 98
Khuzhayorov, B. H. 87
Khuzhayorov, Bakhtiyor 98
Kiladze, Mariya 421
Kokov, Zaur 179
Kucherov, Nikolay 167, 258, 288, 300
Kuchukov, Viktor 258, 288, 312
Kuchukova, Nataliya 258
Kuljonov, Jakhongir 68
Kulterbaev, Kh. P. 131, 137
Kurmaev, Roman 36

L

Lafisheva, M. M. 131, 137
Lapina, Maria 250, 430
Lavrinenko, Irina 489, 509
Lesnikov, Alexander 430
Liksonova, Darya 250

Liutova, Violetta 509
Lutsenko, Vladislav 277
Lyakhov, Pavel 343, 381
Lyakhova, Ulyana Alekseevna 353, 401

M

Madi, Siyanda L. 270
Makhmudov, Jamol 68
Malafey, Oleg 475
Mamatov, Sh 45
Mandritsa, I. V. 189, 270
Mandritsa, O. V. 189
Martirosyan, Elizaveta 36
Minkina, T. V. 189
Misra, Aradhana 217, 233
Mogilny, Anton 430

N

Nagornov, N. N. 373
Nagornov, Nikolay Nikolaevich 401

O

Orazaev, Anzor R. 335, 391

P

Pachev, Urusbi 146
Pavlenko, Tatyana 475
Petrenko, V. I. 189, 205

R

Rebrov, E. D. 157
Rebrov, E. D. 81
Rebrova, I. Yu. 81
Ryabchikova, Valeria 36
Ryabtsev, S. S. 205

S

Samoylenko, Vladimir 167
Sarkar, Ram 450
Sarma, Dikshita 233
Sarma, Kandarpa Kumar 217, 233, 243
Semenova, Natalia 489
Semyonova, N. F. 373
Shalugin, Evgeniy D. 462
Shandilya, Mrinmoy 233
Shaposhnikov, Aleksey 509
Shaposhnikov, Aleksey V. 335
Shebzukhova, Irina 179
Shiriaev, Egor 288, 312
Sinitca, Aleksandr M. 462
Slobodskoy, Vitaly 36
Struchkov, I.V 205

T

Tebueva, F. B. 205

U

Usmonov, Azizbek 68, 98

V

Valuev, Georgy 321
Vershkov, Nikolay 258
Voznesensky, Aleksander S. 462

Y

Yartseva, E. 26
Yartseva, Elena 500

Z

Zhuk, A. S. 441
Zhuravleva, Irina 500, 509
Zokirov, M. S. 45
Zolotarev, Vyacheslav 250