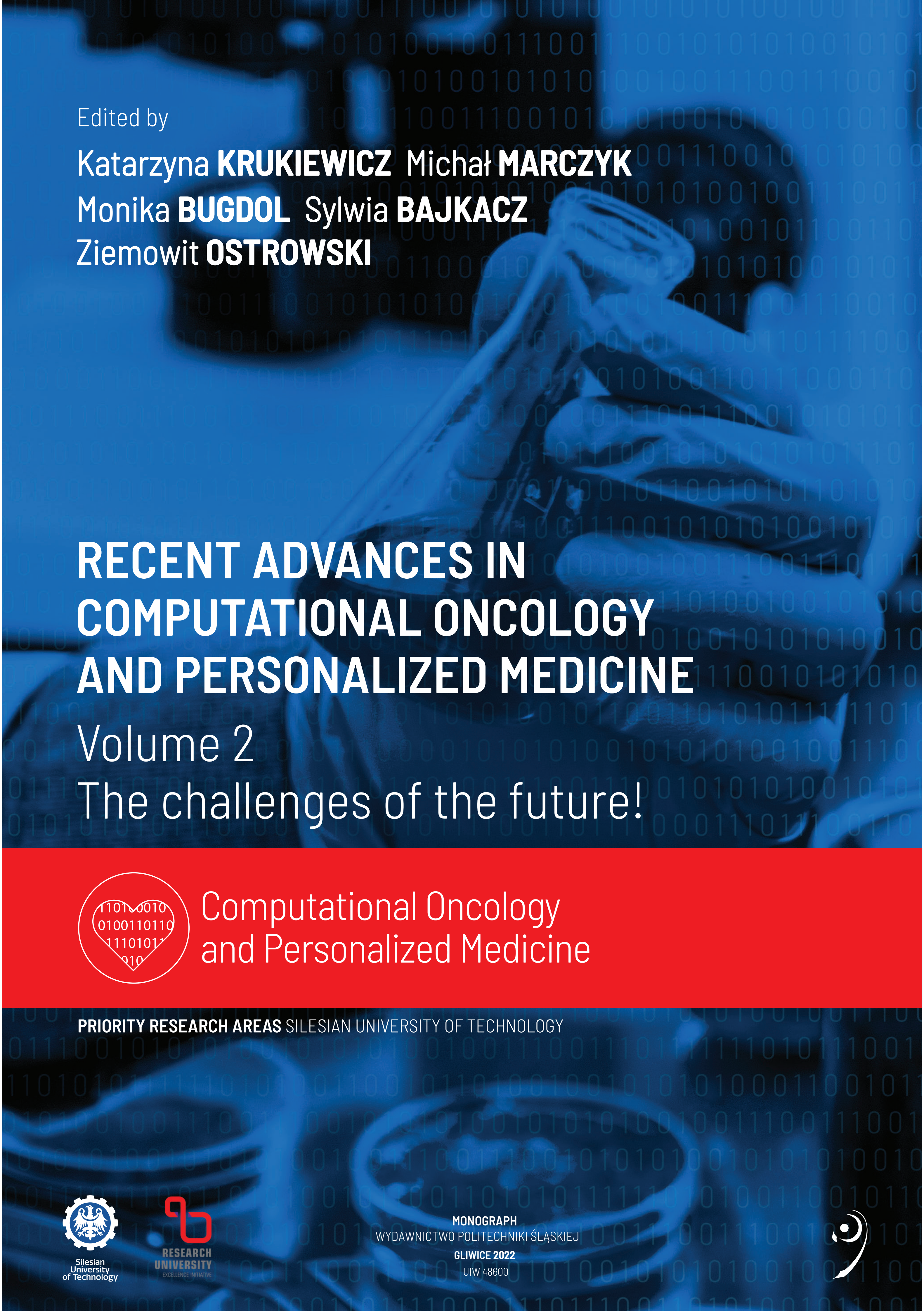


Edited by

Katarzyna **KRUKIEWICZ** Michał **MARCZYK**
Monika **BUGDOL** Sylwia **BAJKACZ**
Ziemowit **OSTROWSKI**

RECENT ADVANCES IN COMPUTATIONAL ONCOLOGY AND PERSONALIZED MEDICINE

Volume 2

The challenges of the future!

Computational Oncology
and Personalized Medicine

PRIORITY RESEARCH AREAS SILESIAN UNIVERSITY OF TECHNOLOGY

Silesian
University
of Technology

RESEARCH
UNIVERSITY
EXCELLENCE INITIATIVE

MONOGRAPH
WYDAWNICTWO POLITECHNIKI ŚLĄSKIEJ
GLIWICE 2022
UIW 48600

Edited by
Katarzyna KRUKIEWICZ Michał MARCZYK
Monika BUGDOL Sylwia BAJKACZ
Ziemowit OSTROWSKI

RECENT ADVANCES IN COMPUTATIONAL ONCOLOGY AND PERSONALIZED MEDICINE

Volume 2
THE CHALLENGES OF THE FUTURE!

WYDAWNICTWO POLITECHNIKI ŚLĄSKIEJ
GLIWICE 2022
UIW 48600

Review
Prof. n. techn. dr hab. n. fiz. inż. lek. Halina PODBIELSKA
Prof. dr hab. inż. Wiesław SZEJA

Published with the approval of the Rector
of the Silesian University of Technology

Publication supported by the Silesian University of Technology Rector's
Pro-quality programme grant No. 04/040/RGM22/0230

Cover design
Maciej MUTWIL

Layout and Typesetting
Joanna Jenczewska-Pajka

ISBN 978-83-7880-876-3

Editors:

Katarzyna Krukiewicz
Department of Physical Chemistry and Technology of Polymers
Silesian University of Technology, Gliwice, Poland

Michał Marczyk
Department of Data Science and Engineering
Silesian University of Technology, Gliwice, Poland

Monika Bugdol
Department of Informatics and Medical Devices
Silesian University of Technology, Gliwice, Poland

Sylwia Bajkacz
Department of Inorganic Chemistry, Analytical Chemistry and Electrochemistry
Silesian University of Technology, Gliwice, Poland

Ziemowit Ostrowski
Department of Thermal Technology
Silesian University of Technology, Gliwice, Poland

Series Editor:

Joanna Polańska
Department of Data Science and Engineering
Silesian University of Technology, Gliwice, Poland

The series *Recent Advances in Computational Oncology and Personalized Medicine* contains publications on research, theory, applications, and design in the field of computational oncology, personalized medicine, biomaterials and medical biotechnology, imaging informatics and telemedicine, biomechanics, analysis and drug design and public health.

The series is published within *The Excellence Initiative – Research University programme* that is implemented to increase scientific excellence and the international importance of the activities of Polish universities. As one of ten Polish universities, *the Silesian University of Technology* obtained the status of a *research university* and commenced the implementation of the programme aimed at increasing its scientific excellence and international significance.

The main objectives of the programme at *The Silesian University of Technology* are:

- Increasing recognition and visibility of the SUT, building the prestige of the University in the group of world-class universities and improving the position in renowned international rankings.
- Conducting high-quality scientific research and implementation of innovations to improve the living and health conditions of people as well as economic development.
- Ensuring high-quality education at the undergraduate and masters levels and the doctoral school, based on research and innovation.
- Supporting the scientific development of employees and PhD students.
- Increasing internationalization of science and education.

LIST OF CHAPTERS

SPIS TREŚCI

PREFACE

The second book in the series continues to report recent research and applied solutions in the area of computational oncology and personalized medicine, being the topic of the *Priority Research Area #1* (POB1) of the *Silesian University of Technology*. The chapters are the result of the research conducted by the staff of SUT and by the partners from both the academic world and socio-economic environment. It aims to present research results and interests and to exchange information on scientific activities related to the scope of POB1.

The idea behind the monograph is to spread among the scientific community the recent advances in computational methods in medical applications. We believe that the new ideas and applications presented within the series would lead to the implementation of new diagnostic tools, drugs or therapies to assist clinicians in their practice and eventually would lead to an increase in life quality.

The chapters of this monograph cover mostly the topics that are the response of the scientific community to the needs raised by clinical partners.

The largest number of chapters represent topics related to **bioinformatics**, including: **medical image processing and analysis** (X-ray microtomography in histopathology; the algorithm for keratin 10 determining in organotypic cell cultures images; and still relevant and important issues sourced in COVID-19: chest X-ray images preprocessing for COVID-19 patients AI-driven classification; denoising autoencoder model for chest X-ray images preprocessing), **classification systems** (feature extraction and dimensionality reduction as the first step of the full hands-free machine learning analysis; neural network for mixed types big data visualization) as well as **cancer-related research** (clonal evolution/somatic mutations in cancers; unsupervised algorithms in the gene expression data of different human cancers) and **general applications** (recurrence and cross-recurrence quantification analyses of the time series in pharmacology and physiology).

The second largest group of chapters is focused on **(bio)materials engineering** (incl. biofunctionalization of surfaces; polycaprolactone in tissue engineering; polyhydroxybutyrate accumulation in bioplastic) and **tissue properties** (subcutaneous adipose tissue).

The third group is related to **numerical modelling in biomedical applications** (incl. skin tumour freezing; cardiovascular biofluid mechanics).

Finally, one chapter covers the up-to-date problem, especially during the pandemic time, of **e-learning course evaluation methodology**.

The publication of the book is one of the activities carried out by the *Silesian University of Technology* and it was published, among others, as a result of *the Excellence Initiative – Research University* programme. As one of ten Polish universities, *the Silesian University of Technology* obtained the status of a research university and commenced the implementation of the programme aimed at increasing its scientific excellence and international significance.

Editors would like to express their gratitude to the authors who have submitted their original research chapters as well as to all the reviewers for their valuable comments. Your effort has contributed to the high quality of the book that we pass on to the readers. We also hope that this second volume in the series would not only ignite a new, joint, interdisciplinary and challenging research but could also result in new methods and diagnostic tools. Such solutions, in future, could also support both clinicians and health agencies in better distribution/usage of available health budgets.

Katarzyna Krukiewicz
Michał Marczyk
Monika Bugdol
Sylwia Bajkacz
Ziemowit Ostrowski
(Editors)

August 2022
Gliwice

DOI:10.34918/85101

Małgorzata SKORUPA[1], Taral PATEL[1], Katarzyna KRUKIEWICZ[1,*]

Chapter 1. DIAZONIUM CHEMISTRY AS A ROBUST APPROACH FOR THE BIOFUNCTIONALIZATION OF TITANIUM SURFACE

1.1. Introduction

Metallic biomaterials are most commonly used as implants for bone replacement or support. These metals include stainless steel, cobalt alloys, and titanium and its alloys which have become widespread since the early 1970s [1, 2]. Titanium is characterized by an excellent corrosion resistance and high strength-to-density ratio, overcoming that of steel. Additionally, it is regarded as an inert material in physiological conditions [2]. This set of conducive, mechanical, physical, and biological properties has made Ti the most attractive metal for biomedical applications, such as dental and orthopedic implants [3]. Another advantage of Ti implants is their non-ferrous character, which allows the patients to be safely examined with magnetic resonance imaging [1]. The resistance to corrosion is associated with a spontaneous passivation of Ti surface [4]. Grade V Ti alloy (Ti-6Al-4V) is more prone to corrosion, resulting in the possible release of cytotoxic vanadium and aluminium in long-term implantations, whereas commercially pure Ti (Grade I-IV) remains more resistant [2, 5].

Surface properties of Ti implants, including surface chemistry, topography and wetting angle, have a decisive effect on the process of osseointegration, affecting cellular response and bacterial colonization [6]. The surface of the implant is especially susceptible to bacterial infection after the surgery, when the tissue gets disrupted and has reduced number of blood vessels [7]. Therefore, the immunological response at the implant/tissue interface is diminished. Additionally, defects in the

---

[1] Department of Physical Chemistry and Technology of Polymers, Faculty of Chemistry, Silesian University of Technology, Gliwice, Poland.

[*] Corresponding author: katarzyna.krukiewicz@polsl.pl.

passive layer and insufficient wear resistance of Ti implants can compromise their corrosion resistance and durability [8]. All these issues can be addressed by appropriate surface functionalization. The efforts to modify Ti surface include chemical and physical methods, e.g. surface texturing, plasma treatment, thermal oxidation, chemical etching, or physical vapour deposition, especially to obtain coatings like TiN or CrN [6, 9].

For decades, scientists have paid close attention to the functionalization of different surfaces in a wide range of applications [10]. Diazonium salts are one of the most versatile compounds in organic science, also they are frequently utilized in a variety of reactions [11]. Because of their high reactivity, diazonium salts are now widely used in surface modification. For instance, Chehimi et al. [12] modified the surface of TiO_2 with diazonium chemistry to initiate radical polymerization of methacrylate. Also Ti6Al4V alloy was modified with diazonium ions to create poly(hydroxyethyl) methacrylate and polyetheretherketone layers [13]. A number of articles has addressed electrochemical deposition of various metals on TiN surfaces as a simple and cost-effective approach [14, 15]. Although there are several coating techniques described in the literature, a diazonium-based surface modification is preferred over the rest, due to its ability to form covalently bonded aryl-layers with desired functionalities, such as alkyl, nitro, cyanide, carboxylic, ester, alcohol, thiol, and halogenated groups [16]. As-formed coatings exhibit good adherence and homogeneity [17]. One of the main benefits of the diazonium-based surface modification is that it may be used on any type of surface, regardless of size, shape, or geometry.

In this chapter, we present the process of surface modification of Ti with a 4-nitrobenzenediazonium salt, resulting in the formation of a layer of nitrobenzene molecules covalently attached to the surface at para position (denoted as Ti-NO_2). To extend the number of possible biofunctionalization pathways, Ti-NO_2 surface was subjected to an additional electrochemical reduction reaction, resulting in the reduction of nitrobenzene moiety to aniline (Ti-NH_2). To verify whether the functionalization procedure was successful, unmodified and modified Ti discs were investigated by means of electrochemical impedance spectroscopy. As-formed Ti-NH_2 surfaces can be further biofunctionalized by a variety of methods to immobilize particular biomolecules (DNA, enzymes, proteins, peptides, etc.), metal complexes, polymers, various nanoobjects and molecular species (carbon nanotubes, fullerenes, metal nanoparticles, etc.), in order to equip Ti surface with unique biological characteristics.

1.2. Materials and methods

1.2.1. Reagents

Ti discs (diameter: 1 in., thickness: 0.02 in., in line with ASTM-B-265/ASME-SB-265 GR 2 specifications), 4-nitrobenzenediazonium tetrafluoroborate, NBF (97%), potassium chloride, KCl (>99.0%), and tetrabutylammonium hexafluorophosphate, nBu_4NPF_6 (>99.0%), were obtained from Sigma Aldrich (Saint Louis, MO, USA). nBu_4NPF_6 was vacuum dried before use. Deionized water (Millipore quality), ethanol (99.8%, Avantor) and acetonitrile (ACN, HPLC grade, Sigma Aldrich) were used as solvents.

1.2.2. Surface modification

The electrochemical functionalization of Ti surface was performed by means of a PARSTAT 2273 potentiostat (Ametek, Berwyn, PA, USA) by a cyclic voltammetry (CV) scanning. A standard three-electrode setup was used, comprising a Ag/AgCl (3 M KCl) reference electrode (ET073, EDAQ, Denistone East, Australia), a platinum plate counter-electrode (1 cm^2) (Mennica Polska, Warsaw, Poland), and a Ti disc working electrode (exposed area of 0.283 cm^2) (Sigma Aldrich, Saint Louis, MO, USA). Ti surface grafting with NBF (5 mM) was carried out in 0.1 M nBu_4NPF_6 ACN solution, within a potential range from -1.0 V to 0.5 V (vs. Ag/AgCl), for 5 CV cycles at a scan rate of 100 mV/s. Further electrochemical functionalization of Ti was performed in the 0.1 M KCl ethanol:water (1:9 v/v) solution, within a potential range from -1.0 V to 0 V (vs. Ag/AgCl), for 5 CV cycles at a scan rate of 100 mV/s.

1.2.3. Surface characterization

Electrochemical characterization of unmodified and modified Ti surfaces was performed by means of an electrochemical impedance spectroscopy (EIS) by collecting EIS spectra in 0.1 M KCl solution with frequencies ranging from 100 kHz to 100 mHz, an AC amplitude of 40 mV (vs. Ag/AgCl) and a DC potential equal to 0 V (vs. Ag/AgCl).

1.3. Results and discussion

Since diazonium chemistry has been known as an efficient approach for surface functionalization [18, 19], grafting and subsequent reduction of 4-nitrobenzenediazonium (NBF) salt was used for the covalent modification of a surface of model Ti disc electrodes (Fig. 1). The first step of functionalization involved an electrochemical reduction of NBF at the surface of Ti electrode. A distinct reduction peak at the potential of -0.4 V (vs. Ag/AgCl) was observed in a cyclic voltammetric (CV) curve associated with this process. NBF reduction peak, however, was observed particularly in the first CV cycle, and disappeared in the subsequent CV cycles, suggesting that Ti surface was successfully modified at the beginning of the reduction process and subsequent CV cycles did not further affect surface chemistry. As a result of the first step of functionalization, the surface of Ti disc was coated with a layer of nitrobenzene molecules covalently attached to the surface at para position (denoted as Ti-NO_2). To extend the number of possible biofunctionalization procedures, Ti-NO_2 surface was subjected to an additional electrochemical reduction reaction in a protic medium (water:ethanol solution), resulting in the reduction of nitrobenzene moiety to aniline (Ti-NH_2).

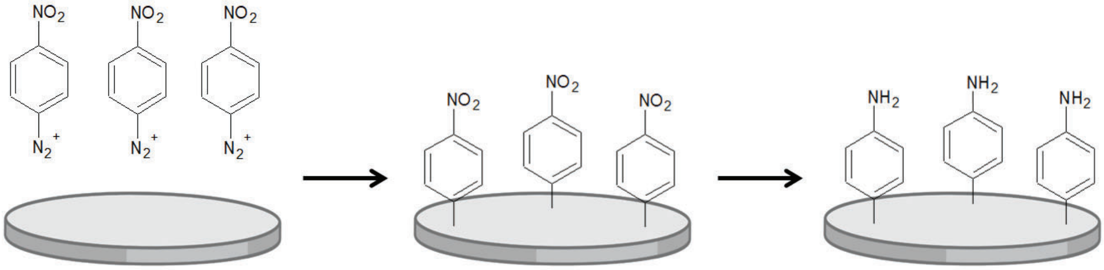


Fig. 1. Schematic representation of the process of Ti surface functionalization with diazonium chemistry (on the example of 4-nitrobenzenediazonium surface grafting)
Rys. 1. Schematyczne przedstawienie procesu funkcjonalizacji powierzchni Ti za pomocą soli diazoniowych (na przykładzie szczepienia powierzchniowego soli 4-nitrobenzenodiazoniowej)

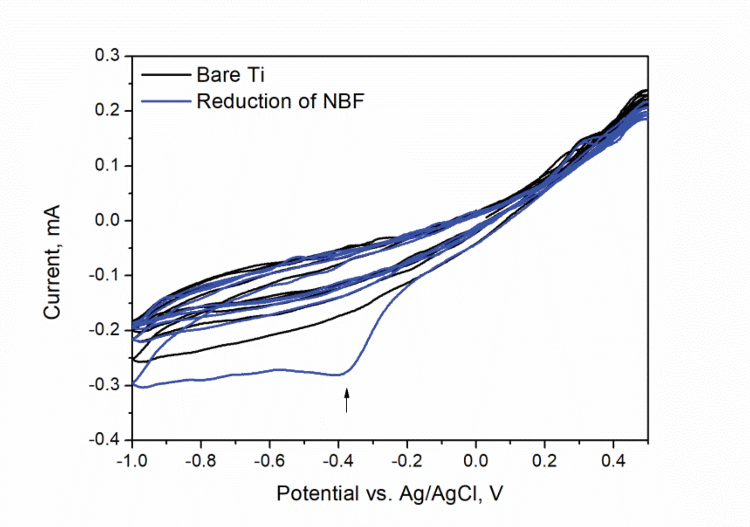


Fig. 2. Cyclic voltammetric curve collected during the reduction of NBF on Ti surface compared with the cyclic voltammetric curve collected during the reduction of Ti in the absence of NBF

Rys. 2. Krzywa woltamperometryczna zarejestrowana podczas redukcji NBF na powierzchni Ti, porównana z krzywą woltamperometryczną zarejestrowaną podczas redukcji Ti w roztworze bez NBF

To verify whether the functionalization procedure was successful, unmodified and modified Ti discs were investigated by means of electrochemical impedance spectroscopy (EIS). Due to the fact that EIS allows to detect minor changes in the mechanism of charge transfer, by comparing the shape of EIS spectra (Fig. 3), it is possible to assess the efficiency of surface modification. EIS data can be analyzed in three different ways. Impedance vs. frequency plots (Fig. 3A) can be used to compare the impedance behaviour of investigated samples in a wide frequency range. By observing the increase/decrease in the impedance values, it could be assessed whether the modification strategy leads to the decrease/increase in the conductivity of the surface, respectively. Phase angle vs. frequency plots (Fig. 3B), on the other hand, can be used to assess the capacitive behaviour of the samples. The number of peaks is related to the number of capacitive processes taking place in the investigated system, and the position of the peak is related to the time constant of the system. Finally, negative imaginary impedance vs. real impedance plots (Fig. 3C), also known as Nyquist plots, combine both the impedance and phase angle into the one plot in the complex plane. The semicircles observed in the Nyquist plots are associated with electrochemical processes such as charge transfer. In general, the larger the diameter of the semicircle, the less conducting is the investigated system.

EIS data clearly indicated the variation in Ti surface impedance at different stages of the functionalization procedure. Bare Ti discs were found to be highly electroactive, exhibiting a low impedance profile and a single phase angle peak at 100 Hz. The impedance profile was found to be typical for metal electrodes, with a low impedance at high frequencies and its increase at low frequencies [20]. The formation of Ti-NO_2 at the first stage of functionalization was found to have a deteriorating effect on surface conductivity by elevating the impedance profile and shifting a phase angle peak to higher frequencies (decreasing time constant). Interestingly, if the reduction of Ti was carried out in the absence of NBF, the resulting material (Ti-red) exhibited the worst electrochemical characteristics among all investigated materials, as evidenced from the largest semicircle observed in the Nyquist plot. Further reduction of Ti-NO_2 to form Ti-NH_2 resulted in the consecutive changes in impedance (decrease in the impedance profile) and capacitive (increase in time constant) behaviour of Ti surface. Basing on the EIS data, it could be concluded that since the electrochemical behaviour of Ti-NO_2 and Ti-NH_2 was different than that noted for bare Ti, both surface functionalization steps were successful in changing the surface characteristics of Ti. The presence of NBF-derived organic moiety on the surface of Ti electrodes was indirectly confirmed by observing the differences in electrochemical behaviour among Ti-NO_2, Ti-NH_2, and Ti-red samples, since the latter one was formed in the absence of NBF.

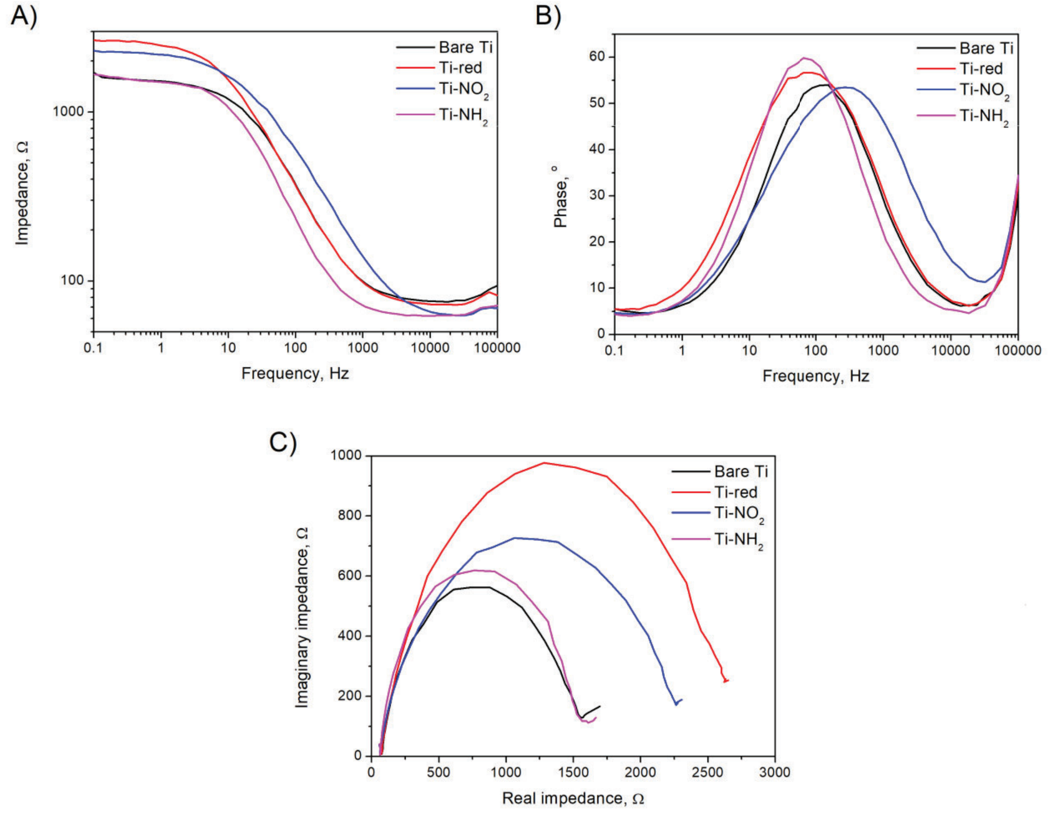


Fig. 3. EIS data in the form of (A-B) Bode plots showing the frequency-dependent behaviour of the (A) impedance modulus and (B) phase angle, and (C) Nyquist plots of bare Ti electrodes, as well as Ti electrodes subjected to reduction, NBF modification and post-modification

Rys. 3. Wyniki EIS w formie (A-B) wykresów Bodego przedstawiających zależność (A) impedancji i (B) kąta fazowego od częstotliwości, oraz (C) wykresy Nyquista dla czystej elektrody Ti, elektrody Ti poddanej redukcji, modyfikacji NBF oraz post-modyfikacji

Diazonium-grafted surfaces can be further biofunctionalized by a variety of methods (Fig. 4), including amide coupling, “click chemistry”, and post-diazotization, among others [21]. Therefore, the biological properties of Ti-NH_2 could be easily tailored to specific needs by the immobilization of particular biomolecules, such as DNA, enzymes, proteins, peptides, etc., in order to equip Ti surface with unique biological characteristics. Additionally, the chemistry of Ti-NH_2 allows for the immobilization of metal complexes, various nanoobjects, and molecular species, such as carbon nanotubes, fullerenes, metal nanoparticles, etc. Besides, covalently attached organic moieties can be used as precursors for surface-initiated polymerization reactions, leading to the formation of polymer brush-like structures of defined physicochemical characteristics and a great potential in biomedical engineering, including biosensing, cell culture and regenerative medicine [22].

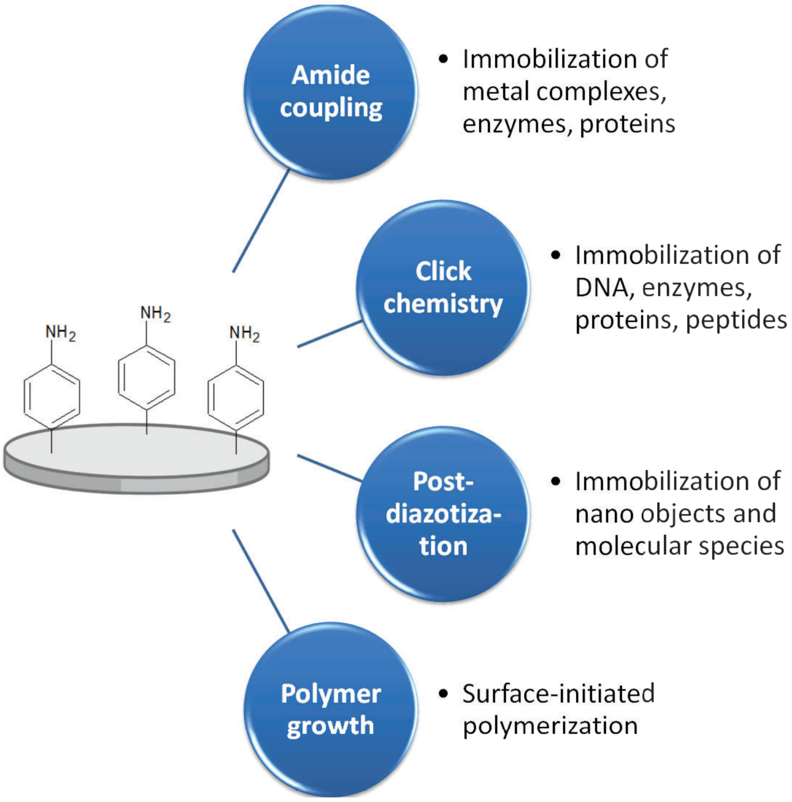


Fig. 4. Further routes of biofunctionalization for Ti-NH_2 surfaces
Rys. 4. Dalsze możliwości biofunkcjonalizacji powierzchni Ti-NH_2

1.4. Conclusions

Diazonium chemistry has been confirmed as an easy and straightforward approach to modify the surface of Ti by means of an electrochemical grafting procedure. As a result, two types of surface chemistries have been formed: Ti-NO_2 and Ti-NH_2, which could be further modified by a variety of methods to immobilize particular biomolecules (DNA, enzymes, proteins, peptides, etc.), metal complexes, polymers, various nanoobjects and molecular species (carbon nanotubes, fullerenes, metal nanoparticles, etc.), in order to equip Ti surface with unique biological characteristics suitable for particular biomedical needs.

Acknowledgements

This research is supported by National Science Centre (Poland) project No. 2019/35/B/ST5/00995.

Bibliography

1. D. Bombač, M. Brojan, P. Fajfar, F. Kosel, R. Turk: Review of materials in medical applications, *RMZ – Mater. Geoenvironment*. (2007) **54**:471–499.
2. C.N. Elias, J.H.C. Lima, R. Valiev, M.A. Meyers: Biomedical applications of titanium and its alloys, *JOM* (2008) **60**:46–49.
3. H.J. Rack, J.I. Qazi: Titanium alloys for biomedical applications, *Mater. Sci. Eng. C.* (2006) **26**:1269–1277.
4. A.T. Sidambe: Biocompatibility of Advanced Manufactured Titanium Implants-A Review, *Mater.* (2014) **7**:8168–8188.
5. M. Hendrickson, S.A. Mantri, Y. Ren, T. Alam, V. Soni, B. Gwalani, M. Styles, D. Choudhuri, R. Banerjee: The evolution of microstructure and microhardness in a biomedical Ti–35Nb–7Zr–5Ta alloy, *J. Mater. Sci.* (2017) **52**:3062–3073.
6. A. Woźniak, M. Adamiak, G. Chladek, M. Bonek, W. Walke, O. Bialas: The Influence of Hybrid Surface Modification on the Selected Properties of CP Titanium Grade II Manufactured by Selective Laser Melting, *Mater.* (2020) **13**:2829.
7. L. Zhao, P.K. Chu, Y. Zhang, Z. Wu: Antibacterial coatings on titanium implants, J. Biomed. Mater. Res. B. *Appl. Biomater.* (2009) **91**:470–480.

8. T. Hryniewicz, R. Rokicki, K. Rokosz: Corrosion and surface characterization of titanium biomaterial after magnetoelectropolishing, *Surf. Coatings Technol.* (2009) **203**:1508–1515.
9. X. Zhao, H. Liu, S. Li, X. Wang, Y. Sheng, P. Zhang, W. Li: Combined effect of TiN coating and surface texture on corrosion-wear behavior of selective laser melted CP-titanium in simulated body fluid*, J. Alloys Compd.* (2020) **816**:152667.
10. V. Mévellec, S. Roussel, L. Tessier, J. Chancolon, M. Mayne-L'Hermite, G. Deniau, P. Viel, S. Palacin: Grafting polymers on surfaces: A new powerful and versatile diazonium salt-based one-step process in aqueous media, *Chem. Mater.* (2007) **19**:6323–6330.
11. S. Bräse, S. Dahmen, C. Popescu, M. Schroen, F.J. Wortmann: The Structural Influence in the Stability of Polymer-Bound Diazonium Salts, *Chem. – A Eur. J.* (2004) **10**:5285–5296.
12. M.M. Chehimi, A. Lamouri, M. Picot, J. Pinson: Surface modification of polymers by reduction of diazonium salts: polymethylmethacrylate as an example, *J. Mater. Chem. C.* (2013) **2**:356–363.
13. M. Sandomierski, T. Buchwald, A. Patalas, A. Voelkel: Improving the abrasion resistance of Ti6Al4V alloy by modifying its surface with a diazonium salt and attaching of polyurethane, *Sci. Reports* (2020) **10**:1–10.
14. J. Vanpaemel, M. Sugiura, D. Cuypers, M.H. Van Der Veen, S. De Gendt, P.M. Vereecken: Electrochemical deposition of subnanometer Ni films on TiN, *Langmuir* (2014) **30**:2047–2053.
15. M.M.O. Thotiyl, T.R. Kumar, S. Sampath: Pd supported on titanium nitride for efficient ethanol oxidation, *J. Phys. Chem. C* (2010) **114**:17934–17941.
16. D. Bélanger, J. Pinson: Electrografting: A powerful method for surface modification, *Chem. Soc. Rev.* (2011) **40**:3995–4048.
17. Y. Cheng, Y.F. Zheng: Surface characterization and mechanical property of TiN/Ti-coated NiTi alloy by PIIID, *Surf. Coatings Technol.* (2007) **201**:6869–6873.
18. T. Menanteau, M. Dias, E. Levillain, A.J. Downard, T. Breton: Electrografting via Diazonium Chemistry: The Key Role of the Aryl Substituent in the Layer Growth Mechanism, *J. Phys. Chem. C.* (2016) **120**:4423–4429.
19. J. Pinson, F.I. Podvorica: Surface modification of materials: Electrografting of organic films, *Curr. Opin. Electrochem*. (2020) **24**:44–48.
20. C. Boehler, S. Carli, L. Fadiga, T. Stieglitz, M. Asplund: Tutorial: guidelines for standardized performance tests for electrodes intended for neural interfaces and bioelectronics, *Nat. Protoc.* (2020) **15**:3557–3578.

21. C. Gautier, I. López, T. Breton: A post-functionalization toolbox for diazonium (electro)-grafted surfaces: review of the coupling methods, *Mater. Adv.* (2021) **2**:2773–2810.
22. M. Krishnamoorthy, S. Hakobyan, M. Ramstedt, J.E. Gautrot: Surface-initiated polymer brushes in the biomedical field: Applications in membrane science, biosensing, cell culture, regenerative medicine and antibacterial coatings, *Chem. Rev.* (2014) **114**:10976–11026.

DIAZONIUM CHEMISTRY AS A ROBUST APPROACH FOR THE BIOFUNCTIONALIZATION OF TITANIUM SURFACE

Abstract

Due to an excellent corrosion resistance and high strength-to-density ratio, titanium is most commonly used as an implant material for bone replacement or support. Surface properties of Ti implants, including surface chemistry, topography and wetting angle, have a decisive effect on the osseointegration, and can be easily tailored by surface modification. In this chapter, we present the process of surface modification of Ti with a 4-nitrobenzenediazonium salt, resulting in the formation of a layer of nitrobenzene molecules covalently attached to the surface at para position. To extend the number of possible biofunctionalization procedures, as-formed surface was subjected to an additional electrochemical reduction reaction, resulting in the reduction of nitrobenzene moiety to aniline. To verify whether the functionalization procedure was successful, unmodified and modified Ti discs were investigated by means of electrochemical impedance spectroscopy.

Keywords: diazonium salts, electrografting, titanium, biofunctionalization

DOI:10.34918/85102

Kornelia JANKOWSKA[1], Marek J. ŁOS[1,*], Andrzej HUDECKI[2], Jolanta HYBIAK[1]

Chapter 2. ADSC DIFFERENTIATION INTO OSTEOBLASTS IN THE PRESENCE OF TISSUE SCAFFOLDS OBTAINED FROM CORE-SHELL MICRO- AND NANOFIBERS WITH ADDITION OF Cu

2.1. Introduction

Biomaterials in regenerative medicine have a variety of use, but the most common is to serve as scaffolds. They are providing physical support, delivery of cells and mobilization of endogenous cells to repair tissues. Thanks to this they play a key role in the reconstruction and replacement of hard to regenerate tissues lost due to trauma or disease [1].

Amongst the most challenging to heal injuries are large bone defects as they are beyond bone's capacity to completely regenerate without medical ingerence. Biomaterials in those defects provide required stabilization and support allowing bones to restore their function. Currently, the gold standards of treatment are bone autografts and allografts. However, they are burdened with numerous disadvantages, such as limited sources, risk of immune rejection or disease transmission [2]. To increase the benefit/risk ratio for the recipient currently used graft could be replaced with polymer scaffolds. In addition to reducing the risk of immune response, lowering the cost of procedure and their overall conveniency they bring more treatment strategies. They allow for performing autologous cell stimulations, stem cell implantation and differentiation [3].

One of the most widely use biomaterial in tissue engineering is polycaprolactone (PCL). It is synthetic polymer characterized by high strength, relatively long

---

[1] Department of Pathology, Pomeranian Medical University, Szczecin, Poland.
* Corresponding author: mjelos@gmail.com
[2] Łukasiewicz Research Network – Institute of Non-Ferrous Metals, Gliwice, Poland.

degradation time, good biocompatibility, and relatively low cost of production. Its properties make it a great candidate to produce improved cell scaffolds. Moreover, with carefully selected additives its features can be easily enhanced making the therapy more successful and convenient for patient [4].

Lately copper has caught the attention of many scientists as a potential additive to biomaterials. Its antibacterial properties may prevent post implant sepsis, which also leads to limitation of antibiotics use during the recovery period. Furthermore, multiple studies show that copper promotes angiogenesis. It might contribute to the copper ability to stimulate osteogenesis, which was also show in some studies. Taking in consideration those properties copper seems to be a good candidate for addition to biomaterials especially those dedicated for bone regeneration [5, 6].

2.2. Material and methods

2.2.1. Biomaterial production

Biomaterials used in this study were produced in two stages. First the solutions for nanofibers were mixed, starting with combining tetrahydrofuran and dimethylformamide with 300 nm copper nanowires in concentration of 1% or 5% in ultrasonic homogenizer. To this solution polycaprolactone was added to obtain PCL/Cu solution. Further, obtained solution was combined with 17% polycaprolactone solution to obtain a core-shell nanofibers called PCL//PCL/Cu/1% or PCL//PCL/Cu/5%.

In the second stage obtained solutions were used to produce fibers using electrospinning technique. Fibers obtained in an electrostatic field falling on the collector surface formed a non-woven fabric. During the works, the conditions for the production of micro and nanofibers were established/optimized. The following manufacturing conditions were used in the preparation of the fibers: (i) solution flow: 10 ml/h; (ii) collector type used: rotary, (iii) collector rotational speed 20 rpm, (iv) voltage: 0.95-1.00 kV/cm; (v) temperature of the solution 24°C, (vi) temperature of the gas (atmospheric air) flowing through the working chamber: 23°C; (vii) gas humidity in the working chamber: 35%.

2.2.2. SEM pictures

The SEM Scanning Electron Microscope was used to study the structure of micro and nanofibers. The samples were tested at a magnification of 5000 to 10000 x. As the tested samples are mostly made of polymers, before testing, the samples were placed in a sputtering machine where, after creating a vacuum, an electrically conductive silver coating with a thickness of several nm was applied. After the sputtering process, the samples were placed in the SEM microscope and the structure research was started.

2.2.3. Cell culture

Experiments were conducted using Adipose Derived Stem Cells (ADSC) which were cultured in Dulbecco's Modified Eagle Medium (DMEM, Lonza 12-604F/U1) with 10% Fetal Bovine Serum (FBS, Corning 35-079-CV) and 1% Antibiotic-Antimycotic (AA, Gibco 15240-062) and incubated at 37°C, 5% CO_2.

2.2.4. Biomaterial extracts preparation

Biomaterials were sterilized by 1h incubation in 70% ethanol, which was followed by rinsing with Phosphate Buffered Saline (PBS, Lonza 0000708833). Biomaterials were left to completely dry for 2-3 h and then incubated for 24 h in standard cell culture medium (DMEM 10% FBS 1% AA) at 37°C, 5% CO_2.

2.2.5. MTT Assay

For MTT test ADSC were seeded in 96-well plate and incubated for 24 hours. After the time cells medium was replaced with fresh portion of the standard medium, biomaterial extracts in concentration of 100% and 50% and 5% DMSO (Corning, 25-950-CQC) solution. The next day media was changed to MTT solution with which cells were incubated for 2 hours. The MTT solution has been prepared by dilution of 3 mg Thiazozyl Blue Tetrazolium Bromide (Sigma, MKCK7253) in 3 ml DMEM without phenol red (Gibco, 2036286). The resulting formazan crystals were then dissolved with isopropanol. The absorbance was measured spectrophotometrically at 570 nm.

2.2.6. ADSC differentiation into osteoblasts

ADSC differentiation into osteoblasts was carried out by seeding cells into 12-well cell plate and incubation till the cell's confluency reached 80%. Then the medium was changed to Induction Medium with or without biomaterial extracts. Cell media was changed for fresh portions every 3-4 days for 3 weeks. The Induction Medium consist of: DMEM, 0.05 mM L-Ascorbic Acid 2-phosphate sesquimagnesium salt hydrate (Sigma, A8960), 0.1 μM Dexamethasone (Sigma, D4902), 10 mM *β*-glycerophosphate disodium salt hydrate (Sigma, G9422), 10% FBS, 1% AA. Cell morphology was observed daily.

2.2.7. Alizarin Red staining

To perform Alizarin Red staining cells were fixed with 2% formaldehyde and incubated with Alizarin Red solution (Millipore, 2003999) for 0.5 h, which was followed by rinsing with distillated water. Cells were observed using light microscope. To obtain quantitative results, Alizarin Red dye was extracted from cells by incubating them with acetic acid, scraping from the bottom of the well, which was followed by incubation at 80°C. Meanwhile the Alizarin Red standards were prepared. The absorbance was measured at 405 nm.

2.2.8. Statistical analyzes

The statistic evaluation was performed using R-studio. The data normality was tested by Shapiro-Wilk test, based on the obtained results Wilcoxon signed-rank test or Student's t-test were carried out.

2.3. Results

2.3.1. PCL//PCL/Cu/1% and PCL//PCL/Cu/5% fabrics

Uniform microfibers of PCL//PCL/Cu/1% and PCL//PCL/Cu/5% were obtained, which formed non-woven fabric (Fig. 1).

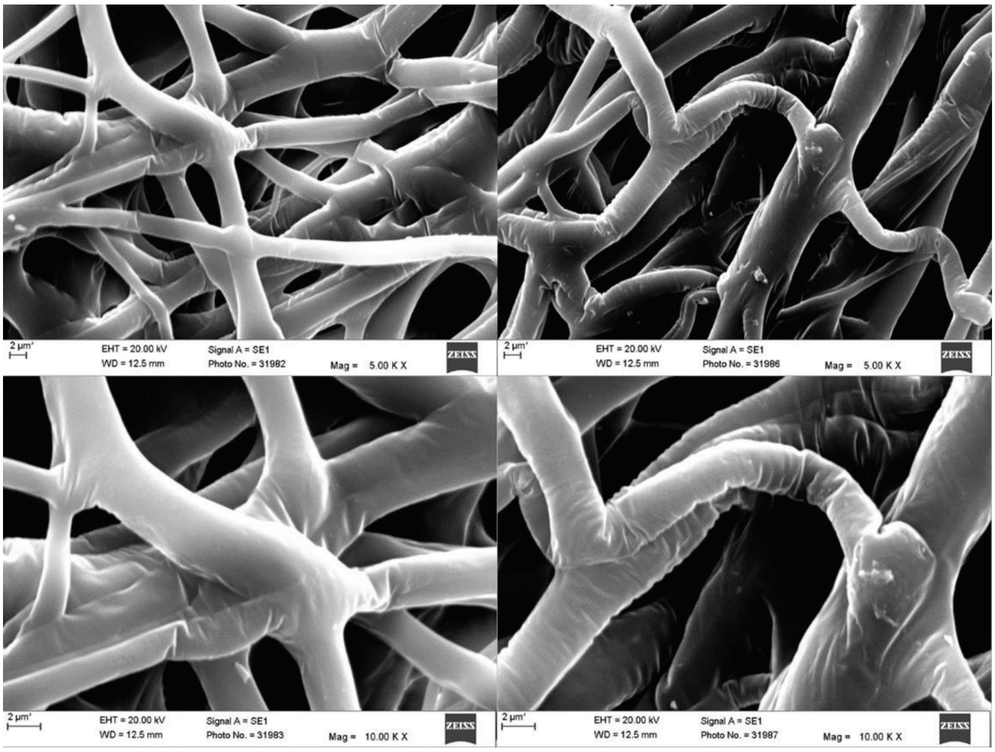


Fig. 1. The structure of the non-woven fabric. Images were taken using a scanning electron microscope (SEM)
Rys. 1. Struktura włókniny. Obrazy wykonano przy użyciu skaningowego mikroskopu elektronowego (SEM)

2.3.2. Biocompatibility of biomaterials extracts

The viability of cells incubated in biomaterials extracts was not markedly affected as at least 94.11% of cells remained viable upon testing (Fig. 2). In other variants of biomaterial extracts ADSC viability was higher than for cells incubated in standard cell culture medium reaching up to 119.13% for 50% PCL//PCL/Cu/5% extract and 111.7% for 100% PCL//PCL/Cu/1% extract.

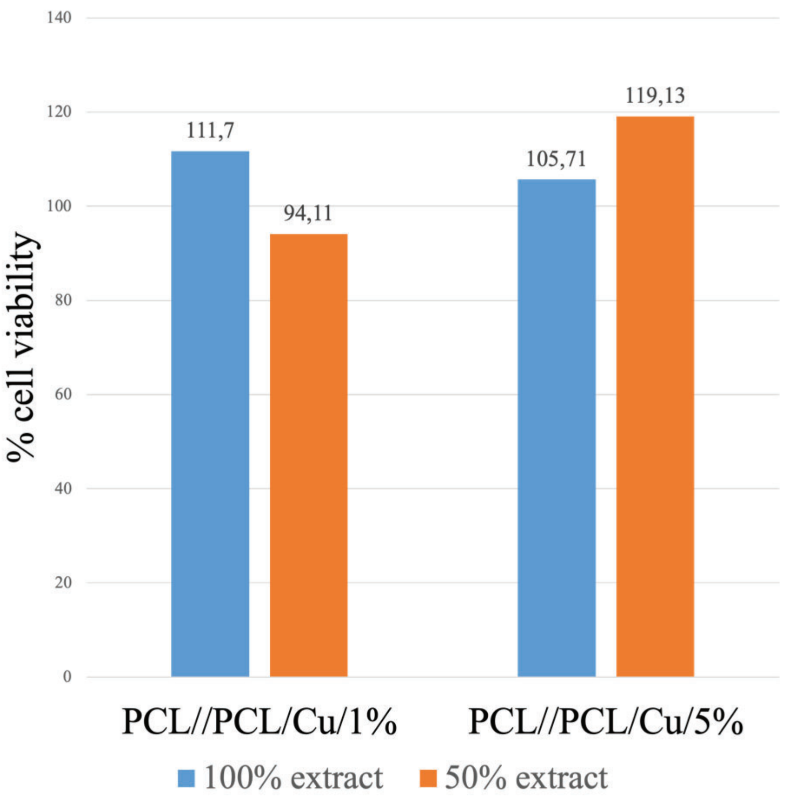


Fig. 2. Biocompatibility of PCL//PCL/Cu/1% and PCL//PCL/Cu/5% extracts
Rys. 2. Biokompatybilność ekstraktów z PCL//PCL/Cu/1% i PCL//PCL/Cu/5%

2.3.3. ADSC differentiation into osteoblasts

As a result of Alizarin Red staining, more stained calcium deposits were observed in the well, where cells were grown in the induction medium with the addition of biomaterials extracts than in wells where cells were incubated in induction medium without extracts (Fig. 3). For cells incubated with PCL//PCL/Cu/5% extract the amount of stained calcium deposits was higher than for cells incubated with PCL//PCL/Cu/1% extract. In well where cells were incubated in DMEM 10% FBS 1% AA no calcium deposits were observed.

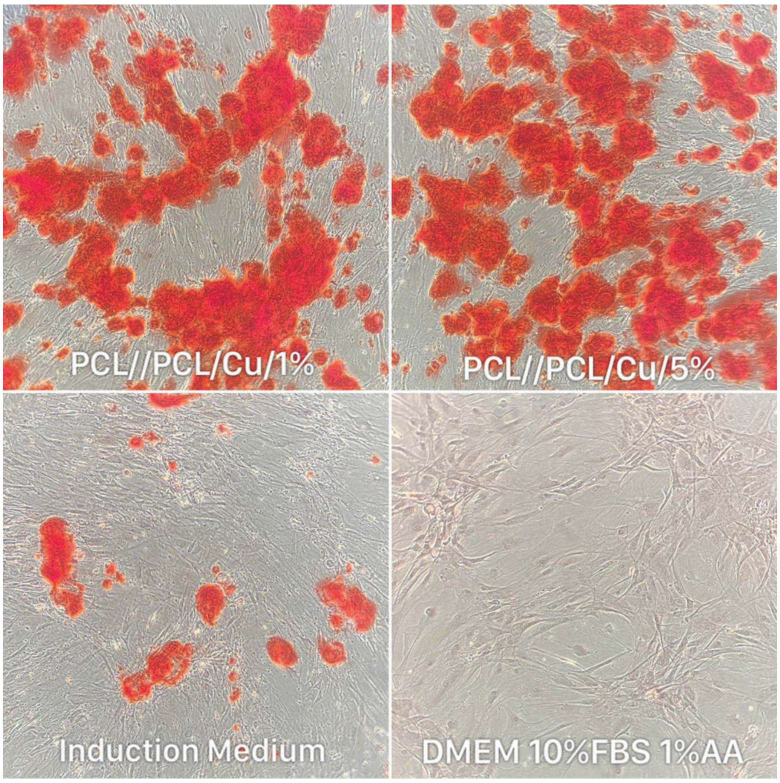


Fig. 3. Calcium deposits stained with Alizarin Red
Rys. 3. Złogi wapnia wybarwione Alizarin Red

The presence of copper significantly (p-value < 0.001) positively affected the differentiation ADSC cells into osteoblasts, as the staining intensity of Alizarin Red was three to four times higher for cells incubated with the addition of Cu-doped biomaterial extracts than for cells incubated without. Moreover, higher Alizarin Red dye concertation was observed in cells incubated with addition of PCL//PCL/Cu/5% extracts than for cells incubated with addition of PCL//PCL/Cu1% extract (Fig. 4).

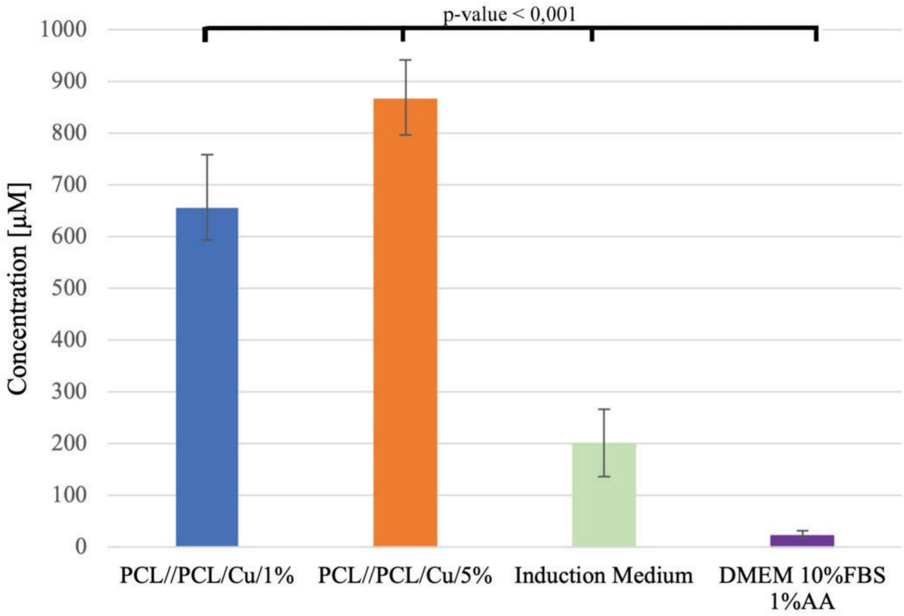


Fig. 4. The intensity of Alizarin Red staining which semi-quantitatively reflects upon the transdifferentiation of ADSC into osteocytes

Rys. 4. Intensywność barwienia Alizarin Red, która w sposób półilościowy odzwierciedla transdyferencjację ADSC do osteocytów

2.4. Discussion and concussion

Polycaprolactone is a well-known biomaterial that thanks to its desired properties found a variety of usage in regenerative medicine. Nonetheless, each new variant of it created by combining it with any additive must be tested for biocompatibility and changes in properties.

Based on performed analyzes it can be concluded that produced PCL//PCL/Cu/1% and PCL//PCL/Cu/5% extracts have no toxic effect on ADSC. Furthermore, the observed increased viability of cells suggests that the addition of copper may stimulate the proliferation of cells. This assumption can therefore be confirmed by the data available in the literature, where authors show that copper stimulates the proliferation of different cell types. Burghardt et al. found that copper in a specific concentration range stimulates the proliferation of mesenchymal stem cells [7]. Alizadeh et al., have shown that copper nanoparticles enhance endothelial cell proliferation and Philips et al. presented that copper cause an increase in dermal fibroblasts proliferation [8, 9].

ADSC differentiation into osteoblasts was performed based on methodology, which was established by combining two methods in order to achieve sufficient efficiency

[10, 11]. Obtained results suggest that the addition of biomaterial extracts significantly stimulates osteogenic differentiation. Taking into consideration numerous reports from the literature on the stimulating effect of copper on osteogenesis it can be concluded that an increased number of calcium deposits are caused due to the presence of copper [12–14]. Observed stimulation may be occurring because of copper's ability to enhance the alkaline phosphatase enzyme activity and osteogenic gene expressions [15, 16].

Taking into consideration all performed analyses it can be concluded that Cu-doped PCL may serve as a promising proto-material for the development of tissue structures facilitating the differentiation of ADSC into osteoblasts and it may find potential application in the production of scaffolds for bone regeneration.

Bibliography

1. K. Sadtler, A. Singh, M.T. Wolf, X. Wang, D.M. Pardoll, J.H. Elisseeff: Design, clinical translation and immunological response of biomaterials in regenerative medicine, Nature Reviews Materials (2016) 1(7):1–17.
2. M. Bahraminasab: Challenges on optimization of 3D-printed bone scaffolds, BioMedical Engineering Online (2020) 19(1):1–33.
3. M. Filippi, G. Born, M. Chaaban, A. Scherberich: Natural Polymeric Scaffolds in Bone Regeneration, Frontiers in Bioengineering and Biotechnology (2020) 8:474.
4. E.H. Backes et al.: Polycaprolactone usage in additive manufacturing strategies for tissue engineering applications: A review, J Biomed Mater Res B Appl Biomater (2022) 110(6):1479–1503.
5. M. Abudhahir et al.: Polycaprolactone fibrous electrospun scaffolds reinforced with copper doped wollastonite for bone tissue engineering applications, J Biomed Mater Res B Appl Biomater (2021) 109(5):654–664.
6. P. Wang et al.: Biological applications of copper-containing materials, Bioactive Materials (2021) 6(4):916–927.
7. I. Burghardt et al.: A dual function of copper in designing regenerative implants, Biomaterials (2015) 44:36–44.
8. S. Alizadeh, B. Seyedalipour, S. Shafieyan, A. Kheime, P. Mohammadi, N. Aghdami: Copper nanoparticles promote rapid wound healing in acute full thickness defect via acceleration of skin cell migration, proliferation, and neovascularization, Biochem Biophys Res Commun (2019) 517(4):684–690.

9. N. Philips, H. Hwang, S. Chauhan, D. Leonardi, S. Gonzalez: Stimulation of cell proliferation and expression of matrixmetalloproteinase-1 and interluekin-8 genes in dermal fibroblasts by copper, Connect Tissue Res (2010) 51(3):224–229.
10. A. Skubis-Sikora, B. Sikora, A. Witkowska, U. Mazurek, J. Gola: Osteogenesis of adipose-derived stem cells from patients with glucose metabolism disorders, Molecular Medicine (2020) 26(1):1–9.
11. Y. Yoshimura et al.: How much medium do you use for cell culture? Medium volume influences mineralization and osteoclastogenesis in vitro, Molecular Medicine Reports, (2017) 16(1):429–434.
12. A. Ewald, C. Käppel, E. Vorndran, C. Moseke, M. Gelinsky, U. Gbureck: The effect of Cu(II)-loaded brushite scaffolds on growth and activity of osteoblastic cells, J Biomed Mater Res A (2012) 100(9):2392–2400.
13. J.P. Rodríguez, S. Ríos, M. González: Modulation of the proliferation and differentiation of human mesenchymal stem cells by copper., J Cell Biochem (2002) 85(1): 92–100.
14. W. Zhang et al.: Graphene Oxide-Copper Nanocomposite-Coated Porous CaP Scaffold for Vascularized Bone Regeneration via Activation of Hif-1α, Adv Healthc Mater (2016) 5(11): 1299–1309.
15. L. Ren, H.M. Wong, C.H. Yan, K.W.K. Yeung, K. Yang: Osteogenic ability of Cu-bearing stainless steel, J Biomed Mater Res B Appl Biomater (2015) 103(7):1433–1444.
16. Y. Yuan et al.: Osteogenesis stimulation by copper-containing 316L stainless steel via activation of akt cell signaling pathway and Runx2 upregulation, Journal of Materials Sciences and Technology (2019) 35(11):2727.

ADSC DIFFERENTIATION INTO OSTEOBLASTS IN THE PRESENCE OF TISSUE SCAFFOLDS OBTAINED FROM CORE-SHELL MICRO- AND NANOFIBERS WITH ADDITION OF Cu

Abstract

Polycaprolactone (PCL) is a biodegradable polymer widely used in medicine as a biomaterial for tissue regeneration. This biomaterial is characterized by high strength, relatively long degradation time and good biocompatibility. Copper in pure form is highly cytotoxic, but at the same time, it is an element with very good electrical and thermal conductivity. In this study, the PLC polymer was combined with copper nanowires and transformed into core-shell fibers upon application of electro-spinning. The end-product formed a highly porous sheet-like structure suitable for serving as, an artificial fibrous cell scaffold. Obtained fibrous scaffolds were examined for the influence on the differentiation of ADSC into osteoblasts.

Here we show shown that: (i) the viability of cells incubated in biomaterial extracts was not markedly affected as at least 94.11% of cells remained viable upon testing, (ii) the presence of copper significantly (p-value < 0.001) positively affected the differentiation ADSC cells into osteoblasts, (iii) the staining intensity of Alizarin Red was three to four times higher for cells incubated with the addition of Cu-doped biomaterial extracts than for cells incubated without.

Based on the above results, we conclude that Cu-doped PLC may serve as a promising proto-material for the development tissue-structures facilitating the differentiation of ADSC into osteoblasts.

Keywords: polycaprolactone, differentiation, core-shell fibers, copper

DOI:10.34918/85103

Daria KOSTKA[1,2,*] Agnieszka GOGLER-PIGŁOWSKA[2], Wiktoria PŁONKA[1],
Damian SOJKA[2], Małgorzata ADAMIEC-ORGANIŚCIOK[2],
Magdalena SKONIECZNA[1], Dorota ŚCIEGLIŃSKA[2,*]

Chapter 3. A TOOL FOR QUANTIFYING THE LEVEL OF ANTIGEN (KERATIN 10) DETERMINED BY THE IHC METHOD IN HUMAN EPIDERMAL EQUIVALENTS OBTAINED IN VITRO

3.1. Introduction

The 3D organotypic co-culture system is an up-to-date culture technique that enables generation of the reconstructed human epidermis (RHE) in vitro. This culture system is considered superior to standard two-dimensional in vitro culture of keratinocytes in plastic dishes. Keratinocytes when grown in air-liquid interface in 3D co-culture with dermal equivalent form complex tissue that replicates the dermal-epithelial crosstalk, polarized protein distribution and architectural features characteristic of each cell layer seen in normal human epidermis in vivo. Despite some drawbacks (lack of immune system cells and melanocytes), the application of RHE model have enabled to address previously inaccessible questions on human keratinocyte differentiation and epidermal regeneration. Furthermore 3D organotypic co-culture method is very powerful when combined with methods of genetic manipulations of keratinocytes such as genes knockout (deletion), knockdown (downregulation) or knockin (insertion). Such a combination provides a strong platform for investigating how the integrity, structure and functions of human epidermis are connected to protein function.

---

[1] Department of Systems Biology and Engineering and Biotechnology Centre, Silesian University of Technology, Gliwice, Poland.

[*] Corresponding author: darikos556@student.polsl.pl

[2] Center for Translational Research and Molecular Biology of Cancer, Maria Skłodowska-Curie National Research Institute of Oncology Gliwice Branch, 44-102 Gliwice, Poland.

[*] Corresponding author: dorota.scieglinska@io.gliwice.pl

One of the steps in the algorithm for evaluation of genetically modified keratinocytes in RHE model is the detection of keratinocyte differentiation markers. This task is performed to gain knowledge on how a particular genetic modification made in cells effects on their potential to differentiate and generate full-length RHE. Alterations in differentiation marker(s) location or changes in their expression levels in RHE model usually represent changes in epithelial stratification and reflect aberrant keratinocyte differentiation. Analysis of differentiation markers expression is routinely executed using immunohistochemistry (IHC) method in order to detect some of marker proteins that are specific for an early, moderate or late keratinocyte differentiation stages. However quantitative analysis of the IHC results in RHE samples can be a challenging task. One of the available IHC staining analysis programs is Fiji with the Immunohistochemistry (IHC) Image Analysis Toolbox. This program was designed for the analysis of images, where the specimen covers the entire photo area. In case of RHE, the specimen occupies a small part of the image. This fact may have significant effect on results as the background signal migh biase the IHC staining signal quantitation. Therefore, this study was undertaken to adress abovementioned issue and develop a tool that could be used to quantify results of IHC detection of antigens in RHE samples.

3.1.1. The structure and functions of the skin and the epidermis

The skin is the most external and one of the largest human organs. Human skin is composed of three layers: epidermis, dermis and subcutaneous tissue. It consists of glands, hair, nails, blood vessels, lymphatic and nerve endings [1]. The function of the skin is to protect the inside of the body against the influence of the environment, physical and chemical damages, infections, UV radiation [1, 2]. This organ is also required to maintain a stable condition of the internal body environment, protect against dehydration and regulate body temperature [3].

The epidermis, a type of stratified squamous keratinizing epithelium, constitutes the outermost layer of the skin. It is composed mainly of keratinocytes organized in hierarchical manner. Stem cells building the top of hierarchy are located in the basal layer, their progeny differentiate along with moving towards the epidermal surface. During epidermal differentiation keratinocytes' morphology transits from large cells with large nuclei of the basal layer through polygonal cells with large multi-lobed nuclei of the spinous layer into spindle-shaped flattened cells of the granular layer and ultimately into enucleated spindle-shaped corneocytes of the stratum corneum [1, 4]. Corneocytes are dead cells filled with keratin fibers embedded in the filaggrin matrix. These cells are built into lipid matrix that form the extracellular shell termed the corneocyte envelope. The lipid coat on the cells' surface is the essential component of the epidermal permeability barrier that protects the body from excessive water loss and absorption of exogenous substances [1, 4].

Epidermal differentiation is a cycle of transformation that lasts from 26 to 42 days. During homeostasis, keratinocytes differentiate from basal proliferating cells to nuclear-depleted squamous corneocytes. The process of keratinocytes differentiation requires the activation of the genes that are implicated in functional and structural cell transformation. Genes related to the differentiation process are concentrated in the Epidermal Differentiation Complex (EDC) [2, 5, 6].

The characteristic components of the epidermis are keratins, a fibrillar protein that builds intermediate filaments. Some of them are considered as differentiation markers in the epidermal keratinocytes. Keratin 10 (K10) represents an early differentiation marker that is expressed in keratinocytes located in the spinous layer. As keratinocytes leave the basal layer and pass to the spinous layer, they inhibit the production of keratin 5 and keratin 14, and activate the production of keratin 1 (K1) and K10. The expression of K1 and K10 is activated under the conditions of the increasing Ca^{2+} concentration [7]. These structural proteins are crucial for proper function and structure of epidermis. For example removal of the K10 gene in mouse model resulted in skin fragility similar to that of patients suffering from epidermolytic hyperkeratosis [8]. The keratinocytes in the upper spinous layer also produce another differentiation markers such as involucrin, a component of the corneocyte envelope, and transglutaminase K, an enzyme that catalyzes a critical step in the formation of the cornified envelope of terminal differentiation [9, 10, 11].

3.1.2. Immunohistochemical reactions

Immunohistochemistry (IHC) is a standard technique that employs antibodies to detect and quantify a specific protein (antigen) within a given tissue. The method is designed in a way that allows microscopic observation of the antigen-antibody complexes. The antigen is specifically recognized by a primary antibody. The detection of a primary antibody-antigen complex in an IHC experiment can be either direct or indirect. Direct detection methods employ primary antibody directly conjugated to a label (dye or enzyme) [12]. Indirect reactions employs additional usage of a secondary antibody that is labelled and directed against the host species of the primary antibody (unlabeled). Indirect reactions are usually used to increase signal (staining) intensity [12].

The color effects in IHC are frequently generated using antibody (primary or secondary) labelled with an enzyme such as the horse radish peroxidase (HRP). Addition of a specific substrate (i.e. a chromogen) gives a colored product. One of the most popular chromogens is 3,3-diaminobenzidine tetrahydrochloride called DAB. To act as a dye, DAB is oxidised and converted to an insoluble polymer, which precipitates as a dark brown pigment at the reaction site allowing visualization of the target molecule [13, 14].

3.1.3. The role of Heat Shock Protein A2 (HSPA2) in the epidermis

HSPA2 belongs to the multigene *HSPA (HSP70)* family of chaperone proteins playing important roles in the maintenance of cellular proteostasis. HSPA2 was initially described as a testis-specific protein, which is crucial for spermatogenesis and male fertility [15]. However, as we showed previously, HSPA2 is also present in selected human somatic tissues including stratified and pseudostratified epithelia. In the epidermis HSPA2 accumulates in basal keratinocytes [16]. The role of HSPA2 in the epidermis and also in other epithelia is poorly understood.

In our preliminary study we showed that HSPA2 can regulate keratinocyte differentiation. We observed that decrease in the HSPA2 protein levels in keratinocytes, resulting from partial inhibition of the *HSPA2* gene expression by the RNAi mechanism, resulted in a more differentiated but still proliferative cell phenotype [17]. In order to examine the effect of a total lack of HSPA2 on keratinocyte differentiation we generated CRISPR/Cas9-mediated knockout of the *HSPA2* gene expression in immortalized human epiderml keratinocyte HaCaT line. Histological and IHC evaluation showed altered stratification of RHE generated from HSPA2-null cells. Among other altered pattern of immunostaining with approved differentiation markers such as K10 and others occured in HSPA2-null RHE. Therefore, semiquantitive analysis of differentiation markers expression would help to describe in more details how a lack of HSPA2 impairs epidermis stratification in RHE model. Thus in this work we aimed at developing a tool that could be useful to quantify results of IHC detection of keratinocyte differentiation markers in RHE samples.

3.2. Materials and methods

The reconstructed human epidermis (RHE) samples were generated from spontaneously immortalized epidermal keratinocyte HaCaT line. We employed CRISPR/Cas9-mediated genetic modification system to generate cells that differed in the *HSPA2* gene expression levels. The HSPA2 protein was absent in CRISPR-A2.2 cells (gene knockout). Normal endogenous levels of HSPA2 were present in CRISPR--CTR control cells. RHE samples were generated in vitro according to protocol described in Gogler-Piglowska 2018 [17]. RHE were fixed in formalin, embedded in paraffin according to standard protocol [17]. The 5 µm thick RHE sections were cut, placed on glass slides, deparaffinized and processed for immunohistochemistry [15]. Endogenous peroxidase activity was blocked for 10 min in 1.5% perhydrol solution. Next, samples were incubated for 45 minutes in 2.5% normal horse serum solution to block endogenous antigens. The ImmPRESS® HRP Universal Antibody (Horse Anti--Mouse/Rabbit IgG) Polymer Detection Kit, Peroxidase, (Vector Laboratories, Burlingame, California, nr MP-7500) was used for detection of antigen-antibody complex accoring to manufacturere's instruction.

The samples were incubated with the primary anty-K10 antibody (BioLegend, San Diego, California, nr 905401) at 1:3000 dilution at 4°C overnight. DAB chromogen was used to develop IHC staining. Samples were inculated with DAB substrate for 3 minutes at room temperature, and then were counter-stained with hematoxylin for 1 min.

IHC staining was observed using a ZEISS AXIOPHOT light microscope at 200x magnification (10x eyepiece, 20x lens), the images were taken using a Zeiss Axiocam 503 color camera and a ZEN 2.6 photo archiving system.

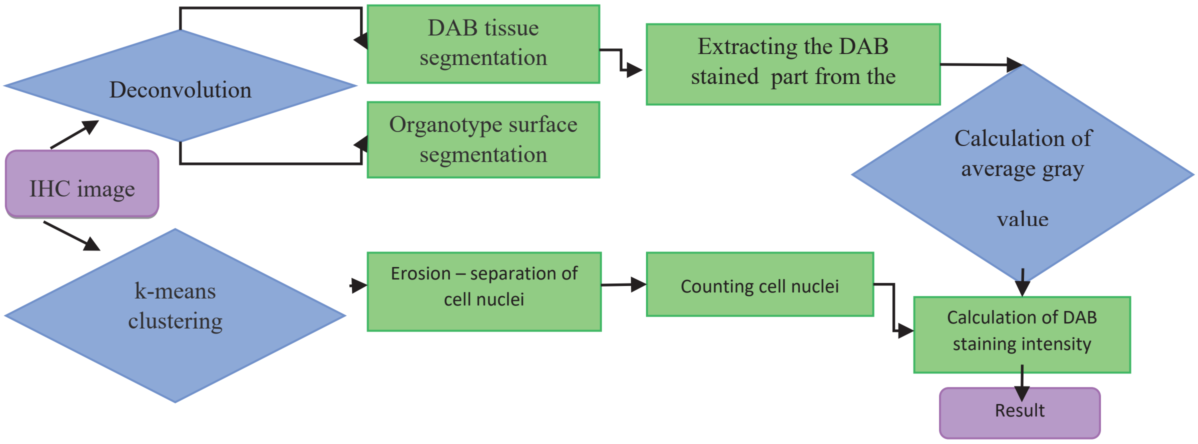


Fig. 1. Flowchart of the IHC quantification program. The program was developed to quantitatively analyze images from IHC staining. Main steps of the algorithm include color image deconvolution of double-stained tissue, calculation the average intensity of pixels in segmented DAB-stained tissue, K-means clustering for nuclei segmentation, counting cell nuclei and calculation the DAB staining intensity normalized by the number of nuclei. The aim of this algorithm is to get quantitative determination of protein expression from IHC images

Rys. 1. Schemat blokowy programu ilościowego IHC. Program został opracowany do ilościowej analizy obrazów z barwienia IHC. Główne etapy algorytmu obejmują dekonwolucję obrazu kolorowego podwójnie zabarwionej tkanki, obliczenie średniej intensywności pikseli w segmentowanej tkance zabarwionej DAB, grupowanie K-średnich dla segmentacji jąder, zliczanie jąder komórkowych i obliczanie intensywności barwienia DAB znormalizowanej przez liczbę jądra. Celem tego algorytmu jest ilościowe określenie ekspresji białek na podstawie obrazów IHC

3.3. Results and discussion

3.3.1. Image Processing

The first step of the algorithm were normalization and deconvolution of images. Image deconvolution is a technique that relies on transformation of color images of multiple-stained tissue (Fig. 2A) into images representing one stain concentration (Fig. 2B, C). The IHC-generated images consisted of three different staining/convolutions: DAB signal, hematoxylin signal and DAB/hematoxylin interference signal. Our aim was to breaking down these tangles without damaging individual signals. Values of parameters P1, P2 and P3 necessary to separate the information contained in the images were taken from the Fiji program for the DAB option (Table 1).

Table 1

Values of parameters obtained from the Fiji program

Signal	P1	P2	P3
Hematoxylin	0.6500286	0.704031	0.2860126
DAB	0.26814753	0.57031375	0.77642715
Residual	0.7110272	0.42318153	0.5615672

The next step was to normalize these values using vector normalization from the group of one signal. Then, a matrix was made of them and used for splitting the signals. Tissue surface was segmented and the average value of pixels intensity was counted. This analysis was conducted using picture in gray scale (Fig. 2).

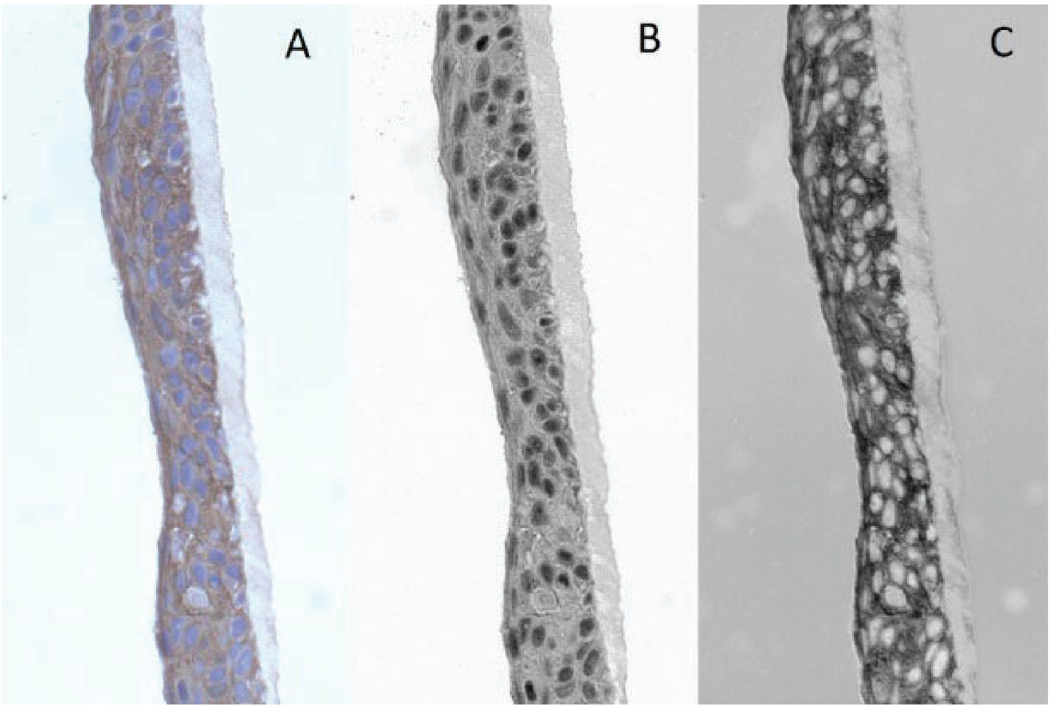


Fig. 2. Deconvolution of IHC image. First step of IHC image processing is deconvolution of signals contained in color IHC image (A). The images obtained after dyeing consisted of three convolutions: DAB signal, hemaotoxylin signal and interference. Implemented program gives individual signals for: nuclei stained with hematoxylin (B) and antigen stained with DAB (C)

Rys. 2. Dekonwolucja obrazu IHC. Pierwszym etapem przetwarzania obrazu IHC jest dekonwolucja sygnałów zawartych w kolorowym obrazie IHC (A). Obrazy otrzymane po barwieniu składały się z trzech splotów: sygnału DAB, sygnału pochodzącego z hematoksyliny i interferencji. Zaimplementowany program daje nam indywidualne sygnały: barwienia hematoksyliną jąder komórkowych (B) I barwienia DAB (C)

The next step was counting average value of pixels intensity in DAB signal area. To get an image of DAB-stained tissue picture containing signal of DAB staining (Fig. 3B) were converted to black and white pixels by image thresholding (Fig. 3A). The next step was multiplication of original image which represent the DAB signal (Fig. 3B) by the black and white mask (Fig. 3A). As a result we get segmented area of DAB signal in grayscale color (Fig. 3B) and this picture was used to farther analysis.

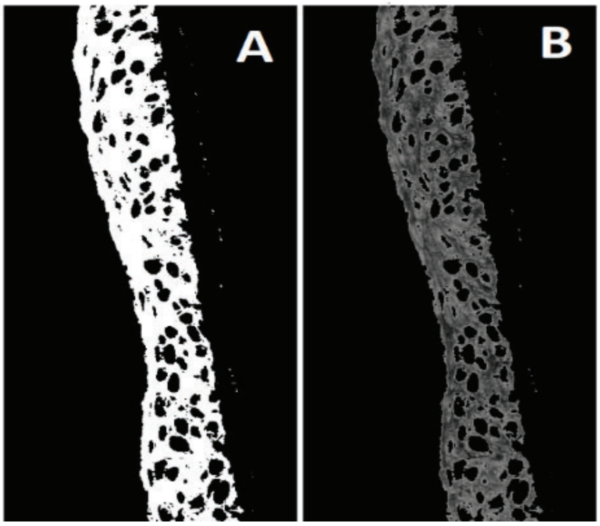


Fig. 3. Segmented area of DAB signal. First to get this result program takes DAB signal and IHC image is converted to black and white pixels by image thresholding (A). Next original image which represent the DAB signal is multiplied by the black and white mask. As a result we get segmented area of DAB signal in grayscale color (B)

Rys. 3. Segmentowany obszar sygnału DAB. Jako pierwszy, aby uzyskać ten wynik, program pobiera sygnał DAB, a obraz IHC jest konwertowany na czarno-białe piksele przez progowanie obrazu (A). Następny oryginalny obraz reprezentujący sygnał DAB jest mnożony przez czarno-białą maskę. W rezultacie otrzymujemy segmentowany obszar sygnału DAB w kolorze skali szarości (B)

Very important part of the image processing workflow was nucleus segmentation and counting. For segmentation purposes k-means clustering was used, which allows to isolate objects by assigning them to one of k groups. The result was an image with 3 colors: black, white and gray. We used three different masks in this step (Fig. 4). Each picture has different clustering result, so the program automatically selects the mask for further analysis.

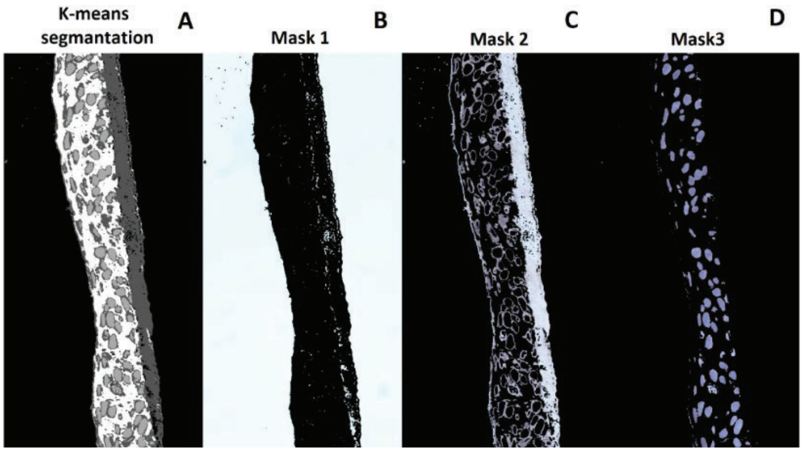


Fig. 4. Results of K-means segmentation and used masks. K-means segmentation was performed using original RGB image. The result was an image with 3 colors: black, white and gray (A). Next we used three different masks (B, C, D). Each picture represent different clustering result. Mask of interest is the one that gives us an image containing blue nuclei (Mask 3). These steps were necessary to allow nucleus counting

Rys. 4. Wyniki segmentacji K-średnich i stosowanych masek. Segmentację K-średnich przeprowadzono przy użyciu oryginalnego obrazu RGB. W rezultacie powstał obraz w 3 kolorach: czarnym, białym i szarym (A). Następnie użyliśmy trzech różnych masek (B, C, D). Każdy obraz przedstawia inny wynik grupowania. Interesująca maska to taka, która daje nam obraz zawierający niebieskie jądra (Maska 3). Te kroki były konieczne, aby umożliwić liczenie jąder

Mask of interest is the one that gives an image containing blue nuclei (Fig. 4D). Next step is getting rid of artifacts. Single pixels also can interfere with the results. The size of the nuclei was irrelevant for analysis, so the solution was the multiple erosion and dilatation of objects (nuclei). The results of the algorithm with the selected nuclei were then visualized (Fig. 5).

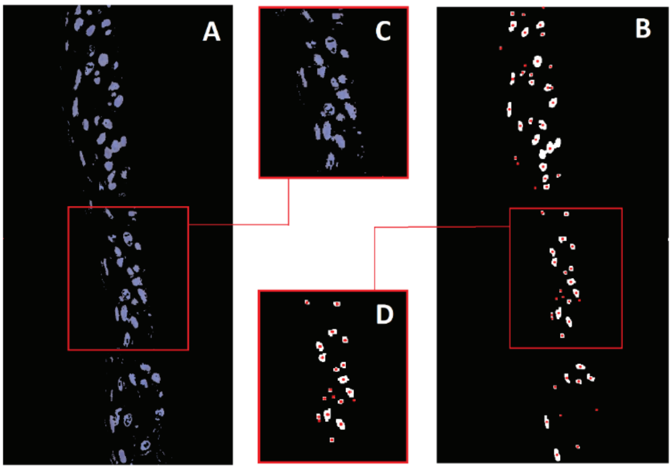


Fig. 5. Selected nuclei. After clustering of image, which we get as result of k-means segmentation, program automatically selected the picture useful for nuclei counting (A). Next image is converted to black and white pixels and each white shape is detected by the algorithm (B). To better visualization enlarged section have also been inserted (C, D)

Rys. 5. Wybrane jądra. Po zgrupowaniu obrazu, które otrzymamy w wyniku segmentacji k-średnich, program automatycznie wybrał obraz przydatny do zliczania jąder (A). Następny obraz jest konwertowany na czarno-białe piksele, a każdy biały kształt jest wykrywany przez algorytm (B). Dla lepszej wizualizacji wstawiono również powiększony przekrój (C, D)

The last step of algorithm is counting intensity of DAB staining normalized by the number of nuclei. Immunohistochemical staining showing the presence of the K10 protein in three-dimensional equivalents of the epidermis obtained from HaCaT cells that differed in levels of the HSPA2 chaperone. The protein was absent in CRISPR-A2.2, while CRISPR-CTRL cells contained normal endogenous levels of HSPA2.

3.3.2. Analysis of DAB-mediated stainig of K10 in RHE samples

K10 staining, visible as a brown color, was observed in all examined RHE samples. The results of IHC detection of K10 in RHE corresponded to the typical pattern for K10 immunostainig in the human epidermis. As expected K10, as an early marker of keratinocyte differentiation, was detected in the cytoplasm of keratinocytes located in the suprabasal layers of RHE. However microscopic evaluation of K10-stained samples showed that K10 staining was clearly stronger in RHE generated from CRISPR-A2.2 cells without HSPA2 protein (Fig. 6) than from HSPA2-expressing CRISPR-CTRL cells.

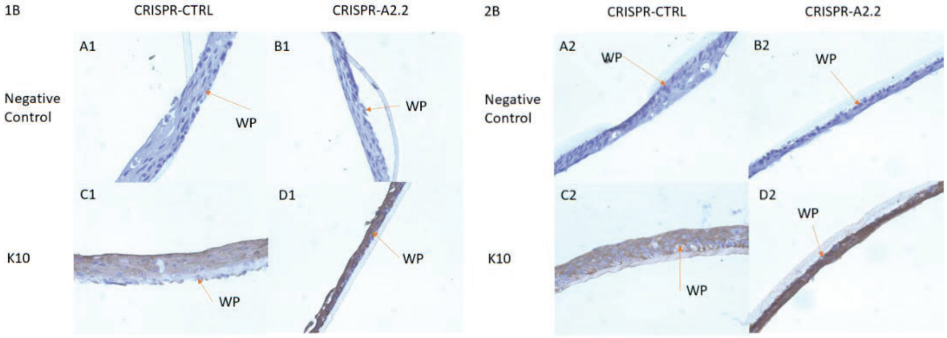


Fig. 6. Immunohistochemical staining showing the presence of K10 protein in RHE obtained from HaCaT cells with varying levels of the HSPA2 chaperone. 1B and 2B are RHE generated in two irrespective biological replicates. Negative control, images without DAB signal (lack of primary antibody) (A1, B1, A2, B2). K10, pictures showing DAB-mediated staining of K10 (C1, D1, C2, D2). CRISPR-CTRL, RHE generated from HSPA2-expressing cells (A1, C1, A2, C2); CRISPR-A2.2, RHE generated from HSPA2-knockout cells (B1, D1, B2, D2). WP- basement membrane. K10 immunoreactivity was clearly stronger in RHE obtained from cells without HSPA2 protein

Ryc. 6. Barwienie immunohistochemiczne wykazujące obecność białka K10 w 3D ekwiwalentach naskórka uzyskanych z komórek HaCaT o różnych poziomach białka opiekuńczego HSPA2. 1B i 2B to powtórzenia użyte w badaniach. WP- warstwa podstawna. Kontrola negatywna oznacza obrazy bez sygnału DAB (brak przeciwciała pierwszorzędowego) (A1, B1, A2, B2). K10 to obrazy z sygnałem DAB w miejscu, w którym występuje keratyna 10 (C1, D1, C2, D2). CRISPR-CTRL pokazuje tkankę z białkiem HSPA2 (A1, C1, A2, C2), a CRISPR-A2.2 pokazuje tkankę bez białka HSPA2 (B1, D1, B2, D2). Barwienie immunohistochemiczne wykazuje różnicę w natężeniu sygnału DAB (odpowiadającego na poziom K10) pomiędzy równoważnikami uzyskanymi z kontrolnych komórek HaCaT (CRISPR-CTRL) a tymi uzyskanymi z komórek z niedoborem białka HSPA2 (CRISPR-A2.2). Immunoreaktywność K10 była wyraźnie silniejsza w organotypach uzyskanych z komórek bez białka HSPA2

Therefore we used the algorithm for quantitative analysis of K10 staining to assess potential differences in K10 expression in quantitative manner. The algorithm for quantitative analysis of DAB-stained RHE was used for two irrespective biological repeats. The 1B repeat group consisted of 11 images taken for RHE generated from CRISPR-CTRL cells and 2 images taken for RHE generated of CRISPR-A2.2 cells. The 2B repeat group consisted of 4 images taken for CRISPR-CTRL and 5 images taken for CRISPR-A2.2 RHE. The mean value for each replicate was calculated. The differences in the intensity of K10 staining between CRISPR-CTRL and CRISPR-A2.2 groups were expressed relative to CRISPR-CTRLs. Results in Fig. 7 show that RHE generated from HSPA2-deficient cells (CRISPR-A2.2) expressed increased level of K10 when compared to RHE generated from CRISPR-CTRL control. This indicates that the lack of HSPA2 protein correlates with increase in K10 expression in RHE.

This results is consistent with our previous findings showing that lowering the level of the HSPA2 protein due to partial suppression of the *HSPA2* gene expression by the RNAi mechanism resulted in a set of keratinocyte phenotypic changes associated with differentiation [17]. The histological and ultrastructural research conducted in parallel with this engineering work have shown that HSPA2-null keratinocytes in comparison to control ones form definitely thinner RHE with residual granular layer (A. Gogler--Pigłowska, work in preparation). Thus, alterations in K10 expression can reflect impaired differentiation of HSPA2-null keratinocytes in RHE model. Therefore, these results confirm that HSPA2 can be an important regulator of keratinocyte differentiation in the epidermis. However, it can be also speculated that higher level of K10 in HSPA2-null RHE can be a direct effect of a total lack of HSPA2 chaperone activity such as abberant K10 folding, altered posstranslational modifications or decreased degradation. This possibilities can be examined in subsequent study.

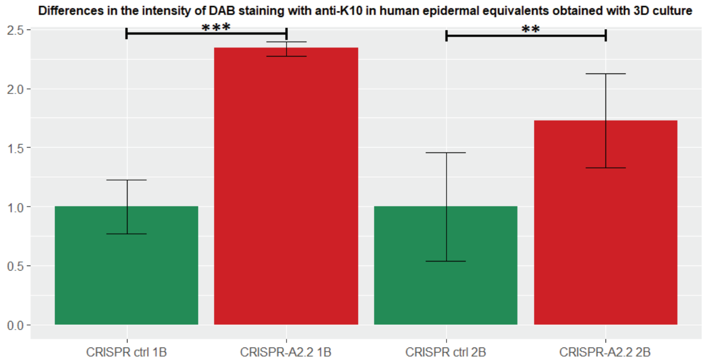


Fig. 7. Differences in the intensity of DAB signal in RHE samples incubated with anti-K10 antibody. The diagram shows results of the program operation for CRISPR-CTRL, with HSPA2 protein CRISPR-A2.2, without HSPA2 protein. Significance was calculated using a one-tailed t-test. For the 1B repeat p-value was 0.00052 and for 2B repeat – 0.0016

Rys. 7. Różnice w intensywności barwienia DAB anty-K10 w ludzkich odpowiednikach naskórka uzyskanych z hodowli 3D. Diagram przedstawia wyniki działania programu dla CRISPR-CTRL, z białkiem HSPA2 i CRISPR-A2.2, bez białka HSPA2. Średnia zmiana w poziomach K10 została obliczona w stosunku do CRISPR-CTRLs, które ustawiono na 1. Jak widzimy, poziomy K10 wzrosły w porównaniu z kontrolą około 2 razy. Istotność obliczono stosując jednostronny test t. Dla powtórzenia 1B p-wartość wyniosła 0.00052, a dla powtórzenia 2B – 0.0016

3.3.3. Evaluation of the possibility of using the algorithm on unstained samples

The program was validated using tissue pictures without DAB staining and without the use of a primary antibody (Fig. 8). When deconvolving pictures without brown DAB response, the signals were split incorrectly. The DAB signal was not visible (the picture is uniformly light gray). This circumstance disturbs the next steps of the algorithm. The aim of the pictures representing DAB staining is to isolate the stained tissue from the image. In a case when DAB signal is absent the program segments the background (Fig. 8 E). It causes an error during calculation of average gray intensity of the area. In the end we obtain very high values of “normalized intensity of DAB staining”, which we can reject basis after inspecting the received images.

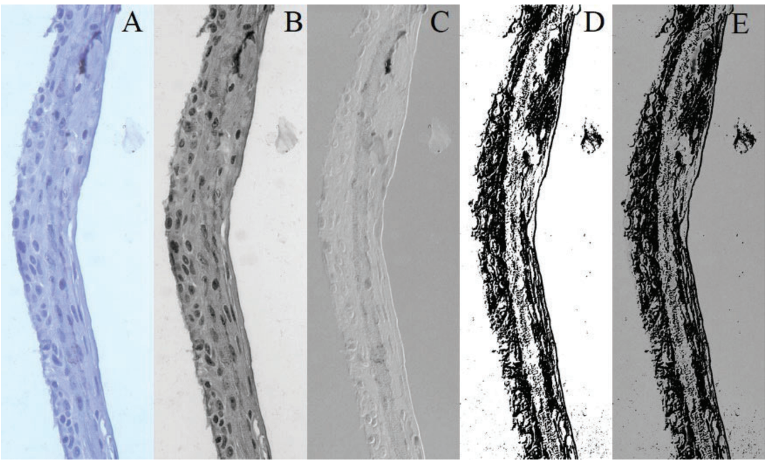


Fig. 8. Results of processing image without DAB signal. The presented images were obtained as a result of the analysis of the image showing no brown color (tissue without primary antibody). We only see purple hematoxylin (A). The C image should show the DAB signal, but it was not present on the specimen, so we cannot see it. This is correct. The presented formulation comes from repeat 1B. Image A shows the original photo. Picture B is the hematoxylin signal, picture C is the DAB signal. In image D we can see image C segmentation and background detail, and in image E grayscale has been superimposed over it

Rys. 8. Wyniki przetwarzania obrazu bez sygnału DAB. Prezentowane obrazy uzyskano w wyniku analizy obrazu nie wykazującego brązowego zabarwienia (preparate bez użycia przeciwciała pierwszorzędowego). Widzimy tylko fioletową hematoksylinę (A). Obraz C powinien pokazywać sygnał DAB, ale na preparacie go nie było, więc go nie widzimy. Prezentowany preparat pochodzi z powtórzenia 1B. Zdjęcie A przedstawia oryginalne zdjęcie. Obraz B to sygnał hematoksyliny, obraz C to sygnał DAB. Na obrazie D widzimy segmentację obrazu C i szczegóły tła, a na obrazie nałożono na niego skalę szarości

3.4. Summary

The created program allowed for quantitative analysis of DAB-mediated IHC staining in samples of reconstructed human epidermis in vitro that occupy only a small area of the image. In this case the freewere Fiji program with the Immunohistochemistry (IHC) Image Analysis Toolbox- IHC_Toolbox.jar extension showed a poor performance in image analysis. The main limitations was that the whole image (the specimen and background) was analysed by the Fiji program as the test specimen. Thus, Fiji program generated erroneous results when tissues occupying only a fragment of the image surface were analyzed. Our program allows to overcome this limitation and provided the tool to calculate DAB staining intensity in tissues that do not cover the entire surface of the image taken. Also, our program accelerates the analysis of the long and narrow specimens (such as the epidermis) that cover only a part of image because it eliminates a need for cutting the specimen-covered area off the each individual photograph. This initial and time-consuming step in image analysis pipeline with Fiji program is hard to accept when a large set of DAB-stained specimens need to be analysed.

Acknowledgement

Many thanks to Professor Dorota Ścieglińska and her team consisting of Dr. Agnieszka Gogler-Pigłowska, Dr. Damian Sojka, MSc Małgorzata Adamiec-Organiściok, who made it possible to perform the scientific research, and also to Professor Magdalena Skonieczna, thanks to whom I learned many laboratory techniques. Many thanks to Dr. Roman Jaksik, who allowed me to discover our potential and helped us write the article.

This work was supported by the National Science Centre, Poland, under research grant project No. 2017/25/B/NZ4/01550 to Dorota Ścieglińska.

Students' Scientific Society of Biotechnology was supported in the 8th edition of funding for student research in Project Based Learning under the Excellence Initiative – Research University program, in accordance with Rector's of the Silesian University of Technology Regulation No. 54/2020 and 55/2020 issued March 13, 2020.

Bibliography

1. T. Wolski, B. Kędzia: Farmakoterapia skóry. Cz. 1. Budowa i fizjologia skóry. Postępy Fitoterapii 1/2019, 2019, s. 61–67.
2. I. Colombo, E. Sangiovanni, R. Maggio, C. Mattozzi, S. Zava, Y. Corbett, M. Fumagalli, C. Carlino, P.A. Corsetto, D. Scaccabarozzi, S. Calvieri, A. Gismondi, D. Taramelli, M. Dell'Agli: HaCaT Cells as a Reliable In Vitro Differentiation Model to Dissect the Inflammatory/Repair Response of Human Keratinocytes. Mediators of inflammation, 2017 (2017): 7435621, 2017.
3. R.B. Presland, B.A. Dale: Epithelial structural proteins of the skin and oral cavity: function in health and disease. Critical reviews in oral biology and medicine : an official publication of the American Association of Oral Biologists, 11(4), 383–408, 2000.
4. W. Sawicki: Histologia dla studentów medycyny. Podręcznik. Wydawnictwo Lekarskie PZWL, Warszawa 1997.
5. M.B. Murphrey, J.H. Miao, P.M. Zito: Histology, Stratum Corneum. StatPearls. 2021.

6. V. Kapuśniak: Proces rogowacenia komórek naskórka, Medycyna Weterynaryjna, 62 (1), 2006.
7. S. Eun Lee, S. Hun Lee: Skin Barrier and Calcium. Ann Dermatol, 30(3):265–275, 2018.
8. J. Reichelt, C. Bauer, R.M. Porter, E.B. Lane, V. Herzog: Out of balance: consequences of a partial keratin 10 knockout. Journal of Cell Science 110, 2175–2186, 1997.
9. S.H. Yuspa, A.E. Kilkenny, P.M. Steinert, D.R. Roop: Expression of Murine Epidermal Differentiation Markers Is Tightly Regulated by Restricted Extracellular Calcium Concentrations In Vitro. The Journal of Cell Biology(3):1207–17, 1989.
10. D.D Bikle, Z. Xie, C. Tu: Calcium regulation of keratinocyte differentiation. Expert Rev Endocrinol Metab, 7(4):461–472, 2012.
11. A.S. Évora, M.J. Adams, S.A. Johnson, Z. Zhang: Corneocytes: Relationship between Structural and Biomechanical Properties. Skin Pharmacol Physiol, 34:146–161, 2021.
12. https://wylecz.to/badania-laboratoryjne/badanie-immunohistochemiczne/, [dostęp: 17.04.2022].
13. V. Dubowitz, C.A. Sewry, A. Oldfors, R. Lane, MUSCLE BIOPSY, S. Elsevier, U. Kämmerer, M. Kapp, A.M. Gassel, T. Richter, C. Tank, J. Dietl, P. Ruck: A new rapid immunohistochemical staining technique using the EnVision antibody complex. The journal of histochemistry and cytochemistry: official journal of the Histochemistry Society, 49(5), 623–630, 2001.
14. S. Magaki, S. A Hojat, B. Wei, A. So, W. H Yong: An Introduction to the Performance of Immunohistochemistry. Biobanking. Methods in Molecular Biology, vol 1897. Humana Press, New York 2019.
15. D. Scieglinska, Z. Krawczyk: Expression, function, and regulation of the testis-enriched heat shock HSPA2 gene in rodents and humans. Cell Stress Chaperones. 2015;20(2):221–35, doi: 10.1007/s12192-014-0548-X.
16. D. Scieglinska, W. Piglowski, M. Chekan, A. Mazurek, Z. Krawczyk: Differential expression of HSPA1 and HSPA2 proteins in human tissues; tissue microarray--based immunohistochemical study. Histochem Cell Biol. 2011;135(4):337–50, doi: 10.1007/s00418-011-0791-5.

17. A. Gogler-Pigłowska, K. Klarzyńska, D.R. Sojka, A. Habryka, M. Głowala--Kosińska, M. Herok, M. Kryj, M. Halczok, Z. Krawczyk, D. Ścieglińska: Novel role for the testis-enriched HSPA2 protein in regulating epidermal keratinocyte differentiation. Journal of Cellular Plysiology, 2017.

IMMUNOHISTOCHEMICAL ASSESSMENT OF THE EXPRESSION LEVEL OF KERATIN 10 IN ORGANOTYPIC CULTURES OBTAINED FROM HACAT CELLS DEFICIENT IN HSPA2 PROTEIN

Abstract

Semi-quantitative IHC is a powerful method that allows to assess protein levels and localization in preserved tissues. This is possible through sophisticated image processing algorithms, which provide the means to quantify the differences in the level of stained protein using deconvolution. Fiji software is an example of such system, however its applications are significantly limited by the requirement to process only those images which are fully filled with the analyzed tissue. When the stained tissue is only part of the image, which is typical in studies of human epidermis, the results are distorted.

The aim of this study was to develop a novel image processing methodology which allows to conduct semi-quantitative analysis of protein levels, based on images which are not completely filled with the stained tissue, and use it to investigate how HSPA2 protein deficiency affects the localization of keratin 10 in human epidermal cells.

Using our automated image processing algorithm we were able to determine, the localization of keratin 10 in organotypic cell cultures. We were also able to assess, the differences in the intensity of keratin 10 staining in the in vitro grown epidermal sections, which allowed us to assess its association with the presence of the HSPA2 protein. The semi-quantitative protein analysis which we propose may contribute to a more detailed understanding of the role of HSPA2 protein in the process of keranocyte differentiation in human epidermis.

Keywords: Keratin 10, immunohistochemistry, DAB, keratinocytes organotypic culture, HSPA2

DOI:10.34918/85104

Wiktoria ŚLIWIŃSKA[1,*], Aleksandra SUWALSKA[1], Michał MARCZYK[1,2,*]

Chapter 4. INFLUENCE OF VARIOUS PREPROCESSING TECHNIQUES AND MODEL PARAMETERS ON NETWORK PERFORMANCE IN COVID-19 DETECTION

4.1. Introduction

SARS-CoV-2 virus, which belongs to the *coronaviridae* family, can attack human organisms [1]. The result of potential infection is COVID-19 disease. Despite the fact, that illnesses due to coronaviruses had occurred for at least 50 years [2], SARS-CoV-2 was the first one that caused a pandemic. In the majority of infected people, COVID-19 causes symptoms similar to a cold, however in some cases it can lead to much more dangerous health changes in the lungs, like consolidations or ground-glass opacities (Fig. 1). The SARS-CoV-2 infection could be detected by Nucleic Acid Amplification Tests (NAAT) or antigen tests. An alternative to confirm the disease is by recognizing the specific changes in the lungs on the chest X-Ray (CXR) or Computed Tomography (CT) images. Since CXR are much cheaper and easier to obtain, this imaging procedure is performed more often [3]. Too intense increase in infections resulting in COVID-19 disease led to the need of studying CXR images on a large scale, so scientists started to find a way of automatization the process, mostly using deep learning-based algorithms.

---

[1] Department of Data Science and Engineering, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland.

[2] Yale Cancer Center, Yale School of Medicine, 06511 New Haven, CT, USA.

[*] Corresponding authors: michal.marczyk@polsl.pl, wiktsli527@student.polsl.pl, ul. Akademicka 16, 44-100 Gliwice, PL.

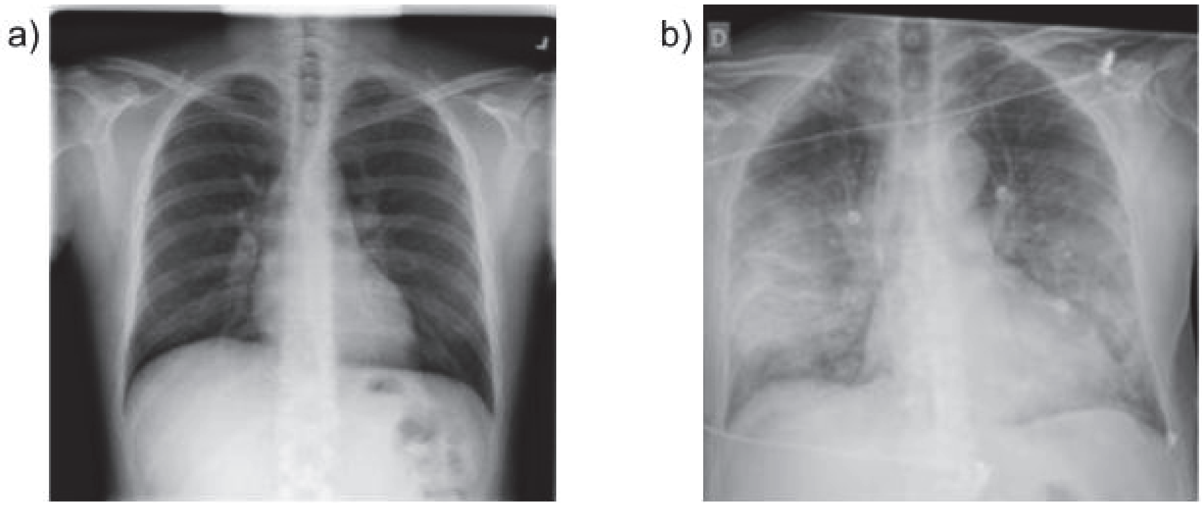


Fig. 1. Comparison of healthy patient lungs (a) and lungs of a patient with COVID-19 (b)
Rys. 1. Porównanie płuc zdrowego pacjenta (a) oraz płuc osoby chorującej na COVID-19 (b)

Deep learning methods are becoming more and more popular, due to their high efficiency and growing collections of well-annotated data. These methods are using artificial neural networks, which are comprised of artificial neurons, that learn from a huge number of data examples. The learning process is choosing the optimal parameters to model the formula, so the cost function would obtain minimal value. Images are mostly analyzed using convolutional neural networks, which can recognize patterns. The performance of the network is primarily evaluated by accuracy, but there are more measures used, like recall and sensitivity.

Deep learning methods were applied in many works concerning patients with pulmonary diseases. Some of them extended the approach to non-viral pneumonia [4], and other solves both, binary and multiclass problems [5, 6]. Researchers in [6] compared different image preprocessing methods, however, the best results were obtained for the network learned with original images. In [7-9] the combination of deep and machine learning methods was checked. The model described in [8] uses ResNet-50. The features were collected from the first fully connected layer and were fed as input to four machine learning algorithms, which made decisions by voting.

The goal of this study is to find the best technique and model parameters to improve the performance of the deep networks for COVID-19 detection. We analyzed the effect of changing the number of convolution layers, the influence of applying the Contrast Limited Adaptive Histogram Equalization algorithm, and the segmentation of the lung region before modeling. Many experiments were performed, from which we selected two models with the best results: one with and one without the application of lung segmentation.

4.2. Materials and Methods

4.2.1. Data

Models were trained with the use of the COVIDx online dataset [10] which contains other publicly available sets with CXR images. The number of images was equal to 10 316. 8 751 of them were images taken on healthy people (no COVID-19 symptoms) and 1 565 were taken on COVID-19 patients, so there was a disproportion between classes. Some of the images had artifacts induced by medical equipment or preliminary assessment, associated with adding arrows indicating changes in lungs.

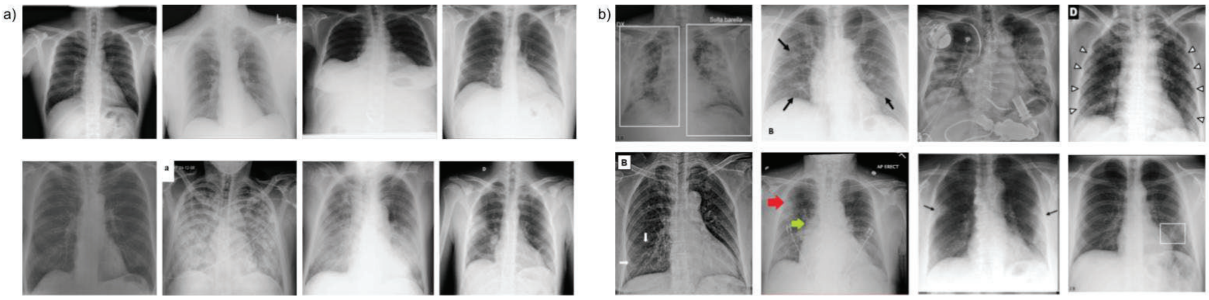


Fig. 2. Example images from COVIDx database without (a) and with artifacts (b)
Rys. 2. Przykładowe obrazy z bazy COVIDx bez (a) oraz z artefaktami (b)

Some networks were trained with input data after lung segmentation. In most cases, it allowed the removal of artifacts from images, because a lot of them were located outside the lung region. Binary image masks were generated with the use of the U-NET network, which is a part of the CIRCA system [11]. In the case of the positive class, where lungs are affected by changes, the difference between them and soft tissues is not sharp, resulting in troubles with segmentation. Therefore, masks with low quality appeared mostly in the COVID-19 class.

To check model generalization, apart from the COVIDx test set, networks were also evaluated using the independent, Spanish clinical dataset (SIIM-Covid19) [12]. We have used 6 333 CXR images. All of them belonged to the COVID-19 class, however, 4 subclasses were indicating the level of lung occupation: 'Negative for Pneumonia', 'Typical Appearance', 'Indeterminate Appearance', and 'Atypical Appearance'.

4.2.2. Preprocessing

All data was loaded as grayscale images and resized to 300 x 300 pixels. The COVIDx dataset was split into training, validation, and test set. In the beginning, 100 images of each class were randomly chosen for the test set. The rest was divided into training (70%) and validation (30%) sets.

In some experiments, the Contrast Limited Adaptive Histogram Equalization (CLAHE) algorithm was applied. This technique is used to enhance the contrast of images, so the changes in the lungs could be more visible. CLAHE is an improvement of standard Histogram Equalization (HE), which only stretches the histogram to obtain more varying pixel intensities. Nevertheless, it works globally, so HE may result in too bright or too dark areas. In the case of CLAHE, equalization is adaptive – it works locally, so the problem mentioned above does not occur. Additionally, the limitation of contrast enables control of the height of the histogram, which helps with noise reduction [13].

To enhance the mask quality, the Convex Hull algorithm was used. It finds the smallest convex set, that contains all points so that the lungs have softer edges and do not have any gaps. It allows for preserving important information from the inside of the lungs. The use of this method was necessary because a lot of patients with COVID-19 disease had advanced changes in their lungs that made segmentation difficult and resulted in low-quality masks.

The data were standardized with the mean and standard deviation of the training set. Spanish set was previously scaled from 32-bit to 8-bit images because networks were trained on 8-bit images from the COVIDx dataset.

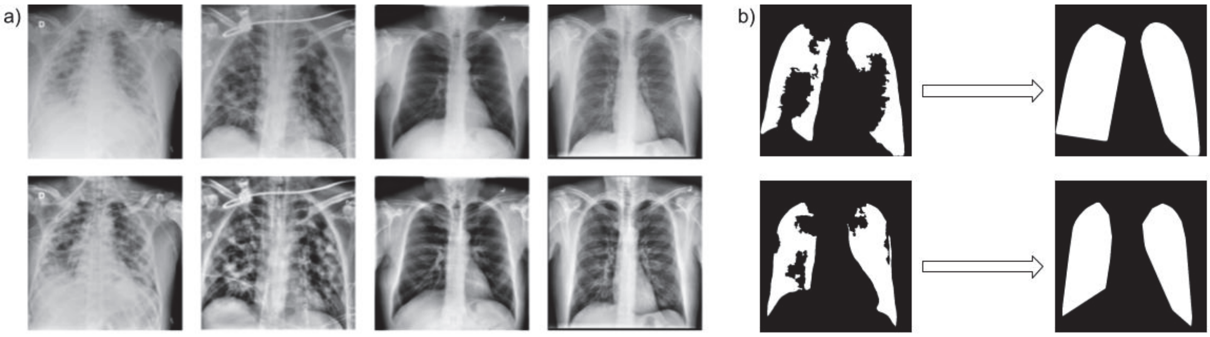


Fig. 3. (a) The effect of using the CLAHE algorithm. Original images are presented in the first row, the second row shows images after CLAHE. (b) The effect of using the convex hull algorithm on damaged masks

Rys. 3 (a) Efekt zastosowania algorytmu CLAHE. Obrazy oryginalne przedstawiono w pierwszym wierszu, wiersz drugi przedstawia obrazy po zastosowaniu CLAHE. (b) Wpływ zastosowania algorytmu convex hull na uszkodzone maski

4.2.3. Experiments

The conducted experiments were divided into two categories: those which were done using input images without segmentation and with it. In total, 96 models were created, 48 for both categories. Their architecture differs in the number of layers, activation function, learning rate, dropout coefficient, and the use of the CLAHE algorithm. Adam Optimizer was used in all cases. In each experiment, the augmentation technique was applied with horizontal flip and rotation up to 15°. The output of every convolutional layer was batch normalized and then passed to the pooling layer that calculates the maximum (max pooling). In the convolutional and pooling layers, padding was set to ‘same’, the step in the convolutional layers was equal to one, and in the pooling layers to two. As final models, two networks with the best performance were selected, one with and one without segmentation.

4.2.4. Evaluation

For network comparison, two metrics were used: accuracy and recall. The first one is the most used and gives us information about the percentage of cases correctly classified. A recall is a metric that informs how good the model is in classifying positive cases. It is especially important in medical image classification because it is better to double-check a patient that we are not sure about than leave him with the disease.

For selected models, Grad-CAM (Gradient-weighted Class Activation Mapping) maps were prepared. Grad-CAMs allow for the visualization of areas that are important for the network during the prediction [14]. For a time, Neural Networks were considered black boxes, since the process of learning was not clear to researchers. The input data were just entered, and the result was recorded, however, we did not know why the network classified this entry into a specific class. With methods like Grad-CAM, it is possible to understand better how the model works, as they enable the verification of network decisions. The influence of changing the number of convolutional layers and applying CLAHE was checked with the use of statistical tests. Additionally, the effect size was checked. In the same way, the dependency between mask quality and the model prediction was verified.

4.3. Results

The architectures of selected networks are presented in Fig. 4. In panel (a) architecture of the network without lung segmentation is presented. The CLAHE algorithm with a *clipLimit* parameter equal to 2 was used to transform the input data. The network uses the Adam optimizer with a learning rate equal to 0.001. It consists of 3 convolutional layers with a sigmoidal activation function. After each convolutional layer, a max-pooling layer was applied. There are 3 fully connected layers, two of them use a sigmoidal function, and the last one uses a softmax activation function. This network obtained an accuracy and recall of 95%. The second network shown in panel (b), which works on images after lung segmentation, also uses the CLAHE with the *clipLimit* equal to 2. This network contains 4 convolutional, 4 max-pooling, and 4 fully connected layers. The activation function in all layers, except the last one with the softmax function, is ReLU. The optimizer used in this network is Adam with a 0.0001 learning rate. The accuracy obtained by the second network was equal to 93% whereas recall was equal to 87%.

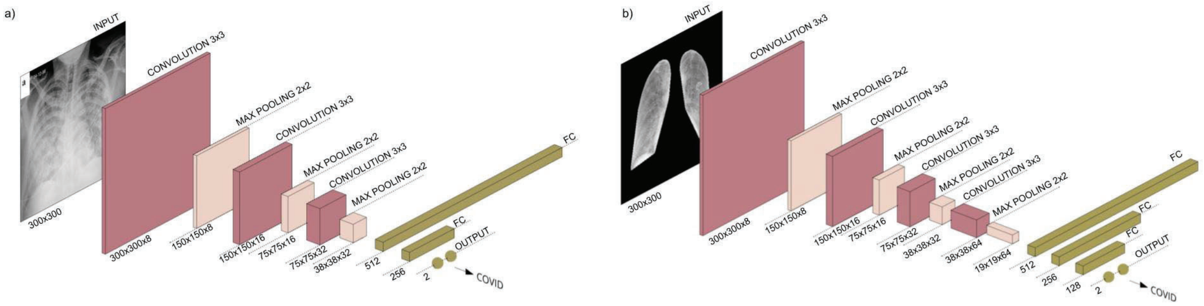


Fig. 4. (a) Architecture of network trained without the use of segmentation. (b) Architecture of network trained with the use of lung segmentation

Rys. 4. (a) Architektura sieci uczonej bez segmentacji płuc. (b) Architektura sieci uczonej na obrazach po segmentacji płuc

Grad-CAM activation maps are shown in Fig. 5. The first column presents original images and the second one Grad-CAM maps for this image created by the first model. In the third column are images after lung segmentation, and in the last column, are Grad-CAM maps for the second model.

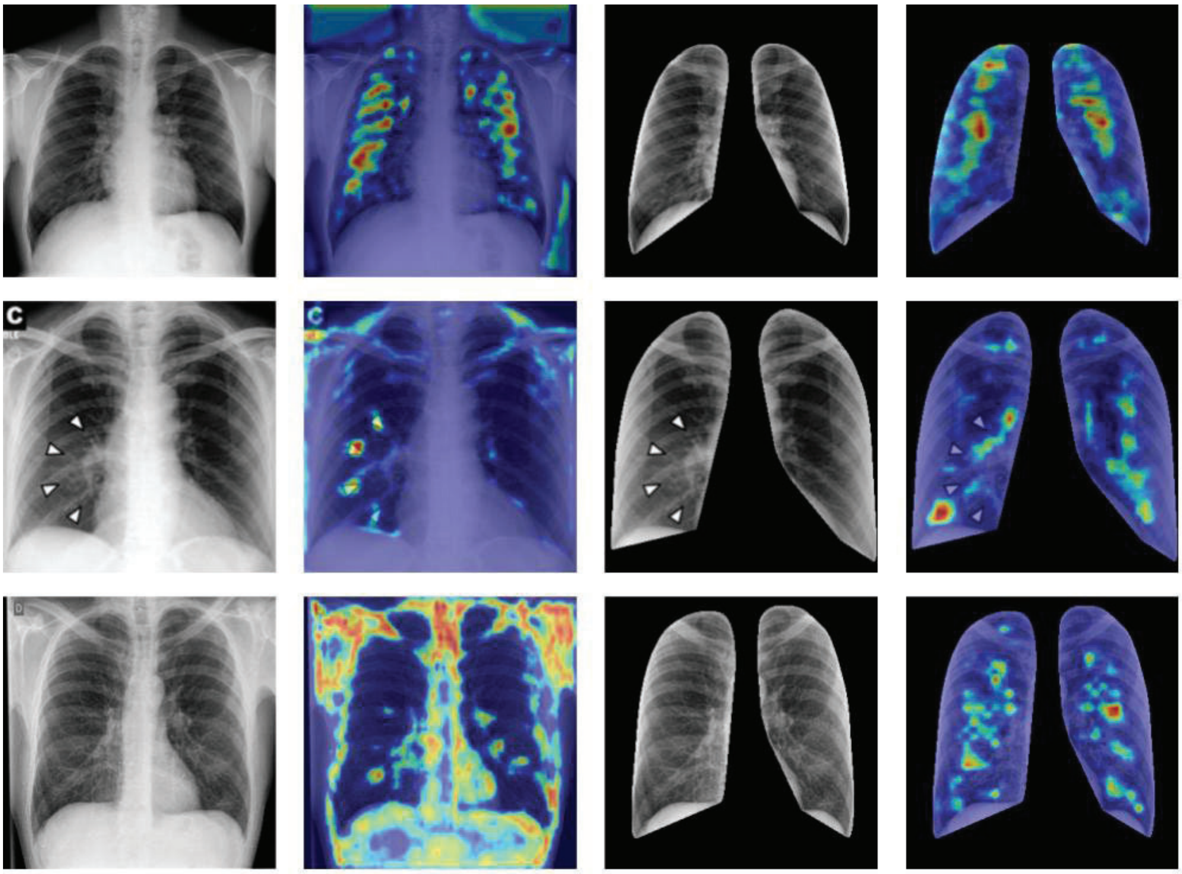


Fig. 5. Grad-CAM activation maps. The first column contains original images, the second one Grad-CAM maps for the first model, the third column images after lung segmentation and the fourth column – Grad-CAM maps for the second model

Rys. 5. Mapy aktywacji Grad-CAM. Pierwsza kolumna zawiera obrazy oryginalne, druga mapy aktywacji dla pierwszego modelu, kolumna trzecia przedstawia obrazy po segmentacji płuc, a czwarta mapy aktywacji dla drugiego modelu

The difference in results for 3 and 4 convolutional layers and results for networks trained with and without the use of CLAHE algorithm was tested by U-Mann-Whitney statistical tests with a 5% significance level and the effect size of those relationships was checked with Glass rank biserial correlation. The same tests were used to verify if the model prediction depends on the mask quality (Table 1). There is no statistically significant difference in medians of the results for 3 and 4 convolutional layers for COVIDx and SIIM-Covid19 databases. Nevertheless, the effect size for SIIM-Covid19 showed a small, negative correlation between the number of convolutional layers and the accuracy of the model. There is a statistically significant difference in medians of the results for networks with the use of CLAHE and without for SIIM-Covid19. The effect size showed a large, positive correlation. However, applying the CLAHE algorithm did not change the accuracy of the model in the COVIDx dataset. The relationship between mask quality and model prediction (Table 1), showed that in both datasets there is a difference between mask quality median for healthy and COVID-19 predictions. The effect size for the COVIDx dataset is medium and for SIIM-Covid19 small.

Table 1

Results of conducted statistical tests

Tested parameter	Dataset	P-value	Effect size
No. of convolutional layers	COVIDx	0.595	-0.0634
	SIIM-Covid19	0.05966	-0.224
The use of CLAHE	COVIDx	0.6494	-0.0543
	SIIM-Covid19	0.00001176	0.52
The mask quality	COVIDx	0.0000172	0.355
	SIIM-Covid19	0.01863	0.112

4.4. Discussion

Artifacts are more likely to occur in the positive group and as a result, the network can learn that this additional information suggests a positive class, which is not desirable. We used the activation maps to enable the visual verification of selected models. The network trained on the original images considered background, body, and artifacts during model prediction, which may lead to erroneous findings (Fig. 5). After segmentation, only the lungs remained on the image, so the network could not learn anything from the background or other body parts. Since a lot of artifacts occurred outside the lungs, the segmentation also removed them. After that, there were too few artifacts for the network to learn them as a class-specific pattern. This situation can be observed in the second row of Fig. 5.

CLAHE algorithm improved results for the SIIM-Covid19 dataset, but there was no statistical dependency for the COVIDx dataset. The effect size of 0.52 for SIIM-Covid19 shows, that there was a large difference – after using the CLAHE, network performance was better. Applying the CLAHE algorithm in most cases improves results [15-17] but as presented in [18], where the accuracy dropped from 93.87% to 92.04% and recall from 92.13% to 89.56% after using the CLAHE as a pre-processing step, the improvement is not the rule. Results for networks with 3 and 4 convolutional layers did not show a statistically significant difference. It may be that the difference of a single layer is not enough to show the dependency and more experiments should be performed. However, there is a small negative correlation for the SIIM-Covid19 dataset, which indicates better results for networks with 3 convolutional layers. In [19] comparison of some ready architectures for the problem of COVID-19 detection was made. VGG16 model was the one with the best performance of 95.88% accuracy. This network contains 13 convolutional layers. Architectures with more convolutional layers obtained the worst accuracy, like ResNet152V2 which consists of 152 layers and obtained an accuracy of 61.76%.

Masks with low quality appeared mostly in the COVID-19 class, so there was a possibility that the network will make decisions based on this information. U-Mann-Whitney statistical tests were conducted in order to verify this hypothesis. They showed that at the 5% significance level in both datasets there is a difference in quality mask medians for classes. However, Glass rank biserial correlation showed that for SIIM-Covid19 this dependency is small, and for COVIDx is medium. The poorly visible dependence on the Spanish set could be caused by a small number of masks of poor quality in relation to the entire collection, the number of which was over 6 000. In the COVIDx set, the relationship was checked on a test set of 200 images.

4.5. Conclusion

Lung segmentation is a very important preprocessing step as it can remove biologically unimportant information from the image. As a result, the network focuses on essential areas of the image. The CLAHE improved the generalization of our models, but there was no dependency on the dataset on which the network was trained. There was no relationship between the number of convolutional layers and network results. Therefore, the use of a network with 3 layers appears to be a better option as the network is less susceptible to overfitting and at the same time faster due to lower computational complexity. Additionally, it is important to check if there are no other dependencies, like in our case, the relationship between mask quality and model prediction.

Bibliography

1. Y.A. Malik: Properties of coronavirus and sars-cov-2, *The Malaysian journal of pathology* (2020) **42(1)**:3–11.
2. Z. Pejsak, K. Tarasiuk, B. Tokarz-Deptuła: Wybrane dane na temat zakażeń koronawirusami, ze szczególnym uwzględnieniem sars-cov-2, *Medycyna Weterynaryjna* (2020) **76(5)**:258–262.
3. A. Jacobi, M. Chung, A. Bernheim, C. Eber: Portable chest x-ray in coronavirus disease-19 (covid-19): A pictorial review, *Clinical imaging* (2020) **64(8)**:35–42.
4. R. Jain, M. Gupta, S. Taneja, D.J. Hemanth, Deep learning based detection and analysis of covid-19 on chest x-ray images, *Applied Intelligence* (2021) **51(3)**:1690–1700.
5. T. Ozturk, M. Talo, E.A. Yildirim, U.B. Baloglu, O. Yildirim, U.R. Acharya: Automated detection of covid-19 cases using deep neural networks with x-ray images, *Computers in biology and medicine* (2020) **121(8)**:103792.
6. C. Doglioni, C. Ravaglia, M. Chilosi, G. Rossi, A. Dubini, F. Pedica, S. Piciucchi, A. Vizzuso, F. Stella, S. Maitan, V. Agnoletti, S. Puglisi, G. Poletti, V. Sambri, G. Pizzolo, V. Bronte, A. Wells, V. Poletti: Covid-19 interstitial pneumonia: Histological and immunohistochemical features on cryobiopsies, *Respiration* (2021) **100(5)**:369–379.
7. L. Hussain, T. Nguyen, H. Li, A.A. Abbasi, K.J. Lone, Z. Zhao, M. Zaib, A. Chen, T.Q. Duong: Machine-learning classification of texture features of portable chest x-ray accurately classifies covid-19 lung infection, *BioMedical Engineering OnLine* (2020) **19(1)**:1–18.
8. P. Saha, M.S. Sadi, M.M. Islam: Emcnet: Automated covid-19 diagnosis from x-ray images using convolutional neural network and ensemble of machine learning classifiers, *Informatics in medicine unlocked* (2021) **22**:100505.
9. D. Wang, J. Mo, G. Zhou, L. Xu, Y. Liu: An efficient mixture of deep and machine learning models for covid-19 diagnosis in chest x-ray images, *PloS one* (2020) **15(11)**:e0242535.
10. L. Wang, Z.Q. Lin, A. Wong, https://github.com/lindawangg/covidnet/blob/master/docs/covidx.md [accessed: 22.05.2020 r.].
11. https://circa.aei.polsl.pl/ [accessed: 15.10.2021 r.].
12. https://www.kaggle.com/c/siim-covid19-detection/data [accessed: 15.10.2021].

13. S. Sahu, A.K. Singh, S.P. Ghrera, M. Elhoseny: An approach for de-noising and contrast enhancement of retinal fundus image using CLAHE, *Optics & Laser Technology* (2019) **110**:87–98.
14. R.R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, D. Batra: Grad-CAM: Why did you say that?, *arXiv preprint arXiv:1611.07450* (2016).
15. S.S. Narli, G. Altan, CLAHE-based Enhancement to Transfer Learning in COVID-19 Detection, *Gazi Mühendislik Bilimleri Dergisi* (2022) 1–11.
16. D. Reynaldi, B.S. Negara, S. Sanjaya, E. Satria: COVID-19 Classification for Chest X-Ray Images using Deep Learning and Resnet-101, *2021 International Congress of Advanced Technology and Engineering (ICOTEN)* IEEE (2021) 1–4.
17. F. Saiz, I. Barandiaran: COVID-19 detection in chest X-ray images using a deep learning approach, *Journal of Interactive Multimedia and Artificial Intelligence*. (2020) 1.
18. A. Tahir, Y. Qiblawey, A. Khandakar, T. Rahman, U. Khurshid, F. Musharavati, M.T. Islamd, S. Kiranyaza, S. Al-Maadeed, M.E.H. Chowdhury: Deep Learning for Reliable Classification of COVID-19, MERS, and SARS from Chest X-ray Images, *Cognitive Computation* (2022) **14(5)**:1752–1772.
19. A. Makris, I. Kontopoulos, K. Tserpes: COVID-19 detection from chest X-Ray images using Deep Learning and Convolutional Neural Networks, *11th Hellenic conference on artificial intelligence* (2020) 60–66.

INFLUENCE OF VARIOUS PREPROCESSING TECHNIQUES AND MODEL PARAMETERS ON NETWORK PERFORMANCE IN COVID-19 DETECTION

Abstract

The COVID-19 pandemic caused a need for an efficient tool to predict patients infected with SARS-CoV-2. One of the possibilities for recognizing the disease is by analyzing chest X-ray (CXR) images, as COVID-19 can cause lung consolidations or ground-glass opacities. However, due to the huge number of infected people, radiologists could not perform real-time analyses, so computer-assisted methods are necessary. This work aims to compare the impact of several techniques on the performance of a deep learning-based approach for COVID-19 prediction. Networks were trained using 10 316 CXRs images. Different preprocessing methods and model parameters were tested, such as the number of convolutional layers, the influence of applying the Contrast Limited Adaptive Histogram Equalization algorithm, and the usage of lung segmentation before model building. Finally, two best-performing networks were selected (one with and one without lung segmentation), which achieved an accuracy of 95% and 93%, respectively. Additionally, networks were evaluated on the independent set of 6 333 images reaching an accuracy of 92.9% and 97.6%, respectively. Grad-CAM and statistical tests were used to analyze the model performance scores. In conclusion, CLAHE mostly improves model generalization. According to the number of convolutional layers, there were no differences in results. During lung segmentation, unnecessary information is removed, which increases the probability of learning only truly important patterns.

Keywords: deep learning, classification, COVID-19, X-ray

DOI:10.34918/85105

Seweryn KALISZ[1,*], Michał MARCZYK[1,2]

Chapter 5. METHODS FOR REDUCTION OF BLUR EFFECT IN THE DENOISING AUTOENCODER MODEL FOR RIB SUPPRESSION IN CXR IMAGES

5.1. Introduction

The development of bioinformatics tools related to image analysis is of high interest to researchers, medical doctors, and engineers since selected technologies could improve clinical performance and efficiency and influence the development of personalized medicine [1]. One of the most widely used medical imaging methods is X-ray imaging, which is characterized by low cost, short processing time, low radiation dose, and high availability [1, 2]. The chest X-ray (CXR) is frequently the first imaging study obtained during a broad range of conditions and remains central to the screening, diagnosis, and management of the disease [2, 3]. Chest radiography provides images of soft tissues such as the heart, airways, blood vessels, and lungs, and additionally the bones of the spine and chest. Overlapping of structures present in the image can complicate the visual interpretation of the detection of abnormalities in radiographs by medical doctors leading to false negative results [4]. A major reason is the presence of bone structures on the radiograph during soft-tissue diagnosis. There exists a technique that enables the separation of soft tissue from bone called dual-energy subtraction (DES) radiography. Some studies showed that suppressing rib shadows on soft-tissue images by using the DES technique has improved the speed and accuracy of radiologists in diagnosing pulmonary nodules [5]. However, compared to traditional imaging, DES requires a higher radiation dose and could result in noisy images with visual abnormalities due to the patient moving [6]. These

---

[1] Department of Data Science and Engineering, Faculty of Automatic Control, Silesian University of Technology, Gliwice, Poland.
[2] Yale Cancer Center, Yale School of Medicine, 06511 New Haven, CT, USA.
* Corresponding author: Seweryn.Kalisz@polsl.pl.

problems, together with lower availability compared to traditional CXR, have interested researchers in the field of image processing to search for solutions allowing the reduction of the presence of bone structures on X-ray images resulting in soft-tissue reconstruction.

Deep learning has become the technique of choice for image analysis tasks in recent years and has had a huge impact on the field of medical imaging [3, 7]. A common method for suppressing ribs from CXR images is based on autoencoder architecture. In literature, three [8] or four [9] layer convolutional denoising autoencoder architectures can be found. Models based on autoencoder architecture are often characterized by a blurring effect on decoded images, which is mostly caused by the occurrence of distance metrics in the loss function [25]. Additionally, several modifications of convolutional neural networks were tested [10, 11], also in the wavelet domain [12]. There are also approaches based on conditional generative adversarial networks minimizing the pairwise image difference and adding Haar 2D wavelet decomposition to improve model convergence [13].

This work aims to review and evaluate methods to improve image quality after the suppression of bone structures by reducing the blurring effect on the example of convolutional denoising autoencoder, composed of 4 layers, and introduced earlier [15]. First, we checked the modification of the image resizing scheme through layers of the autoencoder. Next, we introduce bridge connections between the coder and encoder layers. Finally, we modified the α parameter of the loss function, which was responsible for establishing the weights between the mean square error index and the multi-scale structural similarity index. Also, we have tested the combination of these methods.

5.2. Materials and methods

5.2.1. Data

The Bone Suppression dataset contains 35 pairs of standard CXR images and corresponding soft-tissue images created using DES technology. These data were gathered from a variety of online sources and made available by Innopolis University researchers [8]. There are 11 female pairs and 24 male pairs, each with a different resolution (the largest being 660 x 775 pixels and the smallest being 424 x 465 pixels). JPEG images had a quality compression ratio ranging from 85 to 100 (high-quality images). Several image pairs have arrows or captions superimposed on the CXR image as artifacts. It can be noted that the images are characterized by varying brightness.

5.2.2. Image preprocessing

Considering the different sizes of the images in the dataset all images were cropped to a square to maintain the correct proportion of the structures. The images were then resized to 512 x 512 pixels resolution. Next, the contrast of each image has been enhanced by using Contrast Limited Adaptive Histogram Equalization with a clip limit equal to 2 and grid size equal to 8 x 8 pixels, which specifies the area within which the contrast is increased so that the input and output histograms for the region are almost identical. This operation allows an increase of details in the biological structures shown in medical images [15].

Data augmentation operations were applied to increase the robustness of the model and to increase the size of training data. The resulting images allow reflection of patients' behavior (affine transformations) and image property differences between X-ray scanners (color transformations). In this work, the following operations were used [16]: (i) translation (± 10 pixels); (ii) rotation (± 5 degrees); (iii) scaling (± 10%); brightness change (± 20%); contrast change (± 20%). The parameters of each operation were chosen based on the recent literature [17, 18].

5.2.3. Model architecture

Convolutional denoising autoencoder is the unsupervised learning algorithm that learns to map corrupted data to uncorrupted data by minimizing the loss function between pairs of images [19]. The ribs in the images are treated as noise (corrupted data) which we want to obtain for the soft-tissue image (uncorrupted data). The architecture used is symmetric and consists of 4 layers with several filters (32, 32, 48, 48) and a window size equal to 4 x 4 pixels on the encoder and decoder part. The input image should be grayscale, characterized by dimensions of 512 x 512 pixels, and be normalized to a floating point value in the range <0;1>. Rectified Linear Unit

(ReLU) in each layer was used as an activation function. The padding property of the convolutional block has been set to 'same' for all layers.

The loss function minimizes Mean Square Error (MSE) and maximizes Multi-Scale Structural Similarity Index Measure (MS-SSIM) [20]. MS-SSIM can be defined by the equation:

$$MS - SSIM(I, G) = [l_M(I, G)]^{\alpha_M} \cdot \prod_{j=1}^{M} \left[c_j(I, G)\right]^{\beta_j} \left[s_j(I, G)\right]^{\gamma_j} \tag{1}$$

where *l, c,* and *s* are the luminance, contrast and structure terms at scale *M* and *j*. The exponents *α,* β, and γ are used to adjust the relative importance of different components. Two discrete non-negative signals are represented by *I* and *G*. *I* is a model output and *G* is a ground truth image (soft-tissue image).

The mathematical description of the loss function is given by the following equations:

$$L = \alpha \cdot L^{MS-SSIM} + (1 - \alpha) \cdot L^{MSE} \tag{2}$$

$$L^{MS-SSIM}(I) = \frac{1}{N} \sum_{i \in I} 1 - MS - SSIM(I(i), G(i)) \tag{3}$$

$$L^{MSE}(I) = \frac{1}{N} \sum_{i \in I} (I(i) - G(i))^2 \tag{4}$$

where *N* is several pixels in *I*, *i* is the index of pixels in *I* and *α* is a parameter. The value of parameter *α* was set to 0.84 in the base model, which was empirically determined as the best value for image analysis in [21]. The training was run for 300 epochs in all cases. The initial learning rate was 0.001 for the first 100 epochs. Then it decreased twice every 50 epochs. The model was implemented in Python 3.9.12 using Tensorflow 2.5.0 with CUDA 11.2 and cuDNN 8.2.1 libraries. Figure 1 presents the architecture of the base model. Parameters ABCD in convolutional layers specifying the strides of the convolution along the height and width of the input image.

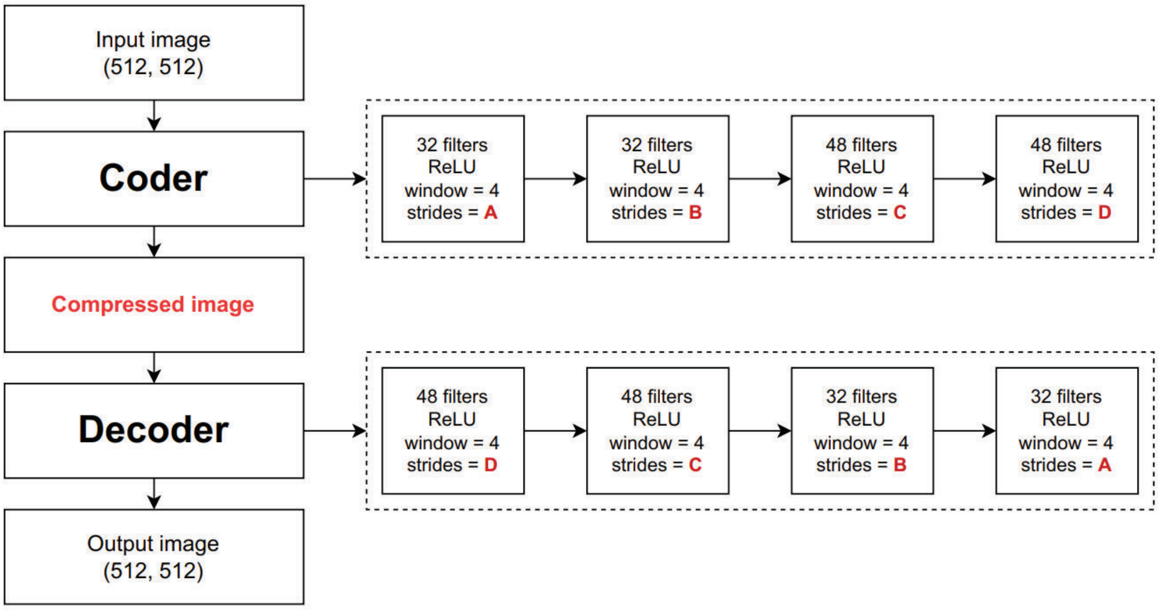


Fig. 1. The base architecture of convolutional denoising autoencoder for bone suppression in CXR images

Rys. 1. Podstawowa architektura konwolucyjnego autoenkodera odszumiającego do tłumienia kości w obrazach rentgenowskich klatki piersiowej

5.2.4. Image resizing

The stride parameter defines the step size for sampling when scanning the input layer to run convolutional operations [22]. In this study, we tested the effect of changing the strides parameter in individual layers, which also changed the compressed images. The compressed image at the input to the decoder is 64 x 64 pixels or 128 x 128 pixels, where it is then increased to its original size by increasing the size of the stride. Table 1 shows the verified configurations. The name Base refers to the configuration used in previously published work.

Table 1

Experimental stride parameters

Configuration Name	A	B	C	D	Compressed Image
Base	1	2	2	2	64 x 64
Strides 1122	1	1	2	2	128 x 128
Strides 1212	1	2	1	2	128 x 128
Strides 1221	1	2	2	1	128 x 128
Strides 2121	2	1	2	1	128 x 128
Strides 2112	2	1	1	2	128 x 128
Strides 2211	2	2	1	1	128 x 128

5.2.5. Bridge Connections

Each convolutional layer learns image features which are called feature maps, which are passed forward using bridge connections (shortcut connections, skip connections) to layers that are not directly adjacent. Compared to the ResNet architecture, the connections are symmetrical between the encoder and decoder. The element-wise addition of the feature maps from a shortcut connection and the connected deconvolutional layer follows [23]. The introduction of symmetric skip connections between the encoder and decoder exhibits two main advantages: (i) improving the results, in deep networks, composed of many layers, where many details are lost during image resizing, (ii) solving the optimization challenge posed by gradient vanishing, resulting in increased performance as the network grows deeper [24]. In this work, the effect of bridging connections between the extreme layers of the autoencoder was examined using different configurations shown in Fig. 2.

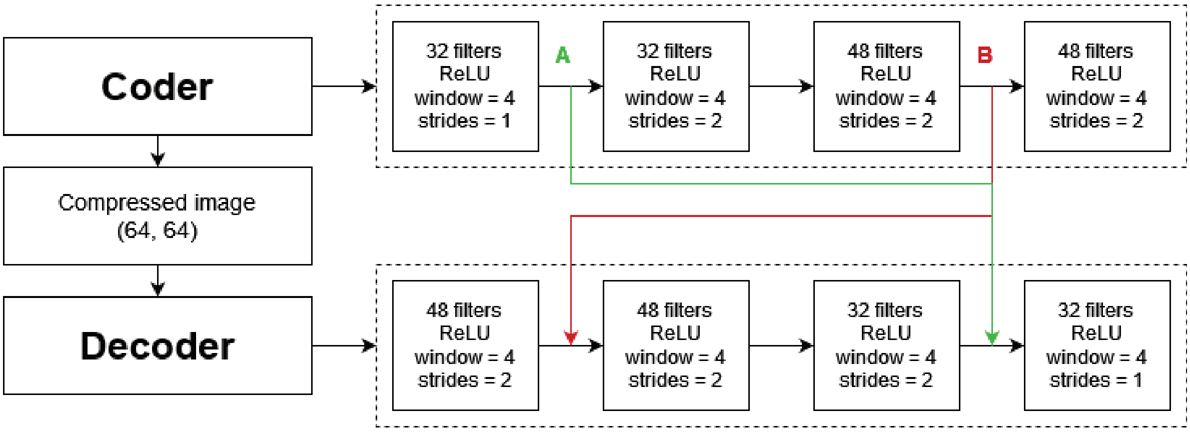


Fig. 2. The base architecture of convolutional denoising autonecoder extended with bridge connections
Rys. 2. Podstawowa architektura konwolucyjnego autonekodera denoisingowego z uwzględnieniem połączeń mostkowych

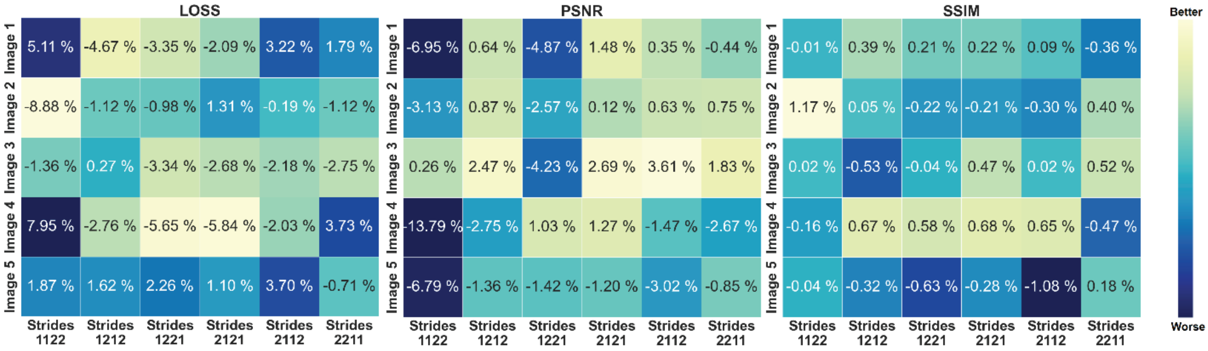
5.3. Experiments and results

The Bone Suppression dataset was divided into a training set containing 25 pairs of images, a validation set, and a test set containing 5 pairs each. New images were generated for the training and validation set by applying an augmentation technique with the same operations for each image in the pair. After this operation, the training set consists of 1025 images, while the validation set consists of 205 images which is 20% of the training set size. The batch size was set to 16. Adam was chosen as the optimizer in the model training. The models were trained for 300 epochs. Each model was trained using the same initial parameters previously generated at random. The model with the best validation loss was selected for each run. Calculations were performed on a desktop computer with an Intel Core I5-10500 CPU, 64 GB Ram, and RTX 3060 12GB graphics card. A single training took approximately 2 hours.

The value of the loss function, Peak signal-to-noise ratio (PSNR), and Structural Similarity Index Measure (SSIM) were used as quality indicators to evaluate the model outcome in comparison to the ground truth image (soft-tissue image measured using the DES method) on 5 pairs of an independent test set data. The values obtained are presented in % and refer to the base architecture introduced earlier [14].

5.3.1. Image resizing

The basic architecture resizes the input image three times by half of the current size, and the compressed image is 64 x 64 pixels. The effect of changing the compressed image size to 128 x 128 pixels in various configurations was tested. The greatest improvement for loss function and image quality metrics was obtained for the following configuration: Strides 2121. Detailed results for all configurations are shown in Fig. 3.



LOSS

	Strides 1122	Strides 1212	Strides 1221	Strides 2121	Strides 2112	Strides 2211
Image 1	5.11 %	-4.67 %	-3.35 %	-2.09 %	3.22 %	1.79 %
Image 2	-8.88 %	-1.12 %	-0.98 %	1.31 %	-0.19 %	-1.12 %
Image 3	-1.36 %	0.27 %	-3.34 %	-2.68 %	-2.18 %	-2.75 %
Image 4	7.95 %	-2.76 %	-5.65 %	-5.84 %	-2.03 %	3.73 %
Image 5	1.87 %	1.62 %	2.26 %	1.10 %	3.70 %	-0.71 %

PSNR

	Strides 1122	Strides 1212	Strides 1221	Strides 2121	Strides 2112	Strides 2211
Image 1	-6.95 %	0.64 %	-4.87 %	1.48 %	0.35 %	-0.44 %
Image 2	-3.13 %	0.87 %	-2.57 %	0.12 %	0.63 %	0.75 %
Image 3	0.26 %	2.47 %	-4.23 %	2.69 %	3.61 %	1.83 %
Image 4	-13.79 %	-2.75 %	1.03 %	1.27 %	-1.47 %	-2.67 %
Image 5	-6.79 %	-1.36 %	-1.42 %	-1.20 %	-3.02 %	-0.85 %

SSIM

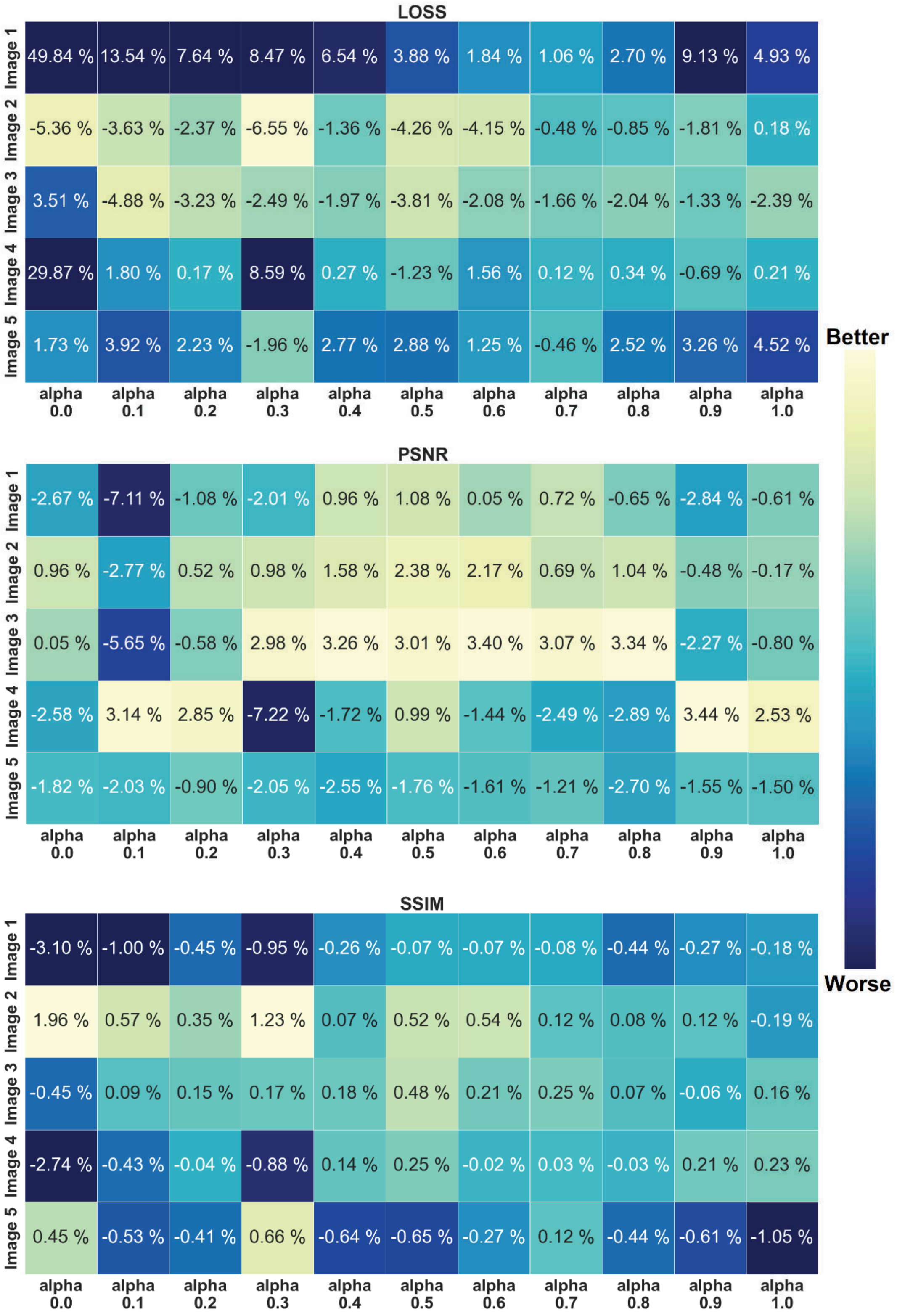
	Strides 1122	Strides 1212	Strides 1221	Strides 2121	Strides 2112	Strides 2211
Image 1	-0.01 %	0.39 %	0.21 %	0.22 %	0.09 %	-0.36 %
Image 2	1.17 %	0.05 %	-0.22 %	-0.21 %	-0.30 %	0.40 %
Image 3	0.02 %	-0.53 %	-0.04 %	0.47 %	0.02 %	0.52 %
Image 4	-0.16 %	0.67 %	0.58 %	0.68 %	0.65 %	-0.47 %
Image 5	-0.04 %	-0.32 %	-0.63 %	-0.28 %	-1.08 %	0.18 %

Fig. 3. Heatmaps with values of the loss function and two measures of image quality, calculated between model output and soft-tissue image: Peak signal-to-noise ratio (middle) and Structural Similarity Index Measure (right) for different strides configuration

Rys. 3. Mapy ciepła przedstawiające wartości funkcji straty oraz dwóch wskaźników jakości obrazu: PSNR oraz SSIM dla różnych konfiguracji parametru „strides”

5.3.2. Loss Function

Parameter α in the loss function is responsible for establishing the weights between the mean square error index and the multi-scale structural similarity index. The parameter value was changed from 0 to 1 with a step of 0.1 using the Base configuration shown in Table 1. The best results were obtained for a parameter α equal to 0.5. The value of the loss function for test images presented in the heatmap was calculated for a baseline parameter α equal to 0.84 for each of the trained models. Figure 4 shows the results obtained.



LOSS

	alpha 0.0	alpha 0.1	alpha 0.2	alpha 0.3	alpha 0.4	alpha 0.5	alpha 0.6	alpha 0.7	alpha 0.8	alpha 0.9	alpha 1.0
Image 1	49.84 %	13.54 %	7.64 %	8.47 %	6.54 %	3.88 %	1.84 %	1.06 %	2.70 %	9.13 %	4.93 %
Image 2	-5.36 %	-3.63 %	-2.37 %	-6.55 %	-1.36 %	-4.26 %	-4.15 %	-0.48 %	-0.85 %	-1.81 %	0.18 %
Image 3	3.51 %	-4.88 %	-3.23 %	-2.49 %	-1.97 %	-3.81 %	-2.08 %	-1.66 %	-2.04 %	-1.33 %	-2.39 %
Image 4	29.87 %	1.80 %	0.17 %	8.59 %	0.27 %	-1.23 %	1.56 %	0.12 %	0.34 %	-0.69 %	0.21 %
Image 5	1.73 %	3.92 %	2.23 %	-1.96 %	2.77 %	2.88 %	1.25 %	-0.46 %	2.52 %	3.26 %	4.52 %

PSNR

	alpha 0.0	alpha 0.1	alpha 0.2	alpha 0.3	alpha 0.4	alpha 0.5	alpha 0.6	alpha 0.7	alpha 0.8	alpha 0.9	alpha 1.0
Image 1	-2.67 %	-7.11 %	-1.08 %	-2.01 %	0.96 %	1.08 %	0.05 %	0.72 %	-0.65 %	-2.84 %	-0.61 %
Image 2	0.96 %	-2.77 %	0.52 %	0.98 %	1.58 %	2.38 %	2.17 %	0.69 %	1.04 %	-0.48 %	-0.17 %
Image 3	0.05 %	-5.65 %	-0.58 %	2.98 %	3.26 %	3.01 %	3.40 %	3.07 %	3.34 %	-2.27 %	-0.80 %
Image 4	-2.58 %	3.14 %	2.85 %	-7.22 %	-1.72 %	0.99 %	-1.44 %	-2.49 %	-2.89 %	3.44 %	2.53 %
Image 5	-1.82 %	-2.03 %	-0.90 %	-2.05 %	-2.55 %	-1.76 %	-1.61 %	-1.21 %	-2.70 %	-1.55 %	-1.50 %

SSIM

	alpha 0.0	alpha 0.1	alpha 0.2	alpha 0.3	alpha 0.4	alpha 0.5	alpha 0.6	alpha 0.7	alpha 0.8	alpha 0.9	alpha 1.0
Image 1	-3.10 %	-1.00 %	-0.45 %	-0.95 %	-0.26 %	-0.07 %	-0.07 %	-0.08 %	-0.44 %	-0.27 %	-0.18 %
Image 2	1.96 %	0.57 %	0.35 %	1.23 %	0.07 %	0.52 %	0.54 %	0.12 %	0.08 %	0.12 %	-0.19 %
Image 3	-0.45 %	0.09 %	0.15 %	0.17 %	0.18 %	0.48 %	0.21 %	0.25 %	0.07 %	-0.06 %	0.16 %
Image 4	-2.74 %	-0.43 %	-0.04 %	-0.88 %	0.14 %	0.25 %	-0.02 %	0.03 %	-0.03 %	0.21 %	0.23 %
Image 5	0.45 %	-0.53 %	-0.41 %	0.66 %	-0.64 %	-0.65 %	-0.27 %	0.12 %	-0.44 %	-0.61 %	-1.05 %

Fig. 4. Heatmaps with values of the loss function and two measures of image quality, calculated between model output and soft-tissue image: Peak signal-to-noise ratio (middle) and Structural Similarity Index Measure (bottom) for various α values in range <0;1>

Rys. 4. Mapy ciepła przedstawiające wartości funkcji straty oraz dwóch wskaźników jakości obrazu: PSNR oraz SSIM dla zmian parametru α w przedziale <0;1>

5.3.3. Bridge Connections

Bridge connections between the symmetrical encoder and decoder layers were introduced into the basic architecture. In this work three configurations were tested: (i) the connection between the outermost layers (connection A in Fig. 2), (ii) the connection between innermost layers (connection B in Fig. 2), (iii) the connection between the outermost layers and the innermost layers (A+B). The best image quality results were obtained for the variant characterized by 2 skipped connections. The detailed results are shown in Fig. 5.

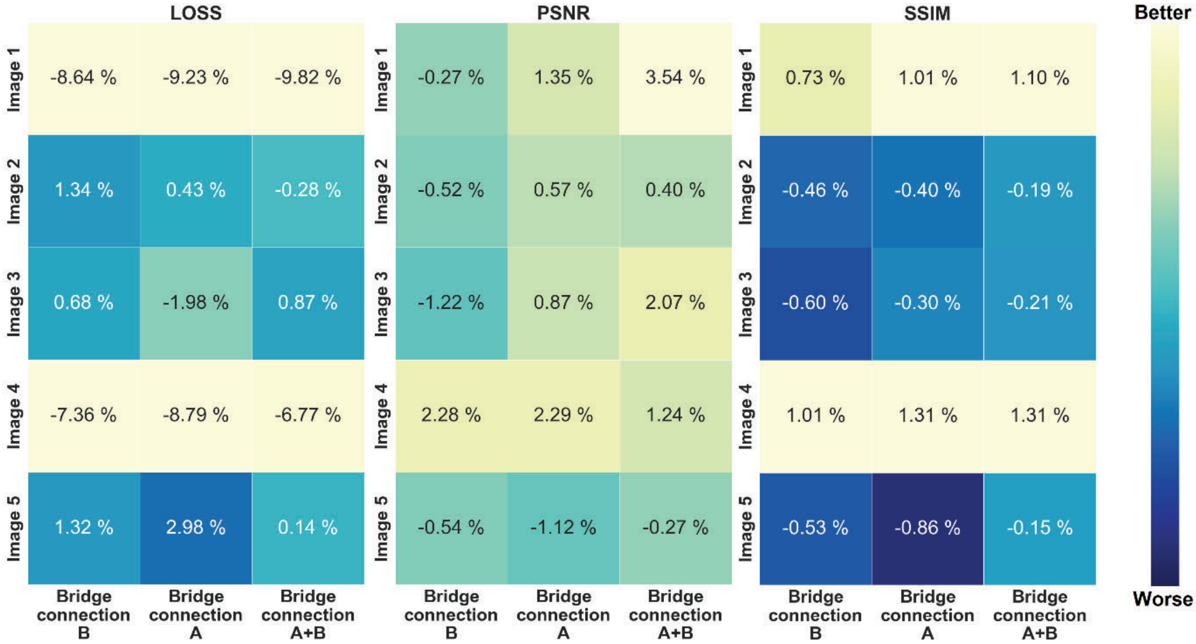


Fig. 5. Heatmaps with values of the loss function and two measures of image quality, calculated between model output and soft-tissue image: Peak signal-to-noise ratio (middle) and Structural Similarity Index Measure (right) for different bridge connection configurations

Rys. 5. Mapy ciepła przedstawiające wartości funkcji straty oraz dwóch wskaźników jakości obrazu: PSNR oraz SSIM dla różnych konfiguracji połączeń mostkowych

5.3.4. Combination of Multiple Methods

The configurations with the best results were selected from the previously described methods. An architecture was created with $\alpha = 0.5$, a strides 2121 configuration, and the use of the bridge connections shown in Fig. 2 (A+B). The applied operations allowed for a significant improvement of the used indices. The values of the obtained results are shown in Fig. 6, while the comparison of the sample images obtained is shown in Fig. 7.

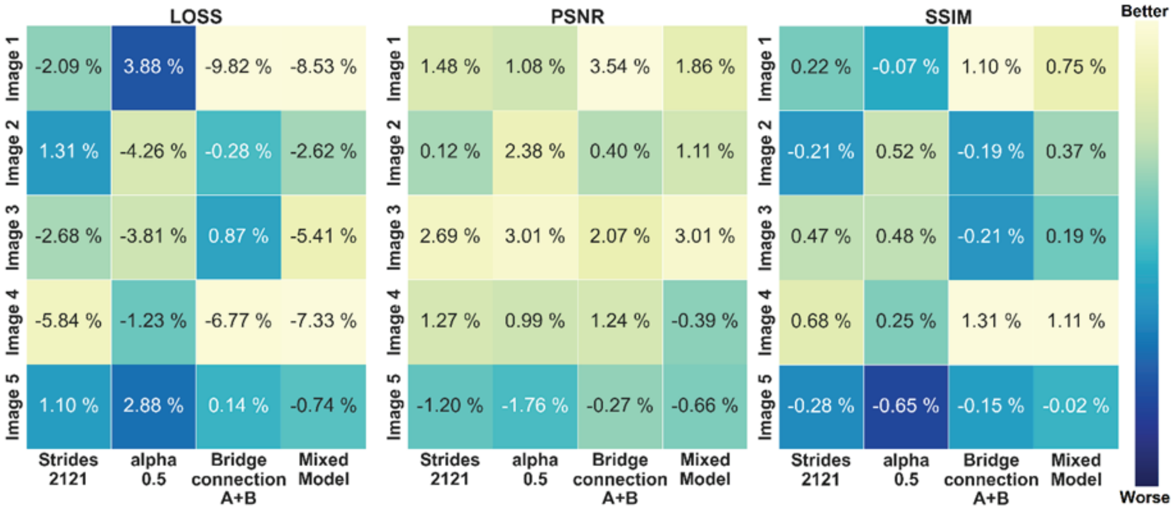


Fig. 6. Heatmaps with values of the loss function and two measures of image quality, calculated between model output and soft-tissue image: Peak signal-to-noise ratio (middle) and Structural Similarity Index Measure (right) for models characterized by the best parameters described in 1.3.1–1.3.4

Rys. 6. Mapy ciepła przedstawiające wartości funkcji straty oraz dwóch wskaźników jakości obrazu: PSNR oraz SSIM dla modeli charakteryzujących się najlepszymi parametrami opisanymi w 1.3.1–1.3.4

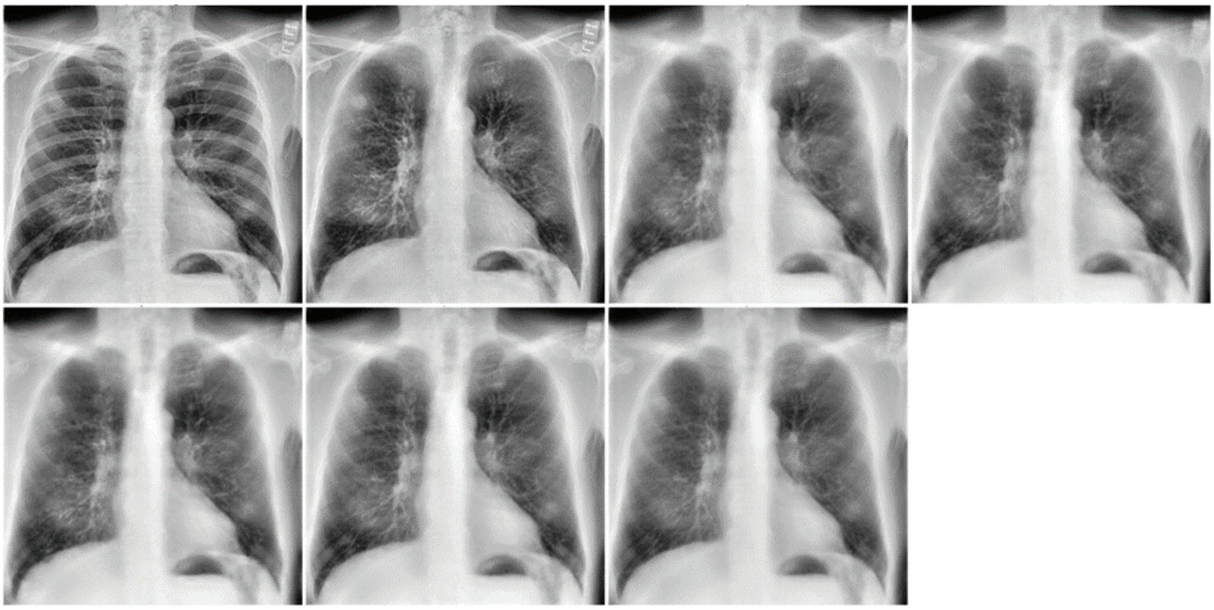


Fig. 7. Comparison of each model based on the output image. Top row: The preprocessed CXR image no. 4 from a testing set (left), soft-tissue ground truth image (middle left), the output of base architecture (middle-right), and the output of the model with strides 2121 configuration (right). Bottom row: output of architecture with α equals 0.5 (left), the output of architecture with two skipped connections (middle), the output of architecture with all presented modifications (right)

Rys. 7. Porównanie obrazów wyjściowych dla każdego z opisanych modeli. Górny rząd: Obraz wejściowy nr. 4 (pierwszy z lewej), obraz tkanek miękkich DES (drugi z lewej), obraz architektury bazowej (drugi z prawej), obraz architektury o konfiguracji strides 2121 (pierwszy z prawej). Dolny rząd: obraz architektury o parametrze α równym 0.5 (z lewej), obraz architektury złożonej z dwóch połączeń mostkowych (środkowy), obraz architektury złożónej ze przedstawionych modyfikacji (z prawej)

5.4. Conclusions

The proposed convolutional denoising autoencoder architecture was updated to eliminate the blur effect, that occurs frequently. A change in the configuration of the 'strides' parameter increases the image size on the input of the decoder, which decreases the overall image compression rate, keeping more important details of the image that are necessary for proper image reconstruction. Adding the bridge connections improves the quality of output images by transferring feature maps between symmetrical encoder and decoder layers. Changing the alpha parameter is needed to tune the loss function for the task-specific images; CXR images in our case. We have tested these methods separately, however, the combination of them brought the highest increase in the quality of the rib-suppressed image in comparison to the base model. An additional advantage of the described methods is the non-impact number of network parameters. The final designed architecture consists of only 4 convolutional layers on the encoder and decoder side. It is a rather simple structure with a small number of parameters, which distinguishes it from existing deep-learning models [9].

The rib-suppressed images presented in Fig. 7 confirm the improved quality of the output image, for each of the methods presented and the final model. In our opinion, these results are sufficient enough to be used as a medical diagnostic support system in a medical facility. An example of the use of the final images obtained with suppressed ribs may be a classification of different lung diseases or tracking disease progression over time.

Future work on the proposed algorithm will focus on separating the region of interest, which in the case of CXRs is the lung region. An architecture consisting of two independent neural networks is planned to be applied. The first network should be responsible for the segmentation of the lung region and could be based on a U-net architecture [26]. The second model will be based on an architecture presented here but trained on segmented lung images.

Acknowledgments

This work was financed by the Silesian University of Technology grant no. 02/070/BK22/0033 for maintaining and developing research potential (SK, MM).

Bibliography

1. C.M. Jones et al.: Chest radiographs and machine learning – Past, present and future. Journal of Medical Imaging and Radiation Oncology 65, 538-544 (2021).
2. S. Raoof et al.: Interpretation of Plain Chest Roentgenogram. Chest 141, 545–558 (2012).
3. E. Sogancioglu, B. Ginneken, K. van Leeuwen, K. Murphy: Deep Learning for Chest X-ray Analysis: A Survey. Medical Image Analysis 72, 102125 (2021).
4. P.K. Shah et al.: Missed non-small cell lung cancer: radiographic findings of potentially resectable lesions evident only in retrospect. Radiology 226, 235–241 (2003).
5. F. Manji, J. Wang, G. Norman, Z. Wang, D. Koff: Comparison of dual energy subtraction chest radiography and traditional chest X-rays in the detection of pulmonary nodules. Quantitative Imaging in Medicine and Surgery 6, (2016).
6. I. Sirazitdinov et al.: Evaluation of Deep Learning Methods for Bone Suppression from Dual Energy Chest Radiography. In Artificial Neural Networks and Machine Learning – ICANN 2020, I. Farkaš, P. Masulli, S. Wermter (eds.), (Springer International Publishing, Cham, 2020), pp. 247–257.
7. G. Litjens et al.: A survey on deep learning in medical image analysis. Medical Image Analysis 42, 60-88 (2017).
8. M. Gusarev, R. Kuleev, A. Khan, A.R. Rivera, A.M. Khattak: Deep learning models for bone suppression in chest radiographs. In 2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). (2017), pp. 1–7.
9. S. Rajaraman, G. Cohen, L. Spear, L. Folio, S. Antani, DeBoNet: A deep bone suppression model ensemble to improve disease detection in chest radiographs. PLOS ONE 17, e0265691 (2022).
10. W. Yang et al.: Cascade of multi-scale convolutional neural networks for bone suppression of chest radiographs in gradient domain. Med Image Anal 35, 421–433 (2017).
11. N. Matsubara, A. Teramoto, K. Saito, H. Fujita: Bone suppression for chest X-ray image using a convolutional neural filter. Physical and Engineering Sciences in Medicine 43, 97–108 (2020).
12. Y. Chen et al.: Bone Suppression of Chest Radiographs With Cascaded Convolutional Networks in Wavelet Domain. IEEE Access 7, 8346-8357 (2019).
13. J. Liang, Y.-X. Tang, Y.-B. Tang, J. Xiao, R.M. Summers: Bone suppression on chest radiographs with adversarial learning. In Medical Imaging 2020: Computer-

Aided Diagnosis. (International Society for Optics and Photonics, 2020), vol. 11314, pp. 1131409.

14. S. Kalisz, M. Marczyk: Autoencoder-based bone removal algorithm from x-ray images of the lung. In 21st International Conference on Bioinformatics and Bioengineering (BIBE). (2021), pp. 1–6.
15. N. Salem, H. Malik, A. Shams: Medical image enhancement based on histogram algorithms. Procedia Computer Science 163, 300–311 (2019).
16. A. Buslaev et al.: Albumentations: Fast and Flexible Image Augmentations. Information 11, 125 (2020).
17. I. Sirazitdinov, M. Kholiavchenko, R. Kuleev, B. Ibragimov: Data Augmentation for Chest Pathologies Classification. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). (2019), pp. 1216–1219.
18. M. Elgendi et al.: The Effectiveness of Image Augmentation in Deep Learning Networks for Detecting COVID-19: A Geometric Transformation Perspective. Frontiers in Medicine 8, (2021).
19. L. Gondara: Medical Image Denoising Using Convolutional Denoising Autoencoders. In 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW). (2016), pp. 241–246.
20. Z. Wang, E.P. Simoncelli, A.C. Bovik: Multiscale structural similarity for image quality assessment. In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. (IEEE, 2003), vol. 2, pp. 1398–1402.
21. H. Zhao, O. Gallo, I. Frosio, J. Kautz: Loss functions for neural networks for image processing. arXiv preprint arXiv:1511.08861, (2015).
22. L. Zaniolo, O. Marques: On the use of variable stride in convolutional neural networks. Multimedia Tools and Applications 79, (2020).
23. L.F. Dong, Y.Z. Gan, X.L. Mao, Y.B. Yang, C. Shen: Learning Deep Representations Using Convolutional Auto-Encoders with Symmetric Skip Connections. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). (2018), pp. 3006–3010.
24. X.-J. Mao, C. Shen, Y.-B. Yang: Image Restoration Using Convolutional Auto--encoders with Symmetric Skip Connections. arXiv preprint arXiv: 1606.08921, (2016).
25. A. Dosovitskiy, T. Brox: Generating Images with Perceptual Similarity Metrics based on Deep Networks. arXiv preprint arXiv: 1602.02644, (2016).
26. O. Ronneberger, P. Fischer, T.Brox, U-Net: Convolutional Networks for Biomedical Image Segmentation. LNCS. 9351. 234-241. (2015).

METHODS FOR REDUCTION OF BLUR EFFECT IN THE DENOISING AUTOENCODER MODEL FOR RIB SUPPRESSION IN CXR IMAGES

Abstract

The most widespread, low-cost, and highly available diagnostic method for detecting abnormalities in the cardiopulmonary system is chest radiography. In many cases, the presence of bone structures on the image makes the correct diagnosis much more difficult. There exist few computational tools for rib suppression on x-ray images of the lungs and the denoising autoencoder model seems to be perfect for this task. However, mostly due to the considerable compression of image size in the model, the blurring effect of pathological structures occurs. This work aims to improve the quality of the lung images obtained from a rib bone suppression algorithm based on a denoising autoencoder. The Bone Suppression dataset, consisting of 35 pairs of chest x-ray images and corresponding soft-tissue images, was used to develop the model. Three methods were proposed to reduce the blurring effect: (i) modification of the image resizing scheme through layers of the autoencoder; (ii) modification of a parameter of the loss function, which was responsible for establishing the weights between the mean square error index and the multi-scale structural similarity index; (iii) introduction of bridge connections between the coder and encoder layers. Different scenarios of proposed corrections were examined, but in general, all tested methods showed increased performance in comparison to the baseline model in terms of peak signal-to-noise ratio and structural similarity. Our results show that these methods significantly reduce the denoising autoencoder blurring effect, improving the quality of output image with suppressed ribs, which could potentially influence the medical diagnosis of the patients.

Keywords: x-ray imaging, blurring effect reduction, bone suppression, autoencoders, deep learning

DOI:10.34918/85106

Aleksandra SUWALSKA[1,*], Marek SOCHA[1], Wojciech PRAZUCH[1],
Joanna TOBIASZ[1,2], Joanna POLANSKA[1], Michal MARCZYK[1,*]
and POLCOVID Study Group

Chapter 6. nUMAP: NEURAL NETWORK BASED UMAP SOLUTION FOR THE MULTI DATASET VISUALISATION

6.1. Introduction

High-dimensional data is common in the field of biomedicine. It is difficult to analyze, therefore it is hard to find its flaws and hidden relations. Dimensionality reduction techniques help researchers to overcome this problem by embedding high dimensional information into the lower-dimensional space. Such embeddings can be visualised and then analysed properly.

One of such embedding techniques is Uniform Manifold Approximation and Projection (UMAP) [1]. Besides embedding it can also be used for clustering and data pre-processing. This method thrives when used with tabular data as an input and despite that it can be used with image data [3-5], it has limited use cases. The common factor of referenced papers is the usage of homogenous image datasets with neutral backgrounds. It is very rare to stumble upon such data in biomedicine.

In the paper [2], a solution to the problem of image data heterogeneity was proposed by introducing a novel method capable of robust transformation of X-ray radiogram into a set of features and UMAP based pipeline capable of embedding features into low-dimensional space. While the method is valid for X-ray images, it requires a definition of a region of interest (ROI). For other biomedical datasets, this may not be possible. In the study, some modifications are introduced to unify the previously

---

[1] Department of Data Science and Engineering, Faculty of Automatic Control, Silesian University of Technology, Gliwice, Poland.
[2] Department of Graphics, Computer Vision and Digital Systems, Faculty of Automatic Control, Silesian University of Technology, Gliwice, Poland.
* Corresponding authors: aleksandra.suwalska@polsl.pl, michal.marczyk@polsl.pl.

proposed method. The proposed pipeline consists of neural network (NN) guided features extraction and UMAP embedding followed by NN universal embedding learning.

Guided feature extraction relies on the information obtained from the latent vector of a pre-last layer of a neural network train specifically for this task [6] (latent space with UMAP in the context of fashion recommendation system), [7] (using latent space vector for UMAP in clustering of genes). UMAP procedure embeds latent space features into the two-dimensional space allowing visualisation and relation analysis. To achieve robust features dependence a regression neural network was trained which learnt the embedding. The pipeline results in a method capable of dealing with numerical, image and mixed data types. The goal of the study was to present different use cases of the proposed method, like discovering the batch effect, analysing the dataset's quality, and explaining the neural network prediction, in order to prove its wide applicability.

6.2. Datasets

6.2.1. Mass cytometry dataset

In the study, a mass cytometry dataset was used that comprised two healthy control samples with the number of bronchoalveolar lavage cells (BALC) equal to 329,228 and 341,007, respectively. The samples come from studies on drug-resistant tuberculosis. Bronchoscopies were performed in the bronchoscopy theatre, ward A5, Tygerberg Hospital (TBH) in Cape Town, South Africa. The cells' signal was measured with the CyTOF2 instrument, at the South African Tuberculosis Vaccine Initiative at the University of Cape Town. For each cell, a set of 32 markers was collected. The dataset was preprocessed (pre-gated) and arcsinh transformed with a co-factor of 5.

6.2.2. Chest X-Ray dataset

The chest X-Ray image dataset was composed of POLCOVID database [2] and COVIDx database [8]. The data from POLCOVID database were collected from 24 Polish hospitals (see POLCOVID Study Group section) during a CIRCA project and consists of 2426 healthy patients, 1147 patients with pneumonia and 1236 patients with COVID-19 (with positive RT-PCR test results). The COVIDx database consists of 8066 healthy patients, 5573 patients with pneumonia and 1763 patients with COVID-19. The images were resized to 512x512 pixels resolution, had their lungs segmented from an images and were scaled.

6.3. nUMAP

The proposed nUMAP method is a modification of the standard UMAP approach (Fig. 1). Since UMAP only accepts numerical data as input, the limitation can be solved with a neural network that can take different types of data as input. The final fully-connected layer provides feature vectors that are combinations of the input data in a numerical form that can be further analysed by the standard UMAP model.

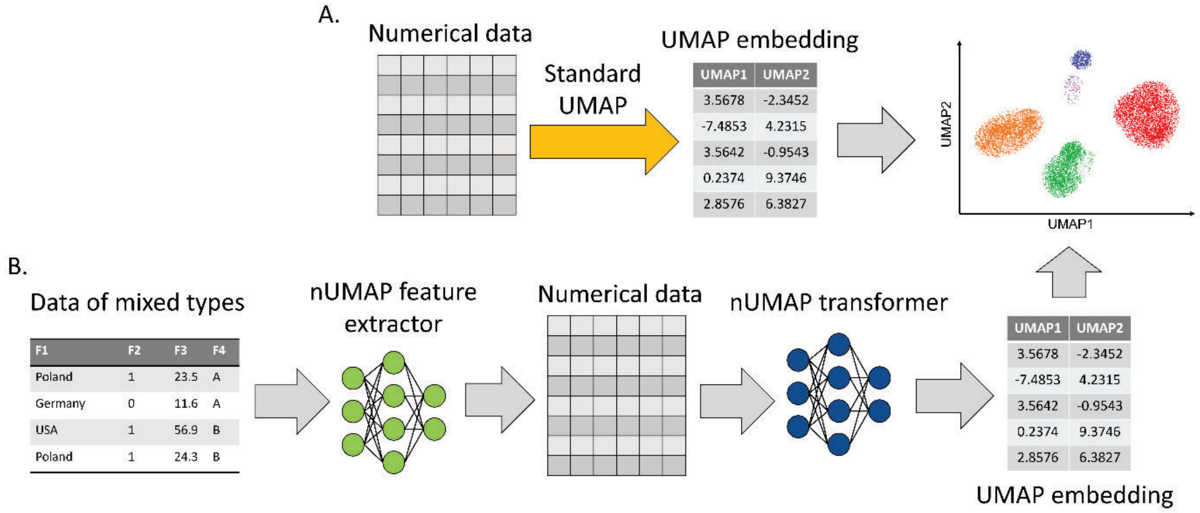


Fig. 1. nUMAP pipeline. A) Standard UMAP model trained on numerical data to generate its embedding that can be visualized in 2D space. B) nUMAP accepts data of mixed types that are processed into numerical representation by the first neural network (nUMAP feature extractor). Then the second neural network transforms the numerical data into embedding, giving identical results to the standard UMAP approach

Rys. 1. Schemat działania nUMAP: A) Standardowa metoda UMAP trenowany na danych numerycznych w celu zwizualizowania ich zredukowanej reprezentacji w przestrzeni 2D. B) nUMAP akceptuje dane o mieszanych typach dzięki zastosowaniu sieci neuronowej (ekstraktor cech nUMAP), która przetwarza je na dane numeryczne. Następnie druga sieć neuronowa transformuje dane numeryczne do reprezentacji o niższej wymiarowości, która jest identyczna do rezultatu otrzymanego ze standardowej metody UMAP

The first neural network is called the nUMAP feature extractor which enables simultaneous processing of different types of data. The nUMAP feature extractor can be any model that can process desired input data into numerical features that can be further analysed. For example, like in the study, it can be a convolutional neural network that accepts images and numerical features as inputs to combine them into new numerical features.

The second part of the nUMAP approach is another neural network, the nUMAP transformer, that learns how to map numerical features into UMAP embeddings. The network is a multilayer perceptron for regression problems, that accepts numerical data as input and UMAP embedding as output. The transformer's architecture used in the study was a simple neural network with two hidden layers consisting of 100 and 50 neurons, respectively. The number of input neurons was equal to the number of numeric features and the number of the output neurons was two (for the reduction into 2D space embedding). The learning rate was set to 0.001.

6.4. nUMAP application examples

6.4.1. Use case 1: visualization of a batch effect in mass cytometry data

Mass cytometry datasets are numerical, therefore the standard UMAP approach can be applied as presented in Fig. 2. The embedding is created and can be visualized in 2D space. The pink area represents the region with the highest density of cells (50% of data) from the sample in reference to all cells (grey points).

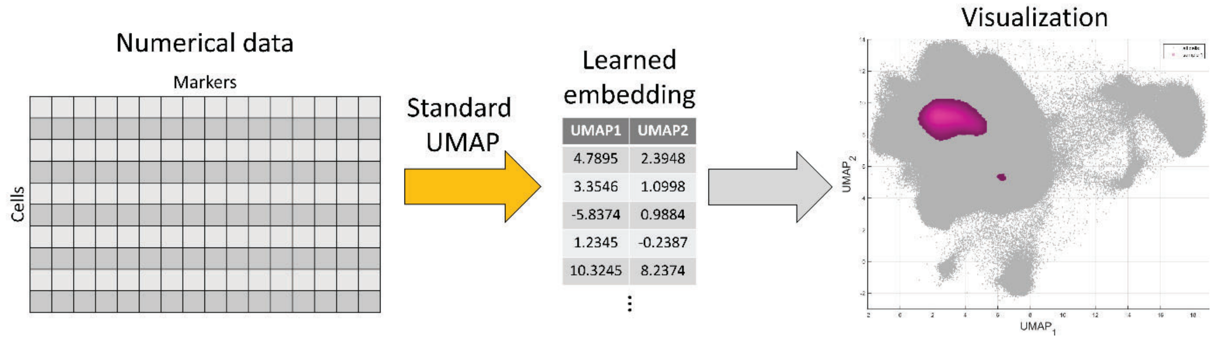
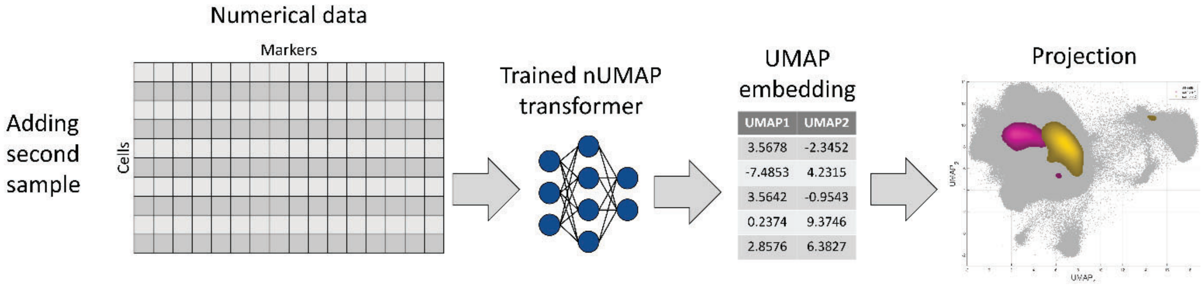


Fig. 2. Standard UMAP approach for mass cytometry data. One sample is transformed into embedding that allows visualizing the dataset in 2D space. Visualization: grey points – all cells; pink area – region with the highest density of cells

Rys. 2. Standardowa metoda UMAP dla danych z cytometrii masowej. Jedna próbka jest przetworzona do dwuwymiarowej reprezentacji, którą można zwizualizować na wykresie. Wizualizacja: szare punkty – wszystkie komórki; różowy region – obszar z największym zagęszczeniem komórek próbki

The problem appears when adding another sample from the same experiment to the first one to represent new data in the learned UMAP embedding. The existing UMAP method that transforms new data with the learned model is very time-consuming for big data. Since mass cytometry datasets may have tens of millions of records (cells), the standard UMAP transformer is inefficient. Therefore, the second part of the nUMAP system applies here as presented in Fig. 3. The trained nUMAP transformer generates embedding quickly and accurately.



UMAP1	UMAP2
3.5678	-2.3452
-7.4853	4.2315
3.5642	-0.9543
0.2374	9.3746
2.8576	6.3827

Fig. 3. nUMAP transformer allows fast embedding creation and projection for mass cytometry (big data). Visualization: grey points – all cells; pink area – a region with the highest density of cells from the first sample; yellow area – a region with the highest density of cells from the second sample that was transformed into the same UMAP space as the first sample

Rys. 3. Transformer nUMAP pozwala na szybkie wygenerowanie dwuwymiarowej reprezentacji danych z cytometrii masowej (duże dane). Wizualizacja: szare punkty – wszystkie komórki; różowy region – obszar z największym zagęszczeniem komórek próbki pierwszej; żółty region – obszar z największym zagęszczeniem komórek próbki drugiej transformowanej do tej samej przestrzeni UMAP co próbka pierwsza

Adding the second sample reveals the presence of a batch effect in the dataset. The batch effect is a technical variation in the data that makes it difficult to reveal biological relationships and should be removed or decreased. Many solutions have been proposed for batch effect removal for mass cytometry. nUMAP makes it possible to visualize the effect of such a batch correction algorithm (Fig. 4).

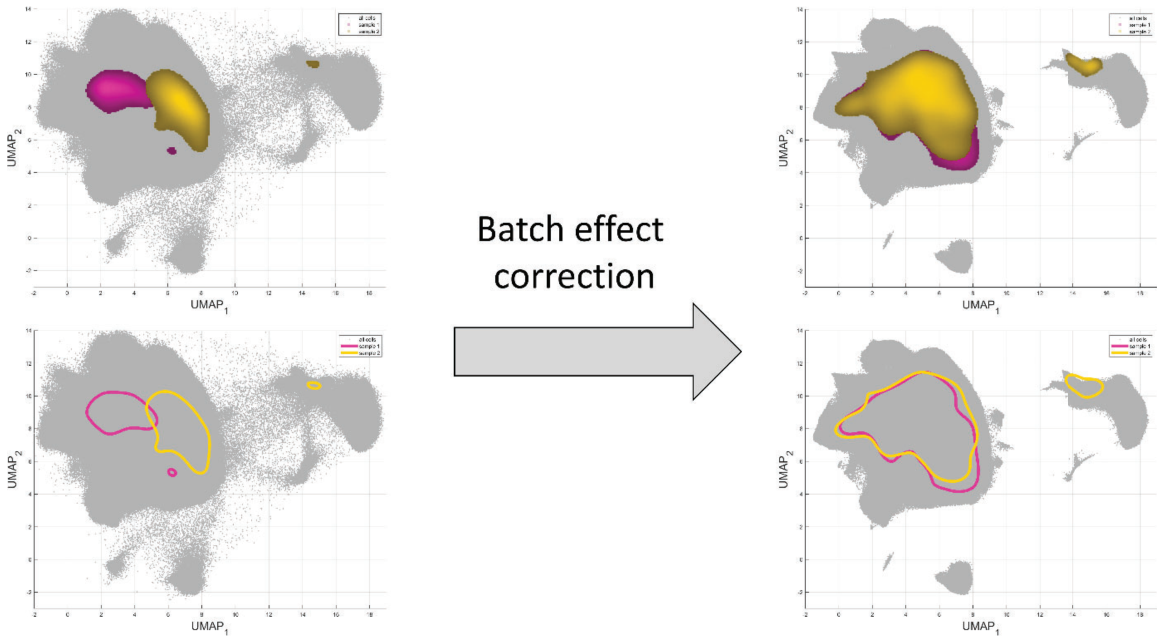


Fig. 4. Visualization of the mass cytometry dataset comprised of two samples before and after batch effect correction (pink and yellow areas representing the regions of their highest cell concentration – for better visualization). After batch effect correction a new UMAP embedding and nUMAP transformer are learned based on the corrected feature values. Since the samples overlap after the batch correction, it is visible that the batch effect was significantly reduced

Rys. 4. Wizualizacja danych z cytometrii masowej złożonych z dwóch próbek, przed i po korekcie efektu paczki (różowe i żółte obszary reprezentują regiony o największym zagęszczeniu komórek – dla lepszej wizualizacji). Wykorzystując wartości po korekcji efektu paczki wytrenowano nowy model nUMAP do stworzenia nowej reprezentacji UMAP. Ponieważ po korekcie próbki nakładają się na siebie, można wywnioskować że efekt paczki został znacząco zredukowany

6.4.2. Use case 2: RTG quality assessment

The X-Ray chest dataset consists of image data (radiograms) and clinical data. The goal was to validate whether the dataset contains hidden biases and flaws which are not related to the lung disease. Therefore a neural network was created to classify the data into three classes: normal (healthy), pneumonia and COVID-19. To verify data robustness, the nUMAP was employed as in Fig. 5. The neural network was used as the nUMAP feature extractor.

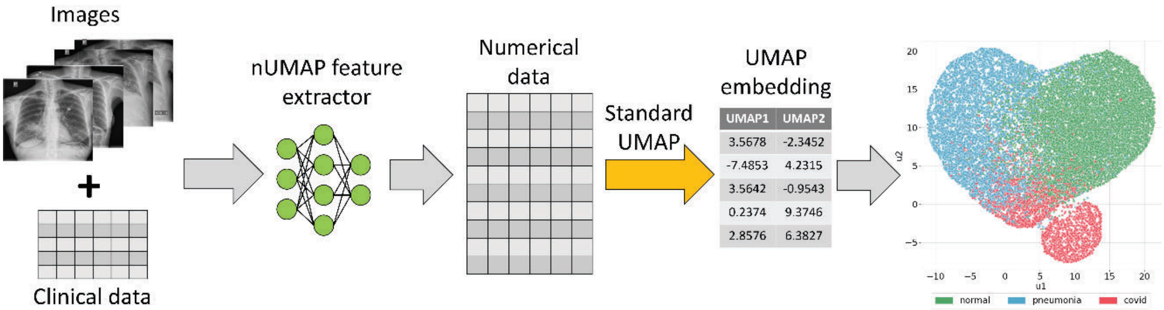


Fig. 5. Image information is merged with clinical data through the nUMAP feature extractor, and the standard UMAP method (with cosine distance metric) is used to project received numerical features on the 2D plane

Rys. 5. Informacja obrazowa jest połączona z danymi klinicznymi z wykorzystaniem ekstraktora cech nUMAP, następnie standardowa metoda UMAP (z metryką odległości cosine) transformuje otrzymane cechy numeryczne do przestrzeni o niższej wymiarowości

Ideally, UMAP embedding would be of compact shape with three fuzzy borders splitting the data into the three classes. While most of the data points follow mentioned behaviour, there is a group of COVID-19 data points which are visibly distinct from the other representatives of the category, forming a separate aggregate of points. Further analysis revealed that data within this 'island' consist of radiograms with relatively low original resolution (Fig. 6).

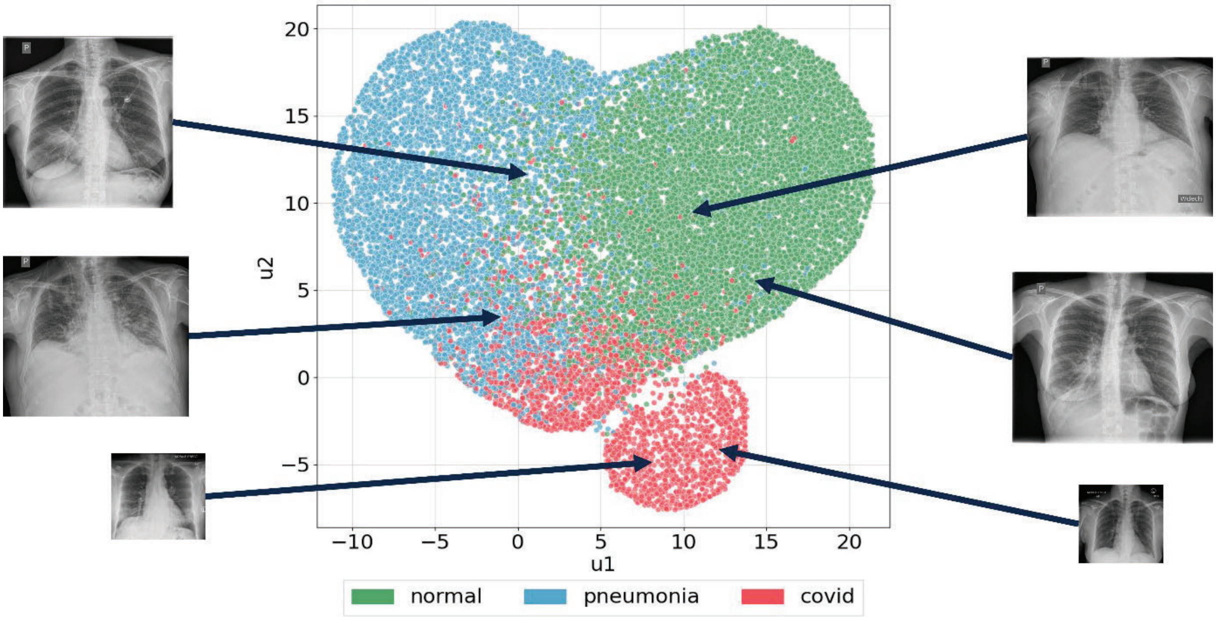


Fig. 6. Example radiograms representing embedded data points on the UMAP 2D plot. The radiograms were scaled following their original proportions to each other. The two smallest images of COVID-19 patients come from a separate aggregate of points

Rys. 6. Przykłady obrazów reprezentujących wartości UMAP po redukcji wymiarowości na wykresie dwuwymiarowym. Radiogramy zostały wyskalowane względem siebie zachowując oryginalne proporcje. Dwa najmniejsze obrazy pacjentów z COVID-19 pochodzą z osobnego skupiska punktów

A resolution of a radiogram has a big impact on its quality. Lower radiogram resolution results in a lower number of pixels which represents lung abnormal changes. Since the neural network input image resolution was 512 x 512 pixels, it was assumed that the island of points is mainly composed of images below the 512 x 512 resolution. To verify the thesis, points on the UMAP visualization were coloured according to the calculated radiogram resolutions (Fig. 7).

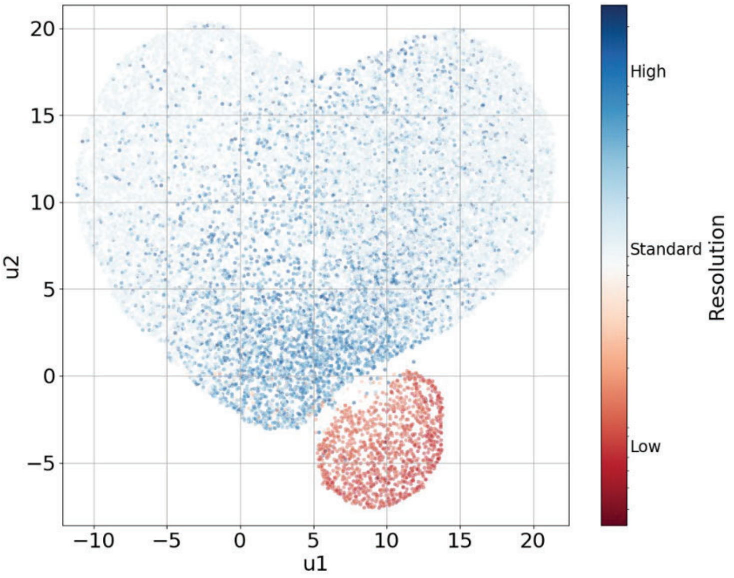


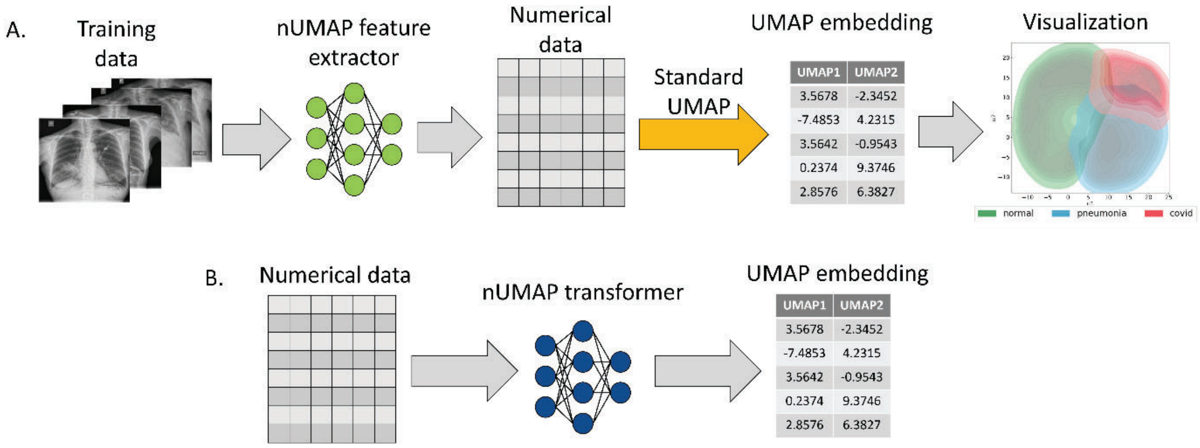
Fig. 7. UMAP embedding representation of X-ray dataset with colour scale pointing to radiogram resolution ranging from about 100x100 to nearly 4000x4000 pixels. Shown on a logarithmic scale

Rys. 7. Reprezentacja przestrzeni UMAP zdjęć rentgenowskich płuc z kolorową skalą wskazującą na rozdzielczość obrazów od około 100x100 do prawie 4000x4000 pikseli. Skala logarytmiczna

The radiograms contained within the smaller aggregate of points are of much lower resolution than radiograms in the compact part of UMAP embedding. This indicates that the neural network converged on the resolution rather than the disease entity. Showing that the problem is hidden in the dataset and calls for data curation.

6.4.3. Use case 3: Explainability of a classifier's prediction

nUMAP can work like an Explainability AI method to clarify the model's prediction. This behaviour is used in the CIRCA classification system [9]. CIRCA nUMAP allows for the projection of new data points, that are analysed through the CIRCA portal, on the learned UMAP visualization. The learned UMAP was created with the use of X-Ray chest images consisting of three patient categories: normal, pneumonia and COVID-19. The nUMAP training process is presented in Fig. 8.



UMAP1	UMAP2
3.5678	-2.3452
-7.4853	4.2315
3.5642	-0.9543
0.2374	9.3746
2.8576	6.3827

UMAP1	UMAP2
3.5678	-2.3452
-7.4853	4.2315
3.5642	-0.9543
0.2374	9.3746
2.8576	6.3827

Fig. 8. Training of the CIRCA nUMAP system. A) Training dataset of X-Ray chest images is used to train the feature extractor. Then numerical features are transformed with the standard UMAP into embedding to create a 2D visualization. B) Numerical data extracted from the nUMAP feature extractor are used to train the nUMAP transformer to generate the same UMAP embedding as the standardapproach

Rys. 8. Trening metody nUMAP systemu CIRCA. A) Zbiór treningowy złożony z obrazów RTG klatki piersiowej jest wykorzystany do treningu ekstraktora cech nUMAP. Następnie cechy numeryczne są poddane transformacji standardową metodą UMAP do dwuwymiarowej reprezentacji. B) Dane numeryczne otrzymane z nUMAP ekstraktora cech są wykorzystane do treningu nUMAP transformatora w celu wygenerowania takiej samej reprezentacji UMAP jak z metody standardowej

When a new chest X-Ray image is loaded into the CIRCA portal, it undergoes a series of preprocessing and classification steps. The goal is to classify the image into the three mentioned categories based on the visual markers. Since the task is difficult due to the heterogeneous nature of COVID-19 changes, displaying the resulting category may not be sufficient. nUMAP tries to explain the prediction by projecting the image into the UMAP visualization that is comprised of three different areas indicating the categories. Moreover, nUMAP revealed subcategories within each category that differentiate the patients based on their degree of advancement of pulmonary changes. The projection process is visualized in Fig. 9.

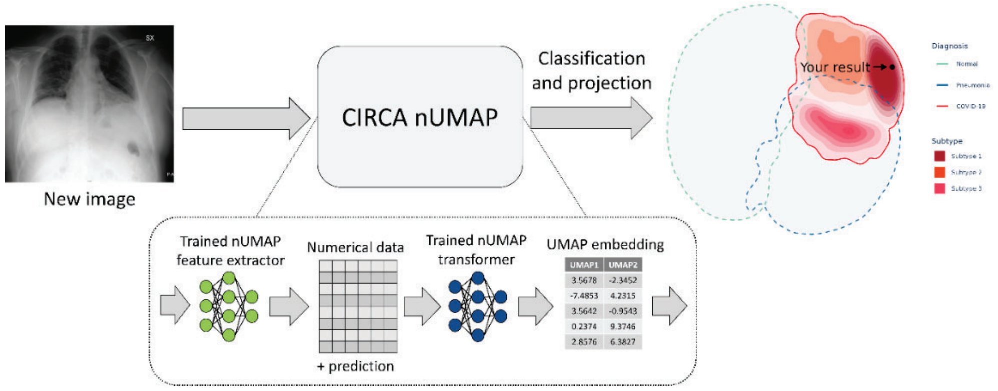


Fig. 9. A new image is processed by CIRCA system that consists of an nUMAP part responsible for projecting the analysis result into learned UMAP representation to visualize it in relation to the training samples. The image is classified as COVID-19 therefore the embedding is placed in the red area (regions indicating normal and pneumonia patients are shaded for better visualization). It can be seen that the result is placed in Subtype 1 of the COVID-19 category, which is the farthest from normal and pneumonia areas – therefore this is the subtype that contains patients with typical COVID-19 changes in the lungs

Rys. 9. Nowe zdjęcie RTG jest przetwarzane przez system CIRCA, który zawiera część nUMAP odpowiedzialną za projekcję do reprezentacji UMAP wyniku analizy w celu jego wizualizacji względem zdjęć ze zbioru treningowego. Obraz został zaklasyfikowany do kategorii COVID-19, dlatego jego dwuwymiarowa reprezentacja została umieszczona w rejonie czerwonym (regiony dla kategorii zdrowe płuca oraz zapalenie płuc zostały zacienione dla lepszej wizualizacji). Wynik został umieszczony w rejonie Podtypu nr. 1 kategorii COVID-19, który leży najdalej od obszarów kategorii zdrowe płuca i zapalenie płuc – w związku z tym jest to podtyp który zawiera obrazy pacjentów z typowymi zmianami COVID-19 w płucach

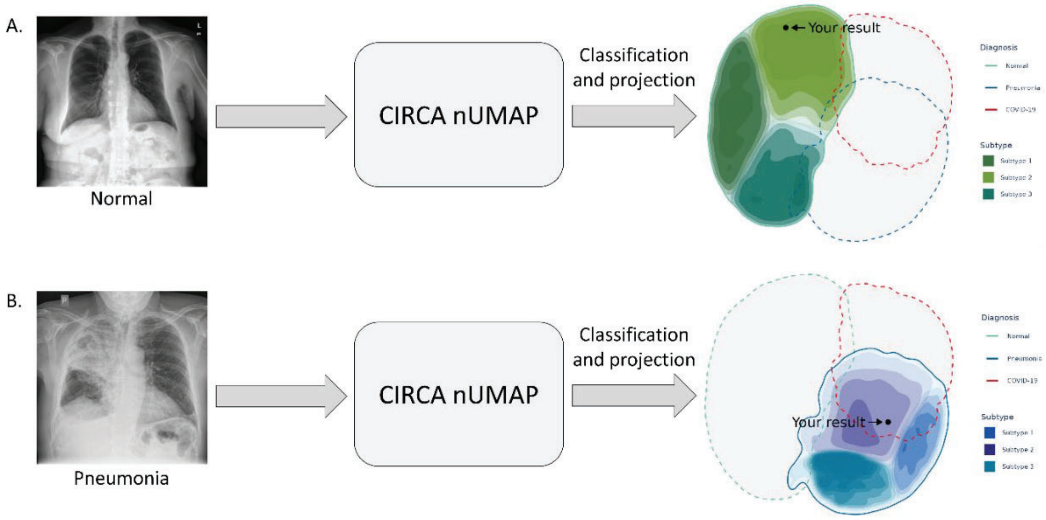


Fig. 10. Examples from CIRCA portal projection of the RTG analysis result in UMAP representation. A) Chest X-Ray image of normal (healthy) lungs is classified and projected with nUMAP into the healthy region of the embedding visualization. B) Chest X-Ray image of lungs with pneumonia changes is classified and projected with nUMAP into the pneumonia region of the embedding visualization

Rys. 10. Inny przykład z systemu CIRCA pokazuje projekcje Przykłady z portalu CIRCA, projekcja wyniku analizy RTG w reprezentacji UMAP. A) Obraz rentgenowski klatki piersiowej prawidłowych (zdrowych) płuc jest klasyfikowany i rzutowany za pomocą nUMAP do rejonu osadzenia kategorii zdrowych. B) Obraz rentgenowski płuc ze zmianami zapalenia płuc jest klasyfikowany i rzutowany za pomocą nUMAP do rejonu osadzenia kategorii zapalenia płuc

6.5. Discussion

Big data analysis is a demanding task that requires appropriate and efficient techniques to gain meaningful insights and results. This often requires presenting data in a lower-dimensional feature space. Methods like UMAP allow for dimension reduction preserving hidden structures in the data, however, they have some limitations.

In the study, a modification of the standard UMAP approach was proposed, that overcomes its limitations. The nUMAP parts can be linked with the standard UMAP approach or customized to a specific problem, as needed.

Firstly, the proposed nUMAP transformer can be trained on numerical features and embedding values from the standard UMAP approach to generate identical embedding but more effectively, especially in the case of big data, like mass cytometry. It solves the problem of the standard UMAP projection of new points into the existing low--dimensional representation which is time-consuming.

Secondly, nUMAP accepts mixed types of data as input due to the use of a neural network that works as a feature extractor. Therefore, different data types can be combined together into a numerical representation that can be used either with the standard UMAP approach or the nUMAP transformer to generate data embeddings. The shown examples of X-Ray data projection prove the usefulness of the approach.

The presented use cases of the nUMAP allowed for deeper insight into the data, revealing problems like batch effect in mass cytometry data or low-quality images of X-Ray dataset that have a great impact on further analysis results. Moreover, the method can be used to explain the classifier’s prediction and therefore the usefulness of the trained classifier.

However, the nUMAP approach also has limitations. Since the feature extractor is a classification network, the method is no longer fully unsupervised and requires some knowledge about the labels (categories) of observations. Future work could focus on the use of a different feature extraction method that accepts mixed types of data as inputs.

6.6. Conclusion

nUMAP is a fast and effective method for embedding generation that can help visualize modifications of big data of mixed types. nUMAP overcomes the standard UMAP limitations and can be used for various purposes.

POLCOVID Study Group

Department of Infectious Diseases and Hepatology, as coordinator: Jerzy Jaroszewicz (Medical University of Silesia in Katowice, Infectious Diseases Hospital No. 1 in Bytom), Jan Baron, Katarzyna Gruszczynska (Department of Nuclear Medicine and Image Diagnostics, Medical University of Silesia in Katowice), Magdalena Sliwinska, Mateusz Rataj, Przemyslaw Chmielarz (Voivodship Specialist Hospital in Wroclaw), Edyta Szurowska (II Department of Radiology, Medical University of Gdańsk), Jerzy Walecki, Samuel Mazur, Piotr Wasilewski (Central Clinical Hospital of the Ministry of Internal Affairs and Administration in Warsaw), Tadeusz Popiela, Justyna Kozub (Collegium Medicum of the Jagiellonian University in Kraków), Grzegorz Przybylski, Anna Kozanecka (Kujawsko-Pomorskie Pulmonology Center in Bydgoszcz), Andrzej Cieszanowski, Agnieszka Oronowicz-Jaskowiak, Bogumil Golebiewski (National Institute of Oncology in Warsaw, Department of Imaging Diagnostics), Complex of Health Care Centres, Mateusz Nowak (Silesian Hospital in Cieszyn), Barbara Gizycka (Single Infectious Diseases Hospital Megrez Ltd. in Tychy: Department of Imaging Diagnostics), Piotr Blewaska (District Hospital in Raciborz), Department of Infectious Diseases and Hepatology, University of M. Kopernika w Toruniu, Malgorzata Pawlowska, Piotr Rabiko, Pawel Rajewski (Collegium Medicum in Bydgoszcz), Department of Radiological and Imaging Diagnostics, Jerzy Walecki (Medical Center for Postgraduate Education, Warsaw), Clinical Department of Imaging Diagnostics, Katarzyna Sznajder (University Clinical Hospital in Opole), Department of Infectious Diseases University of Rzeszow, Robert Plesniak (Medical Center in Lancut), Department of Allergology and Internal Medicine, Marcin Moniuszko (Medical University of Bialystok), Department of Infectious Diseases and Hepatology, Robert Flisiak (Medical University of Bialystok), Andrzej Cieszanowski (Medical University of Warsaw: II Department of Clinical Radiology), Przemyslaw Bombinski (Department of Pediatric Radiology), Agata Majos (Medical University of Lodz: Department of Radiological and Isotopic Diagnostics and Therapy), Michal Mik (Department of General and Colorectal Surgery), Medical University of Wroclaw, Krzysztof Simon (Department of Infectious Diseases and Hepatology), Bartosz Markiewicz (Voivodship Comprehensive Hospital in Kielce: Department of Imaging Diagnostics), Gabriela Zapolska, Krzysztof Klaude, Katarzyna Rataj (Czerniakowski Hospital in Warsaw), Sebastian Hildebrandt, Katarzyna Krutul-Walenciej (Central Clinical Hospital of the Medical University of Gdansk), Adrianna Tur, Grzegorz Drabik (Prognostic Specialist Clinic in Knurow), Damian Piotrowski (Specialist Hospital No. 1 in Bytom).

Acknowledgement

The research leading to these results was partially funded by the National Science Centre, Poland, grant MNiSW/2/WFSN/2020 project name CIRCA – COVID-19 online image diagnostic support service. MM was financed by grant no. 02/070/BK_22/0033. WP and JP were financed by OPUS grant no. 2017/27/B/NZ7/01833. Additionally, AS and WP are holders of a European Union scholarship through the European Social Fund, grant POWR.03.05.00-00-Z305 and JT is a holder of scholarship grant POWR.03.02.00-00-I029. Calculations were carried out using GeCONiI infrastructure funded by NCBiR project no. POIG.02.03.01-24-099/13).

Bibliography

1. L. Mclnnes, J. Healy, J. Melville: Umap: Unifold manifold approximation and projection for dimension reduction, *arxiv preprint* (2018).
2. M. Socha, A. Suwalska, W. Prazuch, M. Marczyk, J. Polanska: UMAP-based graphic representation of POLCOVID chest X-Ray data set heterogeneity, *Recent Advances in Computational Oncology and Personalised Medicine* (2021) **1**:100–114.
3. H. Xiao, K. Rasul, R. Vollgraf: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, *arxiv preprint* (2017).
4. Y. Wang, H. Huang, C. Rudin, Y. Shaposhnik: Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, Atrimap, and pacmap for data visualization, *arxiv preprint* (2020).
5. M. Allaoui, M.L. Kherfi, A. Cheriet: Considerably improving clustering algorithms using UMAP dimensionality reduction technique: a comparative study. *In International Conference on Image and Signal Processing, Springer, Cham* (2020) 317–325.

6. A.R. Sulthana, M. Gupta, S. Subramanian, S. Mirza: Improvising the performance of image-based recommendation system using convolution neural networks and deep learning. *Soft Computing* (2020) **24(19)**:14531–14544.
7. J. Ding, A. Regev: Deep generative model embedding of single-cell RNA-Seq profiles on hyperspheres and hyperbolic spaces, *Nature Communications* (2021) **12**:2554.
8. L. Wang, Z.Q. Lin, A. Wong: COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. *Sci Rep* (2020) **10**:19549.
9. WWW: https://circa.aei.polsl.pl/, access: 16.05.2022.

nUMAP: NEURAL NETWORK BASED UMAP SOLUTION FOR THE MULTI DATASET VISUALISATION

Abstract

High-dimensional data is associated with complex analysis and interpretation of the results. The problem is particularly relevant to biomedical problems like the analysis of cell expression profiles or medical imaging data. Big data contains potential noise that can cause the lack of optimal solutions. One way to deal with the problem is dimensionality reduction and feature selection. However, most of the methods are designed to process numerical data and for images, the methods require vectorization that amplifies the artefacts. Moreover, the techniques accept only one type of data at a time which could be insufficient to discover the real relationships. In the study, a novel method is proposed that is based on the UMAP dimension reduction technique. nUMAP combines the UMAP transformation with neural networks (NN) allowing for the processing of big data of mixed types. The method is based on a sequence of NN-UMAP-NN operations that extract features and create an embedding for effective visualization and inspection of the results. Moreover, the method can be applied to new data without the need for retraining. In the study, three real-world use cases of the nUMAP are presented: detection and correction of a batch effect in mass cytometry data, an inspection of the quality of chest X-Ray images and an explanation of a classifier's prediction. The work proves the effectiveness and wide application of the nUMAP.

Keywords: visualization, UMAP, neural network, embedding, big data

DOI:10.34918/85107

Wojciech SIKORA[1,*], Joanna POLAŃSKA[1]

Chapter 7. MASS CHANNEL SPATIAL DISTRIBUTION AS A TOOL FOR PEAK DETECTION AND ISOTOPE IDENTIFICATION IN MALDI TOF MSI DATA

7.1. Introduction

Mass spectrometry (MS) is a powerful tool widely used in medical research to inspect the molecular composition of samples such as biopsies from patients with suspicion of cancer. The identification of proteins, which are often drug targets and are present at low concentration levels in complicated mixtures, is significantly facilitated by the sensitivity of MS [1]. Information gathered by MS helps with diagnosis, early detection, and drug development not only for cancer but also for other diseases. This technique of molecular mass analysis exists for over a hundred years, and consists of three distinct steps: ionization, ion separation, and ion detection. In the first step, molecules of the sample are given an electric charge (ionization). This charge is used to differentiate between molecules of different masses in the next mass spectrometer module, the mass analyzer (separation). In the final step of MS analysis, the detector measures the quantity of each distinct ion (detection). This gives information about the relative abundance of each molecule in the sample. The result is a function of relative abundance and mass-to-charge (m/z) ratio of ions, called the mass spectrum.

Advances in soft ionization techniques like electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI) and improved mass analyzers made it possible to detect higher-mass molecules such as proteins and peptides, which led to the wide use of MS in the field of Proteomics, [2] the study of proteins, their functions, and their compositions within biological structures.

---

[1] Department of Data Science and Engineering, Faculty of Automatic Control, Silesian University of Technology, Gliwice, Poland.

* Corresponding author: wojciech.sikora@polsl.pl.

Mass spectrometry imaging provides information about the spatial distribution of molecules at the surface of the sample. The general setup of the MSI experiment involves defining an (x, y) grid over the surface of the sample. The mass spectrometer then ionizes the molecules on the surface of the sample and collects the mass spectrum at each pixel on the sample, with the resulting spatial distribution defined by the pixel size [3]. Application of this technique to biological samples from a biopsy can provide information about the spatial distribution of molecules such as peptides and proteins, which in turn can be used to gain even more insight into the condition of the patient, causes, and course of the illness.

Matrix-assisted laser desorption/ionization (MALDI) is a key ionization technique in MS-based proteomics. The advantages of this ionization method are among others high-throughput, high-speed data acquisition, easy sample preparation, and imaging at a high spatial resolution [4]. All these properties make it possible to obtain a large amount of data in a short amount of time. Data that can provide useful insights if properly analyzed. The volume of the data generated by 2D imaging makes manual analysis infeasible therefore in recent years there has been a great increase in the usage of statistics and machine learning methods to analyze such data.

The paper aims to describe a full pipeline for analysis of MALDI time-of-flight (ToF) MSI data that uses the full potential of its data and produces a set of features that is nonredundant, small enough for use in various machine learning techniques, consists of features that can be connected to specific proteins and can be used to rapidly calculate values of the features given new records. The proposed method starts with the aggregation of all mass spectra into a single signal. Then the signal is transformed using a Gaussian filter and divided into small fragments, then considered separately. Data prepared in this way undergoes a feature extraction process where fragments of the aggregated signal are fitted with a gaussian mixture model (GMM) using the expectation-maximization (EM) algorithm, thus generating a representation of all data as a set of potential features defined by the position and the value of the area of its normal distribution. In the dimensionality reduction step, the set of normal distributions describing each fragment is compared using information about spatial distribution provided by MSI. Spatial distributions are compared using Peacock's test statistic [5] and merged if the value of this test statistic is below a data-driven threshold. Further reduction in dimensionality is done by noise filtering. The final step of this proposed pipeline is isoform search where isoforms of the same molecule are reduced to a single feature, once again by comparison of spatial distribution on the sample.

7.2. Analysis of MALDI MSI data

Analysis of MS data can be problematic, especially for MS methods that produce mass spectra for a wide range of m/z with lower resolution such as MALDI MS. With this resolution, ions of specific mass on the mass spectrum, give a gaussian-like shape called a peak. For samples with high complexity (many different molecules in the mixture) such as biopsy samples, those peaks can overlap making peak detection even more difficult. Moreover, molecules of the same mass can generate multiple peaks called isoforms as the result of receiving different charges in the ionization step of mass spectrometry. Finally, such signals are disturbed by many different sources of noise both low and high frequency.

Numerous techniques have been used for dimensionality reduction of MS data. In [6] genetic algorithm is used to iteratively select a small set of m/z values (channels) which using cluster analysis gives the best split of the data into two classes (cancer vs noncancer). A similar approach described in [7] used Artificial Neural Network with backpropagation to classify human tumors. Another heuristic method used for biological MS data feature selection is the Simulated Annealing Algorithm [8]. Exact approach commonly used for dimensionality reduction is Principal Component Analysis (PCA). In [9] PCA is used for dimensionality reduction of MS data, then clustered using Linear Discriminant Analysis. Yihui Liu in [10] uses discrete wavelet transform detail coefficients to extract features from the mass spectrum and then uses support vector machine (SVM) as a classifier. More recently [11, 12] used clustering with k-means based Divik algorithm. All of the above and other dimensionality reduction methods, including most filter, wrapper, and embedded methods of the classical knowledge discovery process, have been used to process and classify MS data. Nevertheless, none of these methods makes full use of the peak detection combined with spatial information that MSI provides.

7.3. Data

The data used for this paper comes from patients with head and neck cancer. A total of 4 samples have been used for acquiring the data. The samples were delineated by an expert pathologist with regions of epithelium and caner (Fig. 1). Each sample has a different size with a total number of over 150 thousand mass spectra. Each mass spectrum consists of over 100.000 m/z channels ranging from 800 DA to 4000 DA.

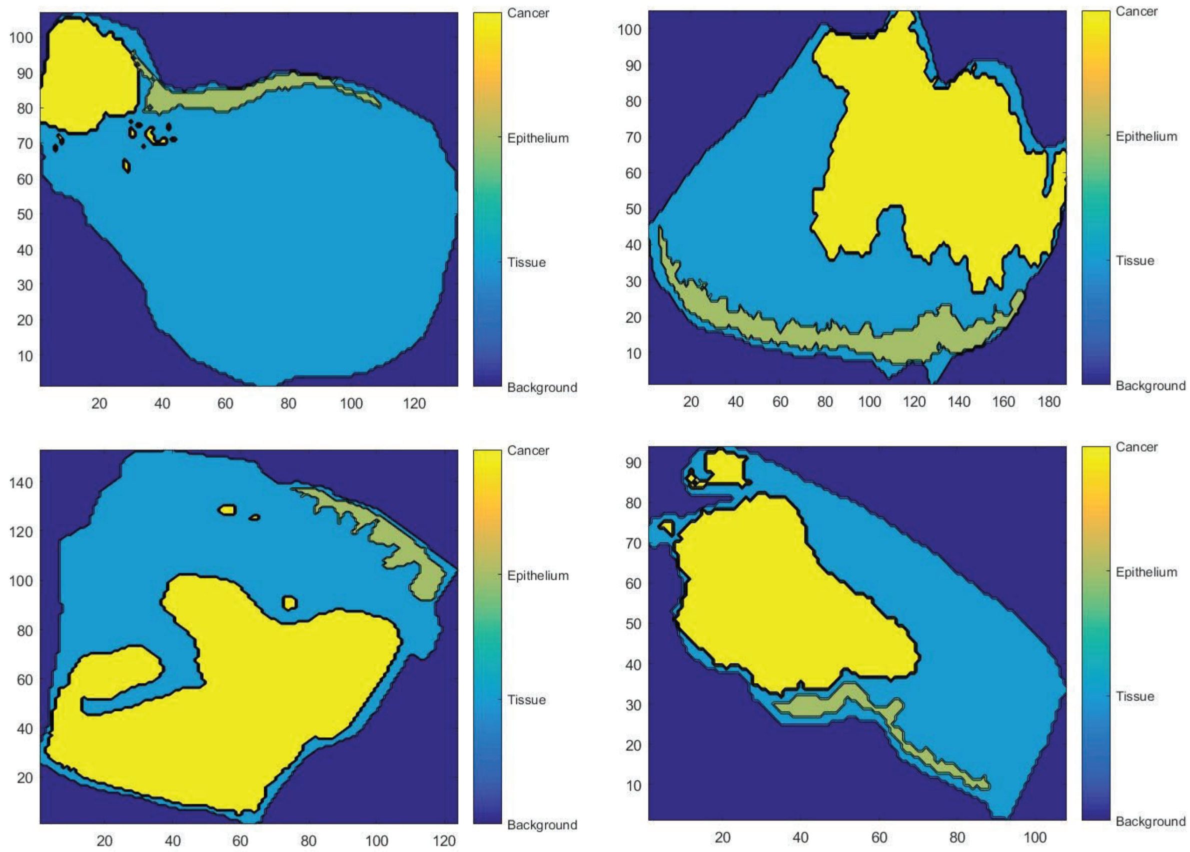


Fig. 1. Biopsy samples with marked epithelium and caner regions
Rys. 1. Próbki biopsji z zaznaczonymi obszarami nabłonka i nowotworu

7.4. Methods

The amount of computations necessary to compare the spatial distribution of two peaks on the sample and the need for a large number of such tests makes it impractical to apply the proposed dimensionality reduction process for each mass spectrum individually. Instead proposed solution is to combine all available data into a single signal and then try to identify peaks correlated with biomarkers on that representation of the data. To do that every value for each channel (m/z value) is aggregated.

7.4.1. Aggregation of mass spectra

The type of the aggregate has an impact on the remaining steps of the feature extraction and dimensionality reduction process. Using average, for example, makes the signal smooth with clear points of division between peaks, however, it can hide meaningful information about potentially very important peaks that have high intensity only in a small number of mass spectra. (Fig. 2a). This problem is solved by choosing maximum as an aggregate, however, in this case, the signal is much rougher and further analysis is harder (Fig. 2b). Therefore intermediate solutions were tired, using the 95th percentile as an aggregate the trade-off between smoothness and information loss is more balanced. An interesting option is also combining the final results derived from each representation. The differences between the representation of all the data and an exemplary single mass spectrum can be clearly seen in the Fig. 2.

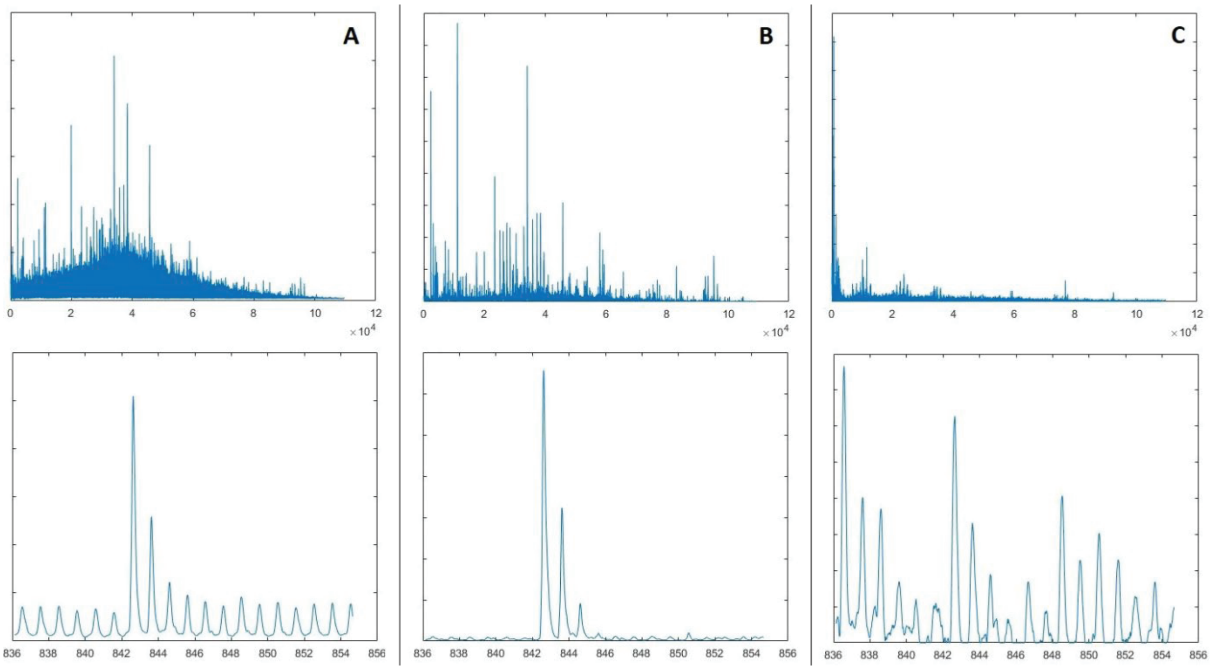


Fig. 2. Visualization of the entire signal (top) and small fragment (bottom) in average representation (A), maximum representation (B), and randomly selected single spectrum (C)

Rys. 2. Wizualizacja całego sygnału (góra) oraz jego niewielkiego fragmentu (dół) dla średniej (A), maksimum (B) oraz losowo wybranego pojedynczego spektrum (C)

Since molecules in MALDI MS are represented by gaussian-like shapes in the mass spectrum or a representation of all mass spectra, the next step of the proposed workflow is to represent the signal as a set of normal distributions around specific m/z values instead of as a list of intensities. This is done by dividing the signal into small parts and fitting them with a gaussian mixture model (GMM).

7.4.2. Signal division

Division of the signal is done by a simple local minima search. For the purpose of fitting GMM into parts, optimal points of division of the signal should be local minima that are relatively close to zero. Parts should also contain a small number of potential peaks ideally one peak or one set of overlapping peaks. Changing resolution of the mass spectra with the increase of m/z value can be problematic. In the lower m/z values, the difference between neighboring channel values is equal to 0.018 DA and it increases gradually to 0.041 DA at the end of the spectrum. In the higher values of m/z, the signal has also much more noise. The optimal process of the division was made by choosing the width of the moving window of the local minima search to be 120 channels. In order to find optimal points of division minima search was done on the signal first transformed with a gaussian filter of small width. This transform removed irregularities in the signal on the scale of a few neighboring channels leaving the shape of peaks unchanged. An additional restriction on the maximal value for the valley in minima search (restriction based on the local neighborhood) ensured that sets of overlapping peaks are not separated into additional fragments but considered together.

7.4.3. Modeling peaks

After splitting the signal into fragments with few potential peaks each, the parts are fitted with the Gaussian Mixture Model using the Expectation-Maximization algorithm in which a randomly selected set of starting conditions iteratively approaches to local optimum. In this case, the starting condition is a randomly selected GMM and the value being optimized is the log-likelihood of the model. Iterations of the EM algorithm are repeated until the shift between iterations, understood as the difference in log-likelihood of iterations is smaller than ε represented as

$$\varepsilon = \left| \frac{L_{i-1} - L_i}{10^3} \right| \tag{1}$$

where:
$\mathcal{E}$ – threshold value for meaningful improvement in EM iteration,
L_i – log-likelihood of iteration i.

Because the EM algorithm is nondeterministic and highly dependent on starting conditions, the algorithm is run multiple times and the run with the best likelihood is chosen.

This process is then repeated with an increased number of elements of the initial GMM. The optimal number of normal distributions in the mixture is chosen by inspecting the gradient of BIC defined as

$$BIC \ = \log(n) * (3k - 1) - 2L \tag{2}$$

where:
BIC – Bayesian information criterion,
N – number of data points,
k – number of elements in GMM,
L – log-likelihood.

BIC is calculated each time and a decision based on the rate of the decrease of this value is made on whether to stop or add another element. Figure 3 shows some randomly selected parts with GMM fitted into them with this method.

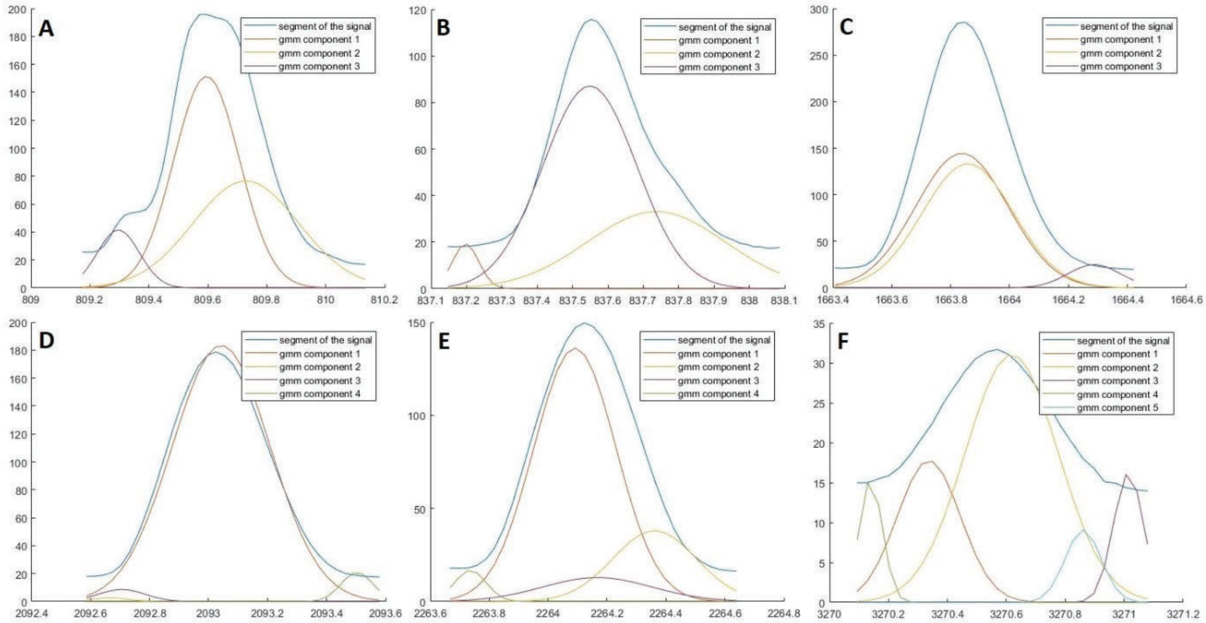


Fig. 3. Randomly chosen signal segments with fitted GMM
Rys. 3. Losowo wybrane segmenty sygnału z dopasowanym GMM

7.4.4. Spatial distribution based peak detection

At this point, all of the data is described by a set of normal distributions, identified by their position, sigma, and lambda values. Those distributions are potential features but many of them describe the same peak. Components 1 and 2 of GMM (Fig. 3c) are completely overlapping and therefore cannot describe different peaks. In Fig. 3b components 2 and 3 seem to be possibly part of a single peak. In Fig. 3a however, components 2 and 3, potentially describe two different, overlapping biomarkers with similar m/z values. In order to decide whether components are describing the same or different peaks, spatial distributions on samples are examined. In [5] Peacock describes a method for comparing spatial distribution by the extension of the Kolmogorov-Smirnov test into two dimensions. If two components within a segment are found to have the same spatial distribution on at least one of the samples, the component with a smaller area is removed from the set, and the area of the larger component is calculated as a sum of the two.

A statistical test used to decide whether two spatial distributions are the same is as mentioned the Peacock's test. However, for this test, no significance levels can be established [5], and therefore there is a need to define a threshold. In order to find the threshold value, more than 60.000 pairs of components were compared. The distribution of those values was fitted with GMM, and the threshold value for similarity of distributions was found (Fig. 4).

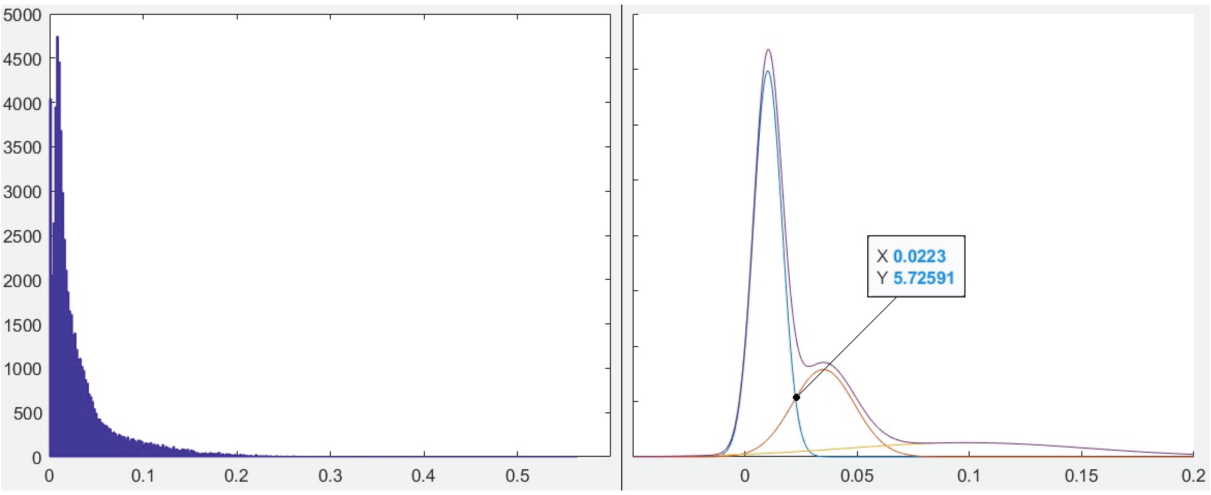


Fig. 4. Distribution of Peacock's test statistic and established threshold value for spatial distribution similarity

Rys. 4. Rozkład statystyki Peacock'a oraz wyznaczona wartość progowa dla podobieństwa rozkładu przestrzennego

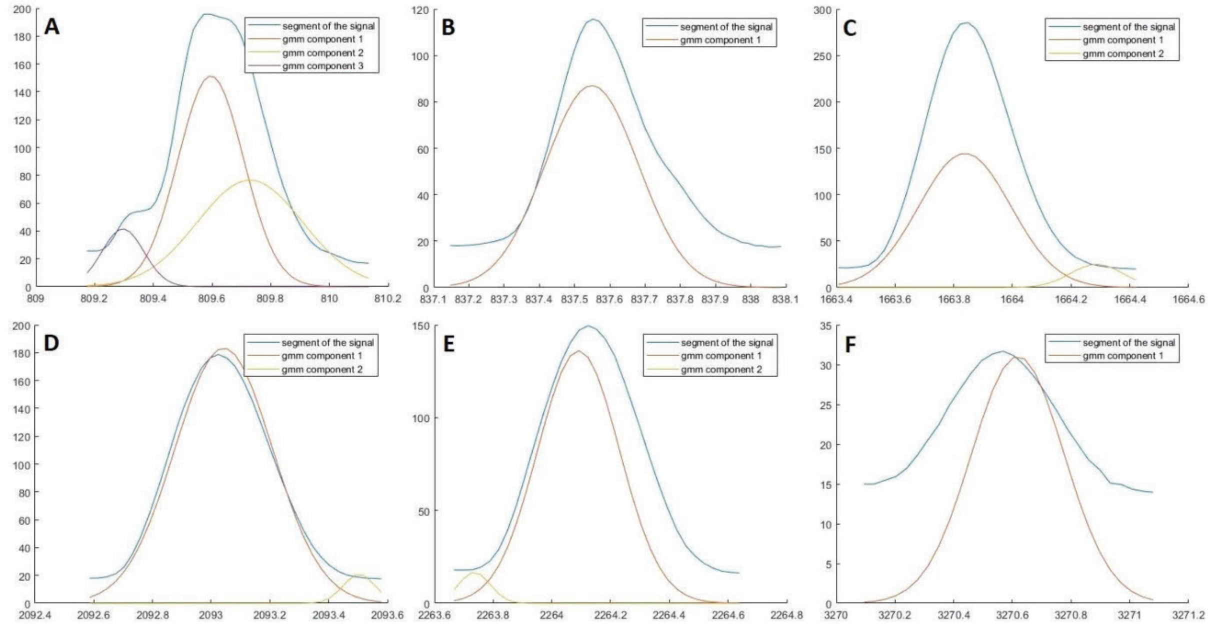


Fig. 5. Signal segments showed earlier with remaining GMM components after spatial distribution comparison
Rys. 5. Wcześniej pokazane segmenty z elementami GMM pozostałymi po porównaniu dystrybucji przestrzennej

Figure 5 shows the results of merging components with a similar spatial distribution. As predicted, components 1 and 2 from Fig. 3c were merged, as seen in Fig. 5c whereas GMM shown in Fig. 3a remains unchanged.

After comparing differences in the spatial distribution in segments shown earlier, we observe a clear reduction in their number (Fig. 5). Only Fig. 5a shows no difference in the number of GMM components. Components 1 and 2 from Fig. 3c were merged as predicted and Fig. 5f shows the resistance of this method to high-frequency noise characteristic of the high m/z part of the spectra. After merging, components associated with noise are also removed based on the signal-to-noise ratio. Finally, one more step is needed to complete the dimensionality reduction process. This step is the identification of isoforms.

7.4.5. Spatial distribution based isotope identification

Typically MALDI ionization results in singly ionized molecules, this, however, is not a rule, and therefore in the mass spectrum, molecules of the same mass can often be observed as a series of consecutive peaks. A series of peaks associated with isoforms of a single molecule is characterized by a 1DA distance between peaks and a similar shape (sigma in the context of normal distributions). The last step of the pipeline is aimed at removing such isoforms and representing them in the final set of features as one feature, described by the localization of the main peak and the combined volume of all peaks.

In order to detect such isoforms, once again spatial distribution on the samples is considered. As it was mentioned, by the nature of isoforms, peaks of such series should be spaced evenly, with 1 DA of the distance between them on the mass spectrum, and should have similar sigma values. Therefore, when considering a peak as a potential start of isoform series, firstly the distance between peeks is considered, and then their shape. Once again thresholds for those values were calculated based on their distribution. A round-robin test with every peak within the 1 DA range through the entire signal would be very time-consuming so the narrowing of potential candidates for isoforms is key.

The search for isoforms starts from the first feature (with the smallest location value) and continues through the entire set of features. Firstly the condition of 1 DA distance is checked, once this condition is fulfilled the similarity of sigma values is checked. When the decision is made that currently considered features could be isoforms, based on those two conditions, the spatial distribution on the samples is evaluated. Figure 6 shows an example, where components 5 and 6 were preliminarily selected as possible isotopes by fulfilling distance and shape conditions but the spatial distribution shows statistically significant differences. Peacock’s test statistic value is above the calculated threshold and therefore those features remain as separate features in the final set. Figure 7 shows an example of features for which Peacock’s test statistic value was below the threshold, and were therefore considered as isoforms of the same molecule and merged.

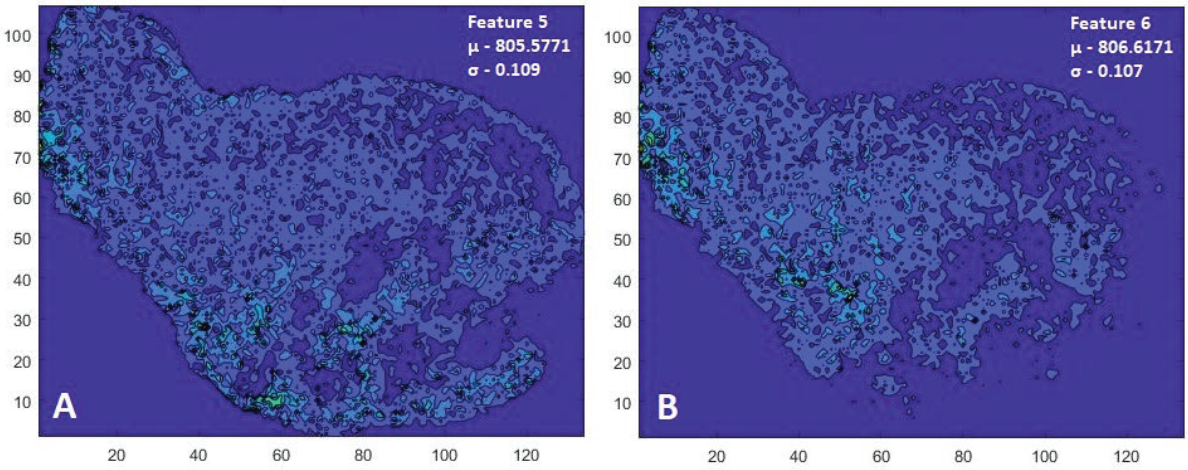


Fig. 6. Visualization of negative outcome of spatial distribution comparison between two features
Rys. 6. Wizualizacja negatywnego wyniku porównania dystrybucji przestrzennej pomiędzy dwoma cechami

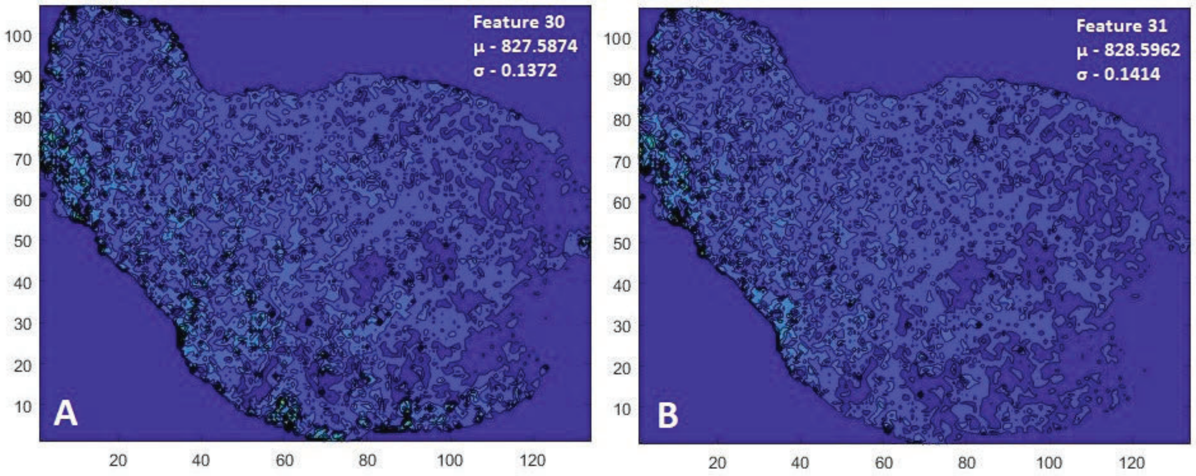


Fig. 7. Visualization of a positive outcome of spatial distribution comparison between two features
Rys. 7. Wizualizacja pozytywnego wyniku porównania dystrybucji przestrzennej pomiędzy dwoma cechami

One molecule can be represented by many isoforms, each separated by 1DA distance on the mass spectrum, so the search continues after the first match until no more matches are found.

7.5. Results and discussion

Computations of the workflo described in this paper,took around 48h to compute on, a computer with 32 GB RAM and i7-11700K 3,6G Hz processor. Considering the amount of the data and the large number of steps in the process, this time is reasonable for this proof of concept and once the features are defined, calculating values for the new data is a matter of seconds. The time needed for the completion of the full pipeline can, and will be optimized. Furthermore, with the usage of parallel computing, this value can decrease drastically, especially for the comparison of spatial distributions which is the most time-consuming step of the workflow.

Table 1

Number of elements in the dataset

Type of aggregation	Number of segments	Number of GMM components	Number of GMM components after merging	Number of GMM components after denoising	The final set of features
Average	3147	10811	4655	2684	1772
Maximum	2414	7939	3485	2060	1405
95th percentile	3131	10806	4662	2596	1629

Commenting on the results of feature extraction and dimensionality reduction, the 14 GB of data, describing tens of thousands of records, with tens of thousands of data points, were reduced to a manageable set of around two thousand features that take no more than 300 MB of memory. Table 1 describes how the number of dimensions of the data changed with each step. In this pipeline, the feature modeling is done using real-life properties of the investigated data, also feature selection and dimensionality reduction steps are based on the physical properties of the data. Thanks to this, we can expect that the final dataset consists of meaningful features, correlated with real properties of investigated data and that this set of features can be a good entry point for further analysis withfeature selection and machine learning.

Bibliography

1. G. Glish, R. Vachet: The basics of mass spectrometry in the twenty-first century, *Nature Reviews Drug Discovery* (2003) **2**:140–150.
2. K. Chughati, R.M. Heeren: Mass spectrometric imaging for biomedical tissue analysis (2010) **110**:3237–3277.
3. A.R. Buchberger, K. DeLaney, J. Johnson, L. Li: Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights, *Anal Chem.* (2018) **90(1)**:240–265.
4. R. Cramer, MALDI MS, *Methods in Molecular Biology* (2009) **564**:85–103.
5. J.A. Peacock: Two-dimensional goodness-of-fit testing in astronomy, *Monthly Notices of the Royal Astronomical Society* (1983) **202**:615–627.

6. E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta: Use of proteomic patterns in serum to identify ovarian cancer, *Lancet.* (2002) **359(9306)**:572–577.
7. G. Ball, S. Mian, F. Holding, R.O. Allibone, J. Lowe, S. Ali, G. Li, S. McCardle, I.O. Ellis, C. Creaser, R.C. Rees: An integrated approach utilizing artificial neural network and SELDI mass spectrometry for the classification of human tumors and rapid identification of potential biomarkers, *Bioinformatics* (2002) **18(3)**:395–404.
8. Y. Li, Y. Liu: A wrapper feature selection method based on simulated annealing algorithm for prostate protein mass spectrometry data, *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, (2008) 195–200.
9. R.H. Lilien, H. Farid, B.R. Donald: Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum, *Journal of Computational Biology* (2003) **10(6)**:925–946.
10. Y. Liu: Feature extraction and dimensionality reduction for mass spectrometry data. *Computers in biology and medicine* (2009) **39(9)**:818–823.
11. G. Mrukwa, J. Polańska, DiviK: Divisive intelligent K-means for hands-free unsupervised clustering in big biological data, *arXiv preprint* (2020).
12. K. Bednarzak, M. Gawin, M. Chekan, A. Kurczyk, G. Mrukwa, M. Pietrowska, J. Polańska, P. Widłak: Discrimination of normal oral mucosa from oral cancer by mass spectrometry imaging of proteins and lipids, *Journal of Molecular Histology* (2019) **50(1)**:1–10.

MASS CHANNEL SPATIAL DISTRIBUTION AS A TOOL FOR PEAK DETECTION AND ISOTOPE IDENTIFICATION IN MALDI TOF MSI DATA

Abstract

Mass spectrometry imaging provides information about biomarkers and their spatial distribution on the samples taken from patients. Analyzing this information can lead to advances in drug discovery and help with diagnosis. The rapid increase in the volume of the biological data gathered by mass spectrometry imaging made it necessary to develop new approaches for analysis. In this paper, we present a data-driven, hands-

-free approach for feature extraction and dimensionality reduction of MALDI ToF MSI data that can be used as the first step of the full machine learning analysis. In our approach, the feature modeling is done using the real-life properties of the investigated data. Also feature selection and dimensionality reduction are based on the physical properties of the data. Thanks to this we can expect that the final dataset consists of meaningful features correlated with real properties of investigated data. The result of the proposed pipeline is a nonredundant set of features small enough for further analysis for example with feature selection and machine learning.

Keywords: MALDI ToF MSI, feature extraction, peak modeling, dimensionality reduction, spatial distribution

DOI:10.34918/85108

Hanna LANGER-MACIOŁ[1,*], Alicja STAŚCZAK[1,*], Karolina WIDZISZ[1],
Wiktoria ŚLIWIŃSKA[1], Kinga LUCIŃSKA[1], Przemysław WENCEL[1],
Barbara STRÓZIK[1], Mariusz FRĄCKIEWICZ[1], Piotr SKUPIN[1],
Dariusz CHOIŃSKI[1], Sebastian STUDENT[1]

Chapter 8. AUTOMATIC ALGORITHM TO MEASURE PHB ACCUMULATION AND SIMILAR STRUCTURES

8.1. Introduction

Along with the rapid technological progress, social awareness and demand for innovative materials, there is a growing need for environmentally friendly alternative to conventional plastics. One of the promising biopolymers is PHA (polyhydroxyalkanoate) and its derivatives, which depending on the hydroxy acid length, can create materials with extremely different properties, from hard and brittle to elastic and sticky. The main advantage is the ability to regulate their features, such as their resistance to mechanical factors or degradation rate, by introducing modifications in the monomer ratio or by combining them with other materials [1]. An additional superiority in ecology terms is the possibility of using waste materials as a substrate. As a result, modern biopolymers are more often used in numerous industry and medicine sectors. The high biocompatibility of PHB (polyhydroxybutyrate) contributes to the great interest in its use for the controlled drug release into the patient's body. The most important aspect in this matter is that the foreign body insertion made of the above-mentioned biopolymer does not cause strong defense reactions in the organism, and its response is mild [2]. Currently, PHA derivatives are also used in the treatment of bone defects and cartilage tissues, cardiovascular engineering [1], as well as in dressing material production and fine powder as a lubricant for surgical gloves [2]. The biocompatible surfactant allowed to

---

[1] Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland.
[*] Corresponding author: alicjastasczak@gmail.com, langerhania@gmail.com.

use the hydrophobic PHB fiber to produce a material that can replace cotton swabs during surgical procedures. This invention eliminates the risk of complications resulting from leaving a fiber fragment inside a patient, as it degrades without accompanying reactions that could be dangerous for the human body. The list of biomaterial applications is growing every year, leading to the emergence of new solutions in terms of production and processing. The process of obtaining biopolymers as a microbial metabolism product also involves quantitative control and microscopic observations to confirm their efficiency. Due to the lack of a tool that may greatly facilitate and control the PHB production process, there is a need to create an algorithm that provides information that allows for the assessment of the material production efficiency.

8.2. Materials and methods

8.2.1. Culture of microorganisms

The culture of *Pseudomonas fluorescens*, capable of biopolymer accumulation under certain conditions, was maintained in two bioreactors operated respectively in a batch and continuous modes. The research was aimed at optimizing the PHB production process, which is one of the intracellular metabolites. To increase the fatty acid conversion into a PHA derivate, it was necessary to maintain stressful conditions for microorganisms and this could be achieved by preventing the microorganisms from gaining access to a carbon source. In the first bioreactor, a fed-batch culture was carried out, involving cyclic metabolite exchange and nutrient additions. For this purpose, each day an amount of reactor medium was decanted and replaced with a freshwater containing the nutrient. The supernatant obtained was taken for further analysis. Homogenous concentrations of oxygen and nutrient were maintained in the entire vessel by using mechanical stirrer and aeration system. Continuous culture experiments were performed in a 5-liter Sartorius Biostat A + bioreactor, providing optimal conditions for *Pseudomonas fluorescens*. The optimal temperature 30°C was controlled by using a heating blanket and deviations from the optimal pH = 7 were adjusted by adding a buffer solution ($NaHCO_3$) to the vessel. Continuous control of breeding parameters was possible due to computer monitoring equipment, which included level sensors, dissolved oxygen, pH, and temperature sensors.

In both bioreactors, the stressful conditions necessary for the production and subsequent accumulation of PHB were maintained by using a medium consisting of peptone and olive oil in varying proportions. The peptone was the main carbon source, while the fatty acid esters contained in the oil were crucial in terms of cell stress, and their presence was evidenced by the increase in lipase activity. Periodic measurements of ammoniacal nitrogen, chemical oxygen demand, lipase activity, and biomass concentration were performed to determine favorable conditions for the target process. The culture condition was controlled on the basis of the concentration of ammoniacal nitrogen. A progressive extinction of microorganisms was manifested by its low value and low nutrient degradation. The degree of nutrient consumption was determined based on lipase activity, as this enzyme is responsible for fat digestion. Measurements on a spectrophotometer were made on a Secoman Uvi Light XT 5 using a p-Nitrophenol. The procedure assumed recording the values at intervals of 15, 25, and 35 minutes from the start and the final result in the form of a waveform for 410 nm and 290 nm. The high lipase activity observed in connection with the determination of the high COD result was interpreted as the significant amount of organic substances and the presence of residual oil in the reactor.

8.2.2. Acquisition of microscopic images

In order to confirm the PHB accumulation in the activated sludge a series of microscopic observations were made. A standard fixation method in the flame was used to prepare the collected sample, then the sample was stained by using two developed methods. The first of staining method uses Sudan III dye and is aimed at the identification of lipids, triglycerides, and lipoproteins. The fixed plates are placed in a humidity chamber, thoroughly rinsed with dye, and left for 30 minutes. The Sudan III (0.1% dye solution in 90% ethyl alcohol solution) is a fat-soluble, nonpolar substance that interacts hydrophobically with hydrocarbon lipid chains, dyeing them red. The second staining procedure was used to observe the PHB granules that are stained in a light orange color. The staining reaction takes place due to the water-soluble substance Nile Blue A (0.05 g of Nile Blue sulfate, 100 ml of 96% ethanol), which is a fluorescent dye from the oxazine group. The procedure involves incubating the samples for 10 minutes in a Coplin staining jar placed at 50°C. Both procedures were terminated by carefully washing the excess dye from the plates and allowing them to dry. For the continuous control of the PHB accumulation process, samples were prepared each day, followed by observations under fluorescence and confocal microscopes. The FV1000 microscope from the Olympus company was used for PHB observations. Figure 1 is an exemplary image obtained during the observation of a specimen, stained with Nile Blue A.

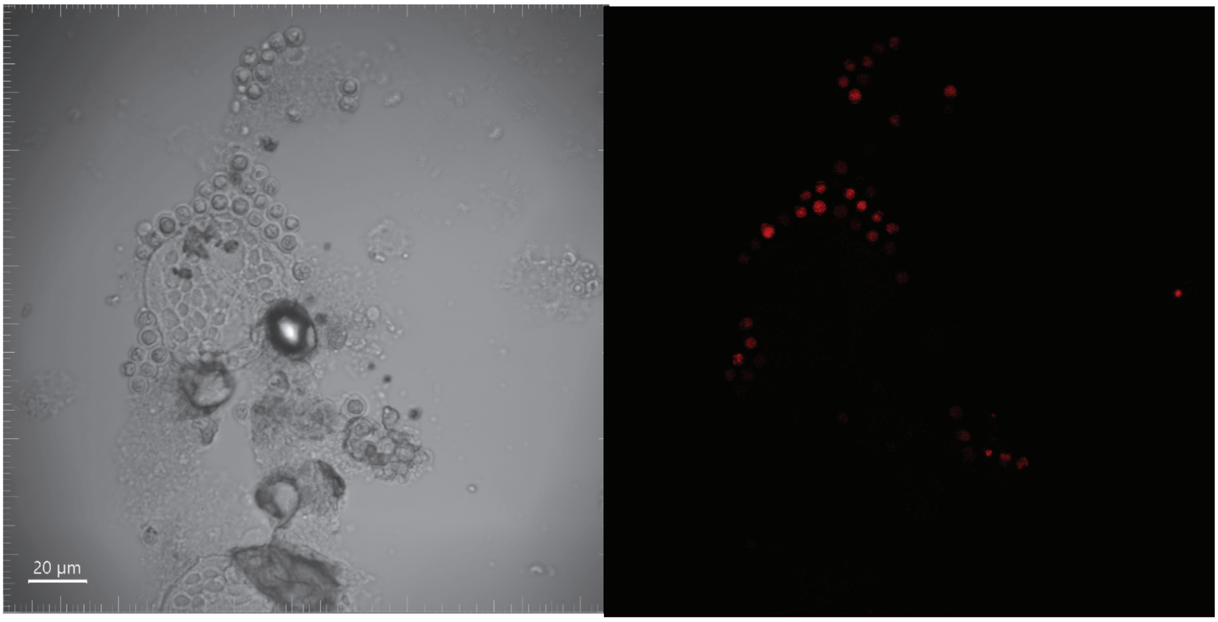


Fig. 1. Comparison of the microscopic images in visible light, and after using a red filter, and excitation with a blue laser. Nile blue A – stained sample was observed using a 60x objective
Rys. 1. Porównanie obrazu próbki mikroskopowej w świetle widzialnym i po zastosowaniu filtra czerwonego oraz wzbudzeniu laserem niebieskim. Próbkę wybarwioną błękitem Nilu A obserwowano przy użyciu obiektywu 60x

8.2.3. Tools and methods used to create the algorithm

The algorithm determining the percentage of PHB in the observed area was written in Matlab programming language, which provides wide packages range for digital image processing. The fact that PHB in the photos is present in the form of the regular objects, similar in shape and color was used. The appropriate RGB channel binarization and morphological operations and filters were applied to identify the background, floc, and granules. To prevent dye misidentification as PHB not rinsed out, the circularity factor was calculated. Too divergent results were treated as misdiagnoses. The main algorithm result is the polymer percentage calculated in the floc and the processed image with identified object's contours marked on it. To visualize the size distribution of the PHB granules, granulometry analysis was performed. Finally, the program returns two charts. The first one is a histogram that shows the area relation of the identified objects, their number with the marked mean, standard deviation, and median. It also presents the surface area distribution. Then, the boxplot displays additional information about outliers. The operation of the algorithm is shown in Fig. 2.

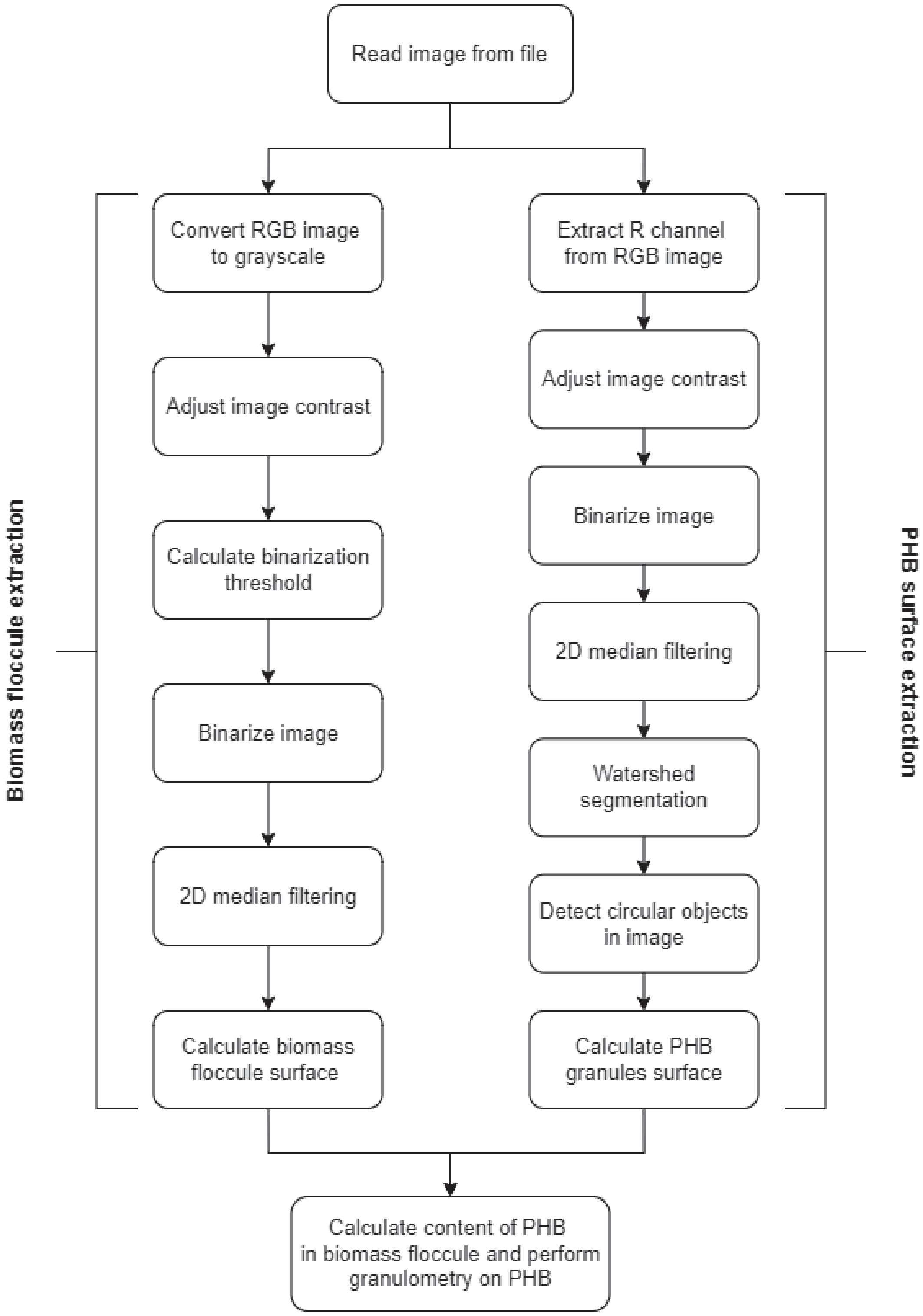


Fig. 2. Operating scheme of the algorithm
Rys. 2. Schemat działania algorytmu

8.3. Results and discussion

The confocal microscopy allows for image generation in three channels, spatial PHB granules observation, and other structures. The proposed algorithm is applied to analyze the microscopic image obtained. In this case, it was used to confirm the presence of PHB granule in the stained samples. The microscopic images show clusters of regular objects with high fluorescence, identified as biomaterial granules. When there are characteristic areas in the image, the algorithm returns a histogram describing the PHB amount over the entire surface, an image with detected, selected objects, and a box plot showing the expected value and standard deviation. In this case, for this image, changing the bins on the histogram will not make the distribution closer to the normal distribution. The program allows for determining the percentage of polymer granules in the identified structure. The algorithm improves the subsequent stages of planning experiments. The result of the microscopic image analysis is shown in Fig. 3. This is a simple way to confirm the accumulation of PHB and estimate its amount in bacterial cells.

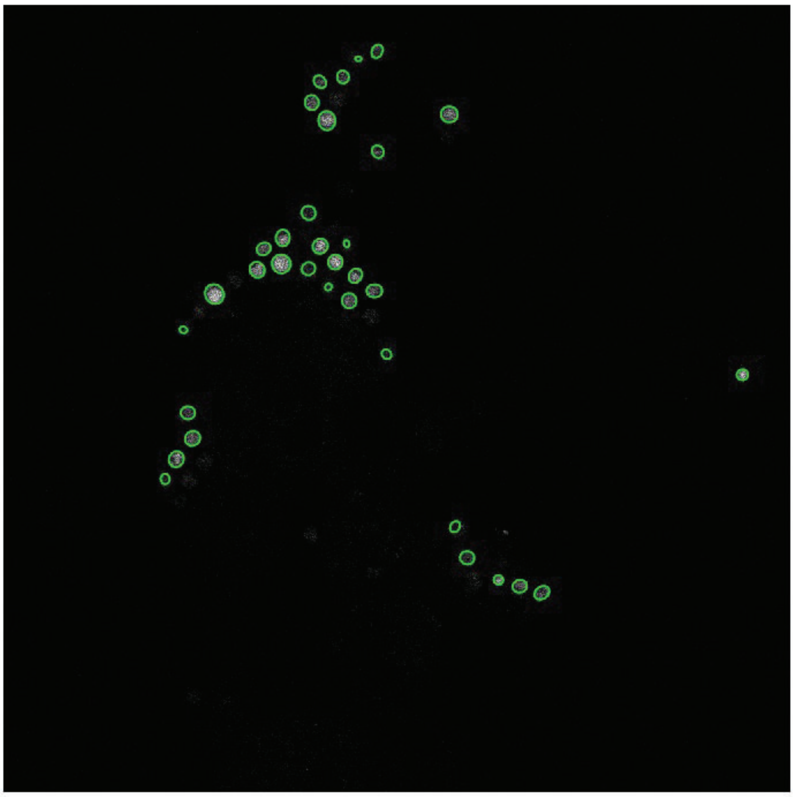


Fig. 3. The microscopic image processed by the proposed algorithm (Fig. 1)
Rys. 3. Rezultat działania algorytmu na obrazie mikroskopowym (Rys. 1)

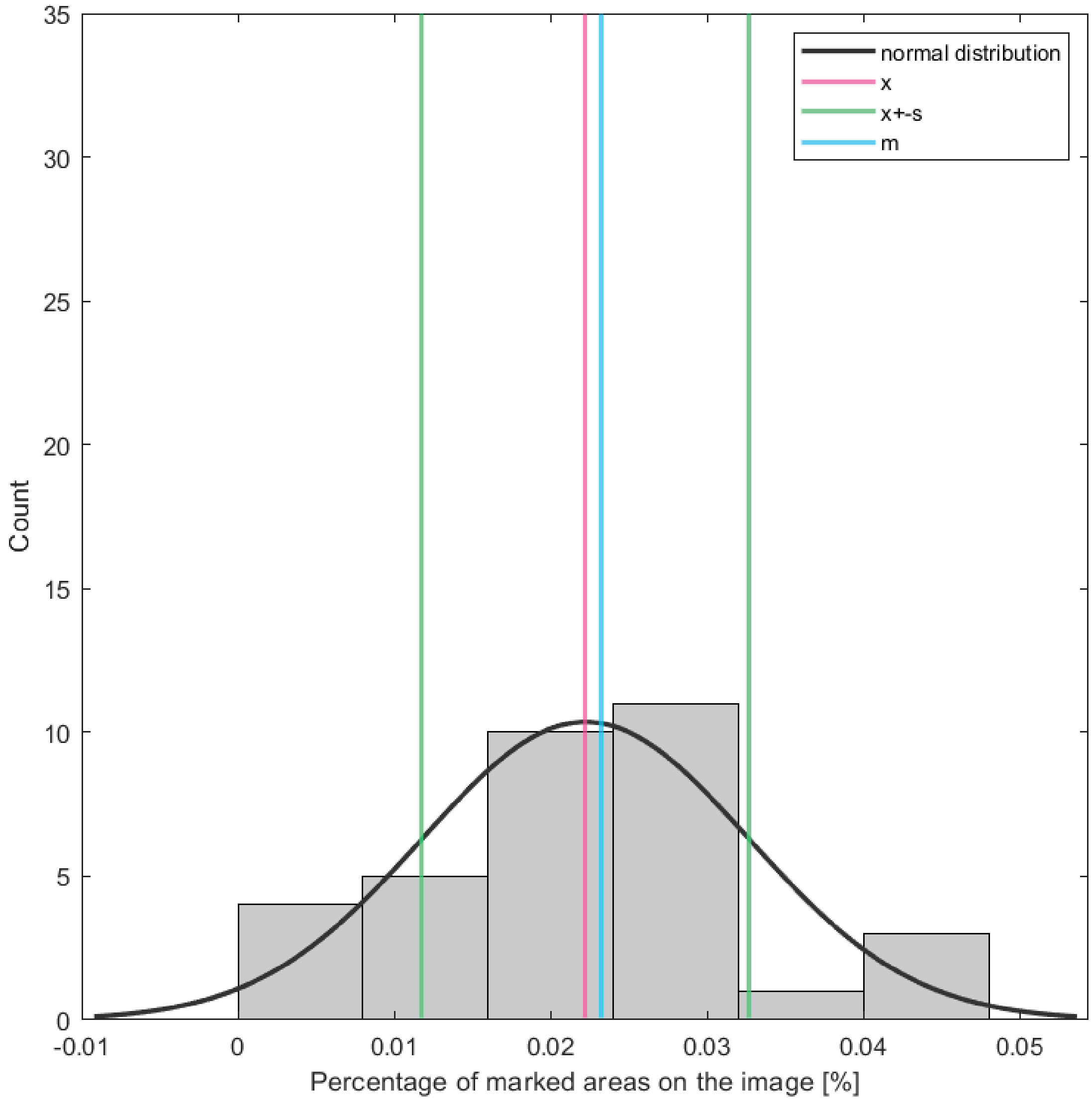


Fig. 4. Histogram showing the percentage of polyhydroxybutyrate in the microscopic image (Fig. 1) and marked with vertical lines: m – median, x – mean, ± s – standard deviation

Rys. 4. Histogram przedstawiający procent polihydroksymaślanu na obrazie mikroskopowym (Rys. 1), zaznaczono liniami pionowymi: m – medianę, x – średnią, ± s – odchylenie standardowe

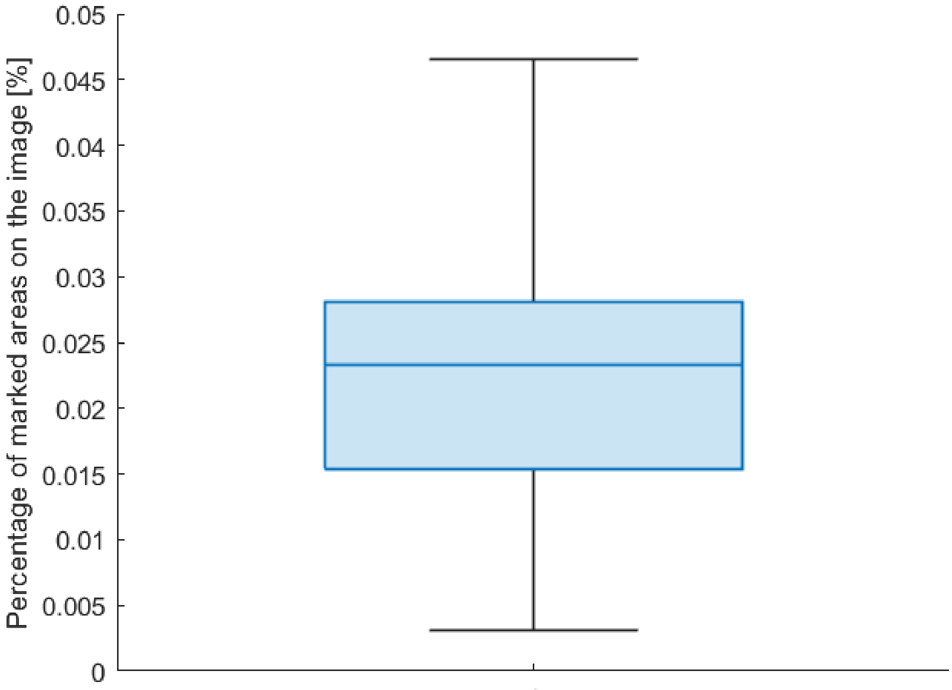


Fig. 5. Box plot showing the percentage of identified polyhydroxybutyrate in the microscopic image (Fig. 1)
Rys. 5. Wykres pudełkowy zawierający procent polihydroksymaślanu na obrazie mikroskopowym (Rys. 1)

The algorithm is universal in a certain way. It can also be used to detect and perform the described operations on objects similar to PHB granules. An example of its application can be the results of an experiment with HELA cells. Carefully prepared HELA tumor cells with fluorescently labeled histone H2B were introduced into the system to assess their cell cycle length and to record monitoring of the movements of individual daughter cells. The research sample was treated with the drug doxorubicin. The experiment was made observations on a fluorescence microscope with automatic image acquisition software. Figure 6 – Fig. 9 present the operations performed on the above-mentioned example to identify cells and their percentage in the photos.

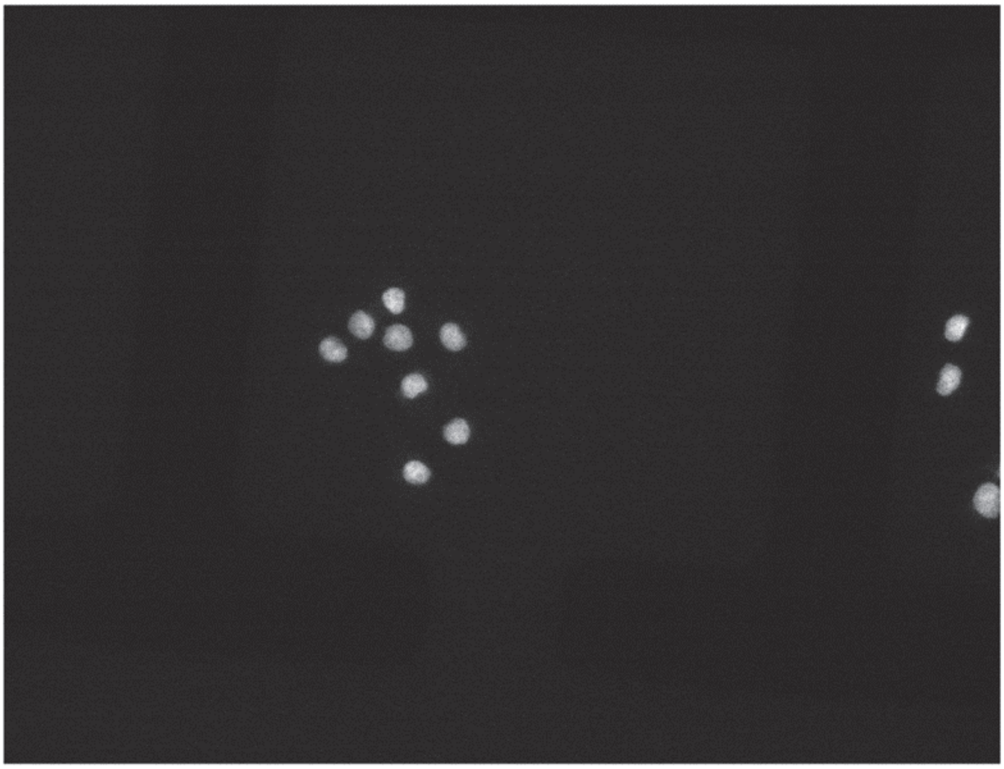


Fig. 6. Sample microscope image with HeLa cells
Rys. 6. Przykładowy obraz mikroskopowy z komórkami HeLa

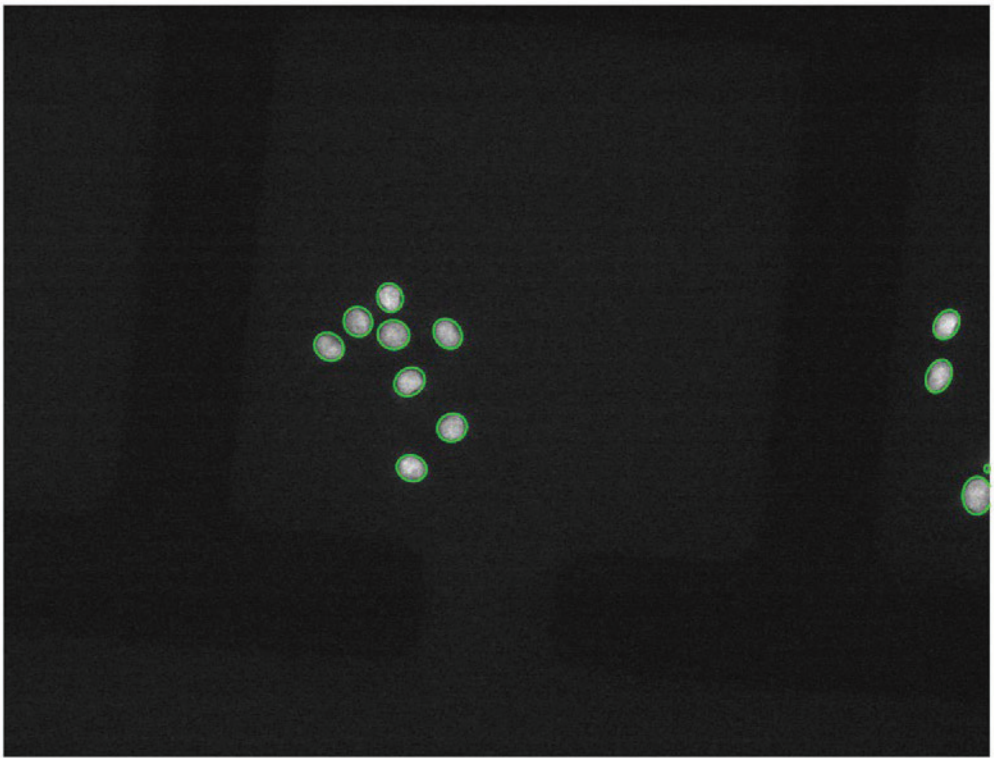


Fig. 7. The processed image after applying the proposed algorithm for the microscopic image (Fig. 6)
Rys. 7. Wynik działania algorytmu na obrazie mikroskopowym (Rys. 6)

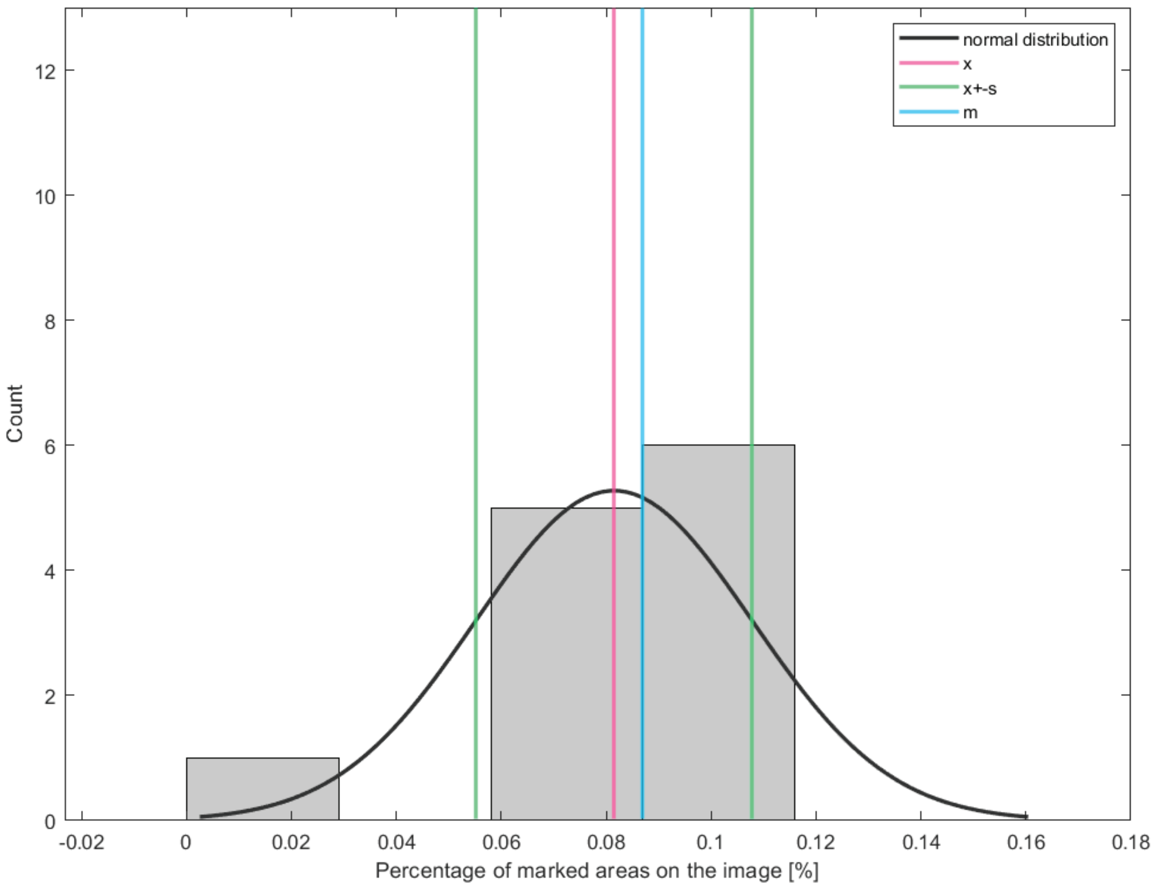


Fig. 8. Histogram showing the percentage of marked areas on the microscopic image (Fig. 6) and marked with vertical lines: m – median, x – mean, ± s – standard deviation

Rys. 8. Histogram przedstawiający procent zaznaczonych obszarów na obrazie mikroskopowym (Rys. 6), zaznaczono liniami pionowymi: m – medianę, x – średnią, ± s – odchylenie standardowe

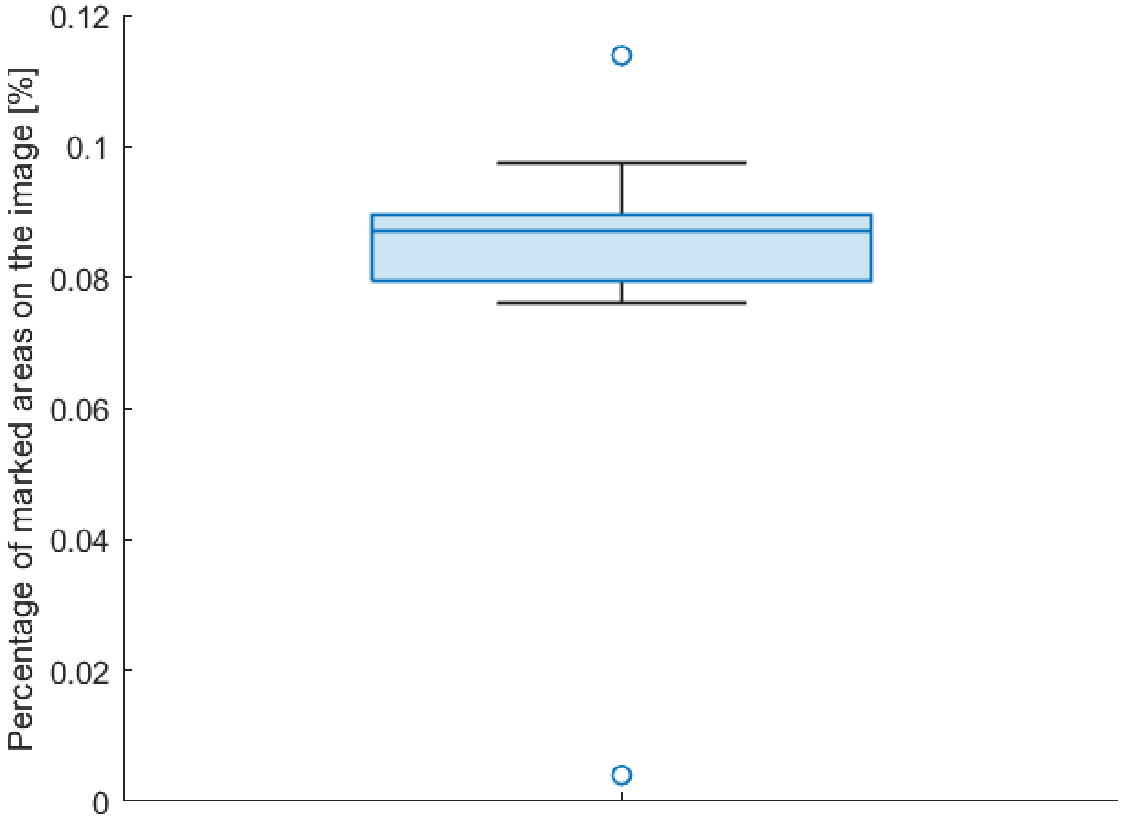


Fig. 9. Box plot showing the percentage of marked areas in the microscopic image (Fig. 6)
Rys. 9. Wykres pudełkowy zawierający procent zaznaczonych obszarów na obrazie mikroskopowym (Rys. 6)

Biomaterials are a promising alternative to plastics used so far, both in application versatility, waste management, and minimizing their generation terms. Some by-products from industrial processes can be used to produce PHA, thus it is possible to reduce problem needed to produce PHA processing costs by up to about 50%. One of the promising alternatives to an expensive substrate are waste substances from the oil industry, which are classified as attractive carbon sources for microorganisms. Another potential suppliers are agriculture and the food industry. By-products or materials rejected during individual processes can be reused in environmentally friendly biopolymer fabrication while reducing their production costs. Potentially suitable for ecological alternative production are banana parts and the peel of numerous fruits, sugar cane fragments discarded in various processing steps, and algae biomass [3]. In the research, olive oil was used for this purpose but leftover fry oil can also be used. Optimizing the PHB production process by reducing the raw materials and energy consumption while maximizing the biosynthesis efficiency requires continuous observations and polymer accumulation control. The proposed algorithm makes it possible to facilitate and automate these tasks, and to estimate the percentage of granules in microscopic images. On this basis, the solutions legitimacy, the microorganism breeding system, or the supplying substrates method was determined. Visualizing the results in graph form showing the object number with size within a given range facilitates work and helps in determining the systematization of the characteristics of the granules. The need to conduct long-term observations generates large amounts of data for analysis. Using a program that performs some of the tasks not only relieves the research team but often enables control activities to be carried out in time, allowing for a quick response to changes observed in the results. In addition, the universality of the algorithm allows it to be used in subsequent studies. The biomaterials topic, and especially PHA and its derivatives, has not been sufficient, and the work continuation may bring huge profits to medicine. Biopolymers have numerous applications in many areas, such as bone and cartilage tissue engineering. One in vivo study on the bone treatment defects has shown that PHB can be applied in implant production and other bone marrow scaffolds. It was a positive tissue adaptive response to the polymer. In addition, a 12-month follow-up showed no chronic inflammatory response. Other applications of PHA derivatives are, for example, devices for the repair of meniscus, articular cartilage, or tendons, as well as bone plates, bone plating systems, orthopedic pins, ligament and tendon grafts, and bone graft substitutes [1]. The antibacterial properties of some biomaterials application in the prophylaxis and postoperative infection treatment have also been shown. For this purpose, a connecting PHB spheres method containing antibiotics with poly(ethylene glycol) (PEG) and titanium, which can be the base material for implants, was developed. Titanium surface with antimicrobial properties in surgical procedures reduces bacteria adhesion and eliminates some of the infection risks [5]. Another department that eagerly uses biopolymer solutions in cardiovascular tissue engineering, deals with a health problem group that is at the forefront in incidence terms. Unfortunately, many cases of cardiovascular disease require surgical treatment, part of which is the patch applied to the patient's body. The material from which the patch is made must meet several requirements, including not being toxic or immunogenic, and also characterized by high durability and flexibility [1]. PHB is a polyester whose thermal and mechanical properties can be compared with conventional plastics [4]. Again, biopolymers can meet the requirements and their application can greatly improve treatment methods and overcome the problems associated with conventional solutions. One of them is the need to repeat the heart valve replacement procedure because if the patient was a child at the first operation time, the implanted implant becomes a mismatch with the organ after some time. By using the PHB to impregnate the valve, an increase in its volume was observed in vivo. In this way, the idea of an implant that adapts to the recipient's body was born, which would exclude the additional risk of postoperative complications. Biopolymers are commonly used mainly in surgery, but not only. In addition to the examples described above, the biopolymers are also useful in the production of sutures and suture fasteners, staples, rivets, screws, surgical meshes, fillers, repair patches, adhesion barriers, guided tissue repair or regeneration devices, nerve guides, pericardial patches, vein valves, ocular cell implants, spinal fusion cages, skin substitutes, and many more [1]. Due to the strength and properties similar to conventional plastics, some biopolyesters can also be a material for film production that is a part of food packaging and the production of disposable cutlery [4]. The wide range of possibilities offered by PHA application derivatives only confirms the belief that biopolymers will be one of the key materials in the field of medicine, health protection and environmental engineering. Thus, optimizing their production process is inevitable. Thanks to the proposed algorithm, it is possible to facilitate the production process and automate some tasks.

Bibliography

1. Q. Wu, Y. Wang, G.Q. Chen: Medical Application of Microbial Biopolyesters Polyhydroxyalkanoates, *Artificial Cells, Blood Substitutes, and Biotechnology* (2009) **37:**1–12.
2. P.A. Holmes: Applications of PHB – a microbially produced biodegradable thermoplastic, *Physics in Technology*, (1985) **16(1):**32–36.
3. R. Sirohi, J. Prakash Pandey, V. Kumar Gaur, E. Gnansounou, R. Sindhu: Critical overview of biomass feedstocks as sustainable substrates for the production of polyhydroxybutyrate (PHB), *Bioresource Technology* Volume 311 (2020), 123536.
4. O. Olejnik, A. Masek, J. Zawadziłło: Processability and Mechanical Properties of Thermoplastic Polylactide/Polyhydroxybutyrate (PLA/PHB) Bioblends, *Materials* Volume 14 (2021), 898.
5. A. Rodriguez-Contreras, M. Soledad Marques-Calvo, F.J. Gil, J.M. Manero, Modification of titanium surfaces by adding antibiotic-loaded PHB spheres and PEG for biomedical applications, *Journal of Materials Science: Materials in Medicine* (2016), 27(8)–124.

AUTOMATIC ALGORITHM TO MEASURE PHB ACCUMULATION AND SIMILAR STRUCTURES

Abstract

Over the years, biomaterials have become more and more popular and find a new applications huge number. This is due to the growing social awareness and the need to implement ecological solutions. Environmentally-friendly production and the possibility of using industrial waste are just some of the biopolymer advantages. A wide property range and modification potential allow them to be used in many industry and medicine sectors. Biomaterials are applied both in surgical tools and accessories production as well as implants permanently inserted into the patient's body. Features such as biocompatibility and high durability eliminate the problems associated with conventional solutions. One of the most widely used biomaterials is PHA derivatives, which can be obtained as a metabolic product of *Pseudomonas fluorescens* bacteria. Process optimization requires the maintenance methods and control procedures necessary for breeding development to determine its effectiveness. One way to confirm the accumulation of PHB is by microscopic observation. An algorithm was developed that automated the identifying polymer granule step. It also improved the process of estimating the PHB amount in an image and allowed for convenient results visualization. The script is also used in other studies where data of this type is analyzed.

Keywords: polyhydroxybutyrate (PHB), polyhydroxyalkanoates (PHA), bioplastic, PHB accumulation, measure PHB accumulation

DOI:10.34918/85109

Bartłomiej PYCIŃSKI[1,*], Karolina KOWOL[1], Jan JUSZCZYK[1], Ewa PIĘTKA[1]

Chapter 9. X-RAY MICROTOMOGRAPHY IMAGING OF FRESH AND FORMALIN-FIXED POULTRY HEART

9.1. Introduction

X-ray microtomography (μCT) is an X-ray technique whose popularity and area of application have been increasing in recent years [1]. This imaging allows obtaining three-dimensional images with a resolution of up to micrometers (thus the technique's name). During the imaging, a set of two-dimensional cone-beam projections is acquired. The final three-dimensional (3D) image is reconstructed from the projections. The μCT technique combines features of both classical X-ray (cone beam, planar projections, image magnification) and computed tomography (3D reconstruction). The main feature that distinguishes this technique is the acquisition geometry (Fig. 1), which involves a rotating table with a sample between the X-ray tube and the flat detector. The closer the table is to the tube, the higher the magnification on the detector and the better the final image resolution. Thus, the highest resolution (on the single micrometers level) is only possible with small samples of dimensions up to few centimeters.

μCT is commonly used in life science studies, especially in tissue engineering. Other areas of interest include botany, biotechnology, or chemistry. Besides, it is employed in biomedical engineering, materials science, electronics, and physics. A survey on the application of μCT in these fields was given in [2, 3].

---

[1] Faculty of Biomedical Engineering, Silesian University of Technology, Zabrze, Poland.

[*] Corresponding author: bartlomiej.pycinski@polsl.pl.

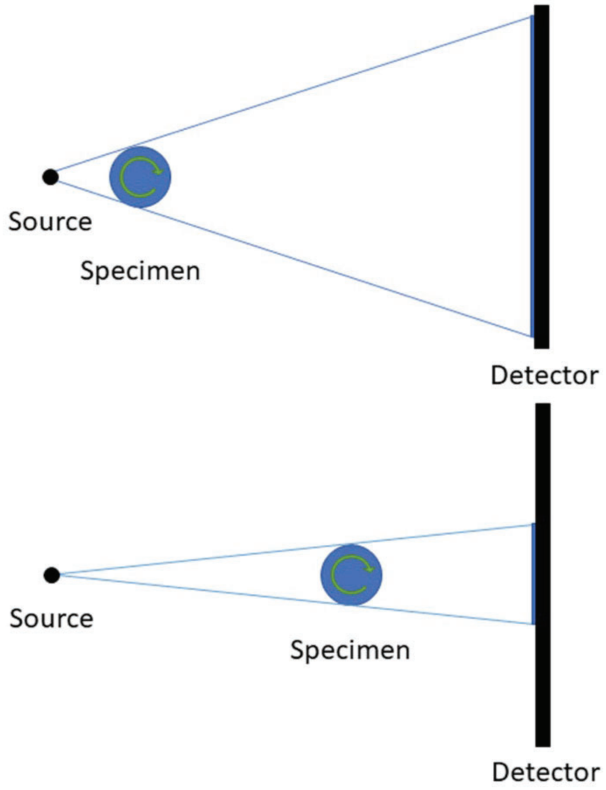


Fig. 1. µCT device components: X-ray source – specimen – detector. The closer the sample is to the source, the greater the magnification of the image
Rys. 1. Elementy aparatu µCT: źródło promieni X – próbka – detektor. Im bliżej źródła jest próbka, tym większe powiększenie obrazu

µCT imaging can also be achieved using a parallel beam of X-rays generated in a synchrotron. Since the X-rays in such machines are monochromatic, the detectors not only measure the intensity of the radiation but also detect the gradient of the phase shift or the transverse Laplacian of the phase [4]. This type of measurement allows for obtaining much higher contrast, especially for soft tissues. [5]. The former µCT technique (based on the X-ray tube) in animal tissue imaging is mainly used for hard tissues (bones, cartilages) or assessment of focal lesions (calcifications, stiff tumors, fibrosis etc). The remaining part of this paper discusses only classical µCT imaging generated by the tube.

It remains an open question whether µCT can – to some extent – replace or complement histopathological imaging. The latter is the gold standard for diagnosing many diseases, but it also has drawbacks: it irreversibly destroys the specimen during preparation and is expensive, and time-consuming. Moreover, few slides are obtained during the preparation, and they are evaluated only in two dimensions. Pathologists have no information about the tissue architecture in 3D space. A solution to this problem could be so-called 3D histopathology, in which consecutive slices are aligned [6]. However, this is still a relatively uncommon technique [7].

During histological preparation, the specimen is processed in numerous consecutive steps [8]. Immediately after being collected, it is fixed in formaline (4% buffered water solution of formaldehyde). The next steps contain immersion in increasing solutions of propanol or ethanol, then in 100% xylene. The final step is embedding in paraffin, which leads to formaline-fixed, paraffin-embedded (FFPE) samples.

This study aims to compare μCT images of fresh and formalin-fixed (FF) poultry heart specimens. These are preliminary measurements that are currently being developed. Other similar studies have focused on imaging FFPE samples alone or comparing FFPE with FF images [8-11] or on involving staining to enhance tissue contrast [12-14].

9.2. Materials and methods

9.2.1. X-ray imaging

All the images were acquired with the ProCon X-Ray CT-Compact plus system (https://procon-x-ray.de), deployed and adjusted by Casp System. The machine is controlled by a dedicated software provided by the manufacturer. The application allows the manipulation of a wide range image acquisition parameters. The parameters selected for this study are given in Table 1, they follow the Standard Operating Procedure of Microcomputed tomography [15]. No pixel binning was applied. The acquired projections were reconstructed using X-AID v. 2020.3.3, the software that the device manufacturer recommends. The FDK algorithm, the ring artifact correction, and the Hamming filter were used for both examinations to reconstruct final 3D images. The resolution of the reconstructed images was 19.2 μm in all three dimensions.

Table 1

Details of X-ray μCT machine

Voltage [kV]	Current [μA]	Power [W]	Exposure time [s]	FOD [mm]	No of projections	Total duration [h:m:s]
40	160	6.4	2	100.15	2880	3:12:29

9.2.2. Intensity range of µCT reconstructions

In the employed system, the obtained intensity values are represented as 16-bit unsigned integers ranging from 0 to 65 535. They do not directly map onto the Hounsfield scale. Therefore we can directly compare only the results obtained with exactly the same acquisition parameters. In this study, this requirement was fulfilled.

To convert intensity values to Hounsfield units (HU), the µCT system has to be calibrated with a calibration phantom [16]. The phantom consists of a container with water and air. After obtaining final images, HU can be determined according to the equation:

$$\mathrm{HU}(x) = 1000 \cdot \frac{\mu_x - \mu_{\mathrm{H_2O}}}{\mu_{\mathrm{H_2O}} - \mu_{Air}} \tag{1}$$

where μ_x, $\mu_{\mathrm{H_2O}}$, μ_{Air} are the linear attenuation coefficients for the examined voxel, the water and air, respectively.

The values of HU are computed for a specific acquisition parameter set and when they changes, the calibration shall be repeated. The detector's response linearity should also be verified.

9.2.3. Poultry tissue

In the experiments, we used a poultry (hen) heart obtained from a meat store. The organ was stored for 24-48 hours at 4-6°C after the preparation of the hen. Two µCT examinations were performed. The first, without additional preparation, in the horizontal position of the specimen. Then, immediately after imaging, the heart was placed in formalin (4% buffered water solution of formaldehyde) at 20°C for 48 hours. After this time, the heart was picked up, rinsed with tap water, and mildly dried. The second imaging was then performed with identical acquisition parameters as before (Fig. 2). The specimen was disposed of after use.

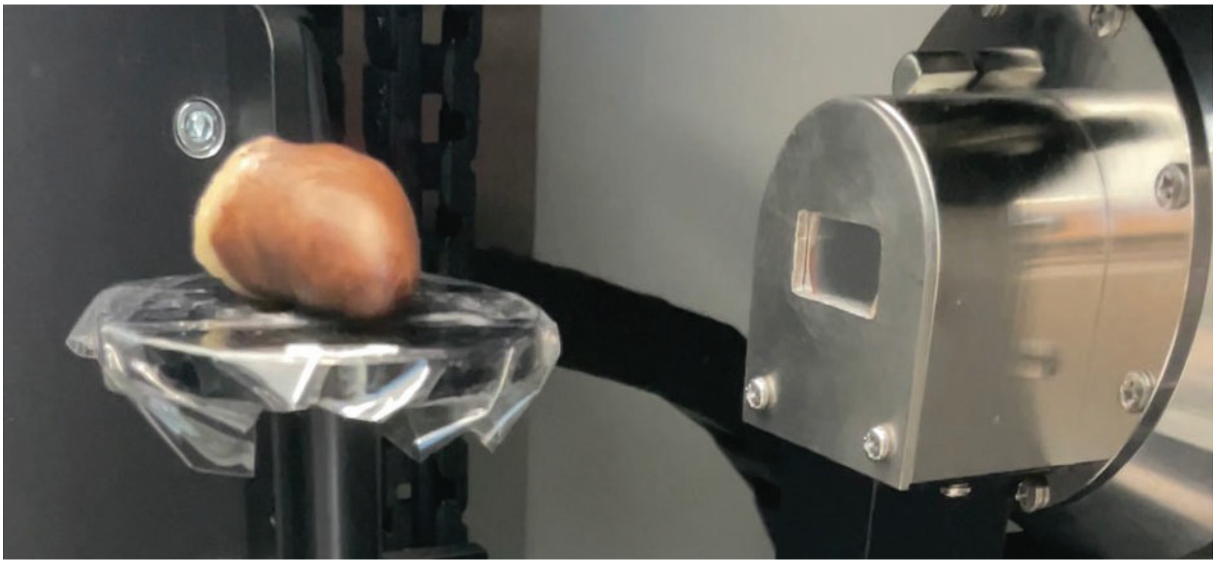


Fig. 2. The FF heart on a table. On the right side, there is an X-ray tube
Rys. 2. Serce stabilizowane formaliną na stoliku. Po prawej stronie przedstawiona jest lampa RTG

9.3. Experiments and results

The reconstructed 3D volumetric images were statistically analyzed in a Python 3.9.6 environment and visualized in 3D Slicer (https://slicer.org) software. The following parameters of the fresh and FF specimen were assessed: the heart volume, its dimensions along main axes (defined as the longest vertical and horizontal axes, and the longest axis perpendicular to the previous two), and average as well as extreme intensities. Additionally, signal-to-noise ratio (SNR) and contrast-to-noise ratio (CNR) were assessed according to the following equations [17]:

$$\mathrm{SNR} = \frac{I}{\sigma} \tag{2}$$

$$\mathrm{CNR} = \frac{|I_1 - I_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}} \tag{3}$$

where I, I_1, and I_2 indicate the mean intensities of homogeneous components within the specimens, and σ, σ_1 and σ_2 are the standard deviations of these components.

In both images the heart was segmented, and its volume was classified as one of two main tissues: muscle or fat. The Otsu segmentation method was employed. Basic descriptive statistics were calculated for the tissues and background. CNR and SNR were calculated for arbitrary chosen homogenic volumes of size 2 x 2 x 2 mm.

9.3.1. Results

The fresh heart's volume was equal to 9.67 cm^3, while the FF specimen was 8.54 cm^3 (decrease of 11.67%). Muscle took 92.94% of the fresh heart and 92.50% of the FF. The remaining part of the heart was classified as fat. The dimensions of the fresh and FF organ are given in Table 2. Descriptive statistics for the image intensities were calculated and summarized in Table 3. The reconstructed voxel intensity values were not converted to the Hounsfield scale, hence the reported values remain dimensionless and their significance is useful for mutual comparison only.

Table 2

Dimensions and volumes of the specimens

	X	**Y**	**Z**	**volume**
Fresh	41.36 mm	16.61 mm	27.65 mm	9.67 cm^3
Formalin fixed	37.65 mm	16.12 mm	26.78 mm	8.54 cm^3
Change	-8.96%	-2.99%	-3.12%	-11.67%

Table 3

Descriptive statistics of the reconstructed volumes

	Fresh heart			**Formalin-fixed heart**		
	muscle	**fat**	**background**	**muscle**	**fat**	**background**
mean	43 396	37 697	25 307	40 049	33 561	19 067
std	1 276	1 145	8 643	1 434	1 494	7 245
min	36 120	31 896	0	31 906	26 448	0
max	63 057	43 893	39 227	64 550	41 457	34 194
median	43 422	37 799	28 084	40 062	33 642	21 588
SNR	1 156	1 084	9	780	505	7

Experiments revealed that all average values of the tissue intensity decreased in the FF specimen. Additionally, the differences between the mean values (i.e., the intensity contrast) increased. These features are presented in Fig. 4. Values of SNR are smaller in the FF specimen. CNR values between muscle and fat equal to 3.32 and 3.13 for fresh and FF specimens, respectively.

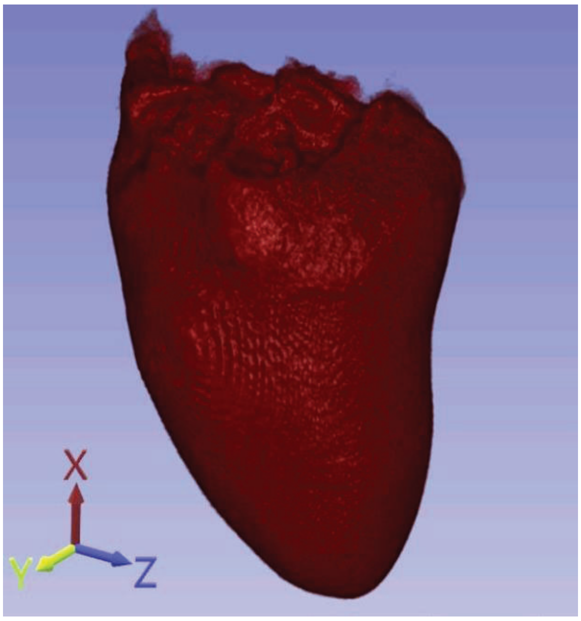


Fig. 3. Volumetric rendering of the fresh heart
Rys. 3. Wizualizacja wolumetryczna świeżego serca

The obtained results indicate that the process of formalin fixation of the poultry heart does not negatively affect the imaging of the tissue. Muscle and fat can be distinguished clearly in both images, as presented in Fig. 4 and Fig. 5. Fresh tissue is more flexible than FF specimen, hence there are motion artifacts in the reconstruction

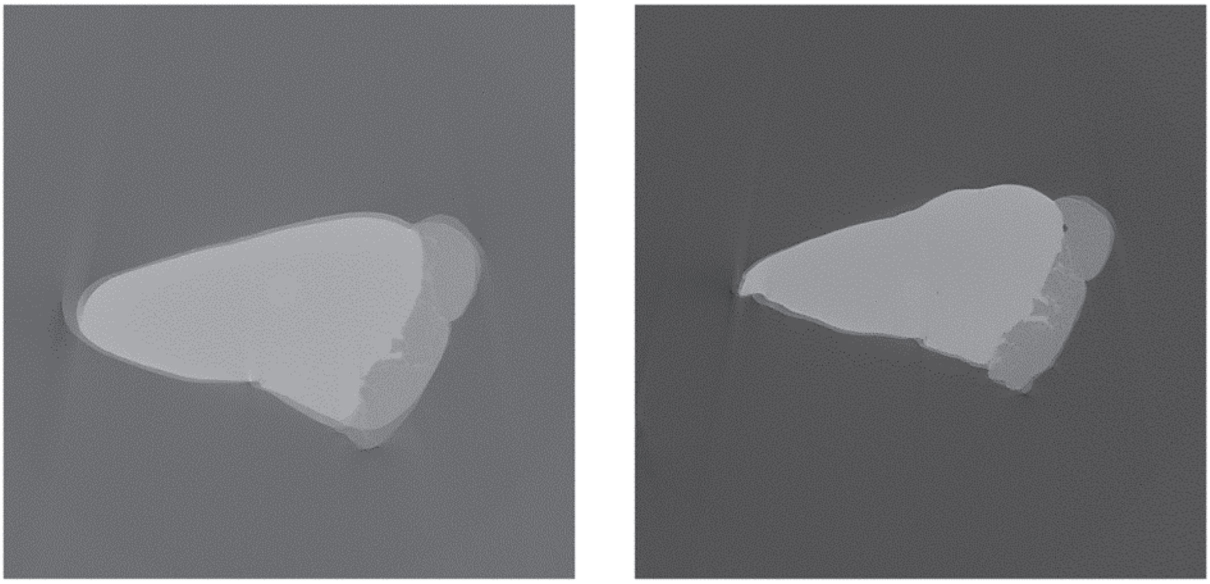


Fig. 4. Images of fresh (left) and FF (right) heart. In the FF heart, higher contrast and smaller volume may be noted. In the fresh image, movement artifacts in the most lower part of the tissue are visible.
Rys. 4. Obrazy serca świeżego (strona lewa) i stabilizowanego w formalinie (strona prawa). W obrazie serca stabilizowanego widoczny jest większy kontrast i mniejsza objętość. W obrazie serca świeżego widoczne są artefakty ruchowe w najniższym fragmencie tkanki.

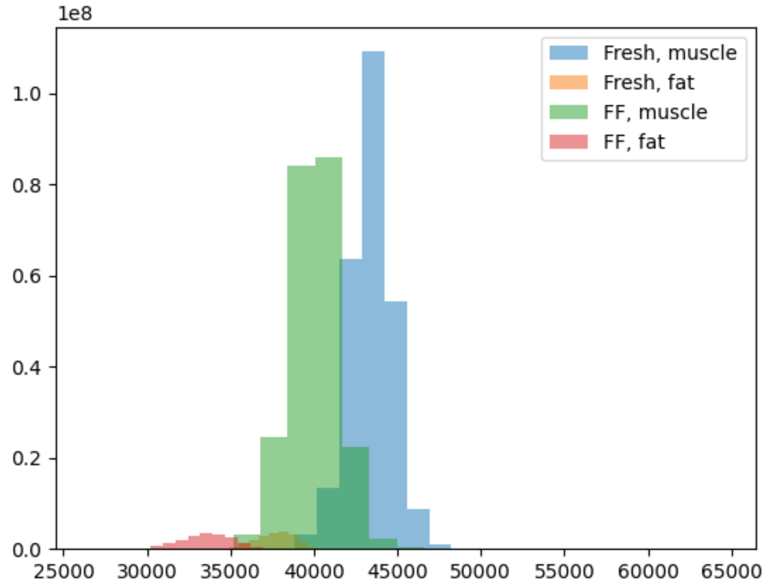


Fig. 5. Histograms of intensities of the fresh and FF heart tissues
Rys. 5. Histogram intensywności wokseli tkanek serca świeżego i stabilizowanego w formalinie (FF)

9.4. Conclusion

In the paper, 3D μCT imaging of a poultry heart was presented. The images were acquired from fresh and FF tissues. Fixation was performed in a standard way in a formaline solution. The results show that stabilization does not decrease the quality of imaging, yet slightly improves the contrast. FF images also feature significantly reduced motion artifacts. The change in the organ volume is small and regular, and does not affect the quality of the images.

Bibliography

1. Lin A.S.P., Stock S.R., Guldberg R.E.: Microcomputed Tomography, [in:] Hawkes P.W., Spence J.C.H. (eds.): *Springer Handbook of Microscopy*. Cham: Springer International Publishing, (2019): 1205–36.
2. Ritman E.L.: Current Status of Developments and Applications of Micro-CT. *Annual Review of Biomedical Engineering* (2011); **13**: 531–52.
3. Orhan K. (ed.): *Micro-Computed Tomography (Micro-CT) in Medicine and Engineering*. Cham: Springer International Publishing, (2020).

4. Larsson D.H., Vågberg W., Yaroshenko A., Yildirim A.Ö., Hertz H.M.: High-resolution short-exposure small-animal laboratory X-ray phase-contrast tomography. *Scientific Reports* (2016); **6**: 39074.
5. Betz O., Wegst U., Weide D. et al.: Imaging applications of synchrotron X-ray phase-contrast microtomography in biological morphology and biomaterials science. I. General aspects of the technique and its advantages in the analysis of millimetre-sized arthropod structure.*Journal of Microscopy* (2007); **227**: 51–71.
6. Pichat J., Iglesias J.E., Yousry T., Ourselin S., Modat M.: A Survey of Methods for 3D Histology Reconstruction. *Medical Image Analysis* (2018); **46**: 73–105.
7. Pyciński B., Yagi Y., Walts A.E., Gertych A.: 3-D Tissue Image Reconstruction from Digitized Serial Histologic Sections to Visualize Small Tumor Nests in Lung Adenocarcinomas. *Information Technology in Biomedicine*. Vol. 1186. Cham: Springer International Publishing, (2020): 55–70.
8. Rodgers G., Kuo W., Schulz G. et al.: Virtual histology of an entire mouse brain from formalin fixation to paraffin embedding. Part 1: Data acquisition, anatomical feature segmentation, tracking global volume and density changes. *Journal of Neuroscience Methods* (2021); **364**: 109354.
9. Kawata N., Teplov A., Ntiamoah P., Shia J., Hameed M., Yagi Y.: Micro-computed tomography: A novel diagnostic technique for the evaluation of gastrointestinal specimens. *Endoscopy International Open* (2021); **09**: E1886–9.
10. Scott A.E., Vasilescu D.M., Seal K.A.D. et al.: Three Dimensional Imaging of Paraffin Embedded Human Lung Tissue Samples by Micro-Computed Tomography. *PLOS ONE* (2015); **10**: e0126230.
11. Khimchenko A., Deyhle H., Schulz G. et al.: Extending two-dimensional histology into the third dimension through conventional micro computed tomography. *NeuroImage* (2016); **139**: 26–36.
12. Busse M., Müller M., Kimm M.A. et al.: Three-dimensional virtual histology enabled through cytoplasm-specific X-ray stain for microscopic and nanoscopic computed tomography. *Proceedings of the National Academy of Sciences* (2018); **115**: 2293–8.
13. Senter-Zapata M., Patel K., Bautista P.A., Griffin M., Michaelson J., Yagi Y.: The Role of Micro-CT in 3D Histology Imaging. *Pathobiology: Journal of Immunopathology, Molecular and Cellular Biology* (2016); **83**: 140–7.
14. Walton L.A., Bradley R.S., Withers P.J. et al.: Morphological Characterisation of Unstained and Intact Tissue Micro-architecture by X-ray Computed Micro- and Nano-Tomography. *Scientific Reports* (2015); **5**: 10074.

15. Teplov A., Tabata K., Fu X. et al.: Development of Standard Operating Procedure (SOP) of Micro-computed tomography (micro-CT) in Pathology. *Diagnostic Pathology* (2019); **5**.
16. Pauwels R., Jacobs R., Singer S.R., Mupparapu M.: CBCT-based bone quality assessment: are Hounsfield units applicable? *Dentomaxillofacial Radiology* (2015); **44**: 20140238.
17. Clark D.P., Ghaghada K., Moding E.J., Kirsch D.G., Badea C.T.: In vivo characterization of tumor vasculature using iodine and gold nanoparticles and dual energy micro-CT. *Physics in Medicine and Biology* (2013); **58**: 1683–704.

X-RAY MICROTOMOGRAPHY IMAGING OF FRESH AND FORMALIN-FIXED POULTRY HEART

Abstract

X-ray microtomography (μCT) is an imaging technique of growing popularity in recent years. It can visualize small samples with a resolution of the final image reaching micrometers. It remains an open question whether μCT can – to some extent – replace or complement histopathological imaging. This study aims to compare μCT images of fresh and formalin-fixed poultry heart specimens. The following features were calculated: the heart volume, its dimensions along main axes, distributions of voxel intensities. The obtained results indicated that the process of formalin fixation of the poultry heart did not negatively affect the imaging of the tissue, and muscle and fat can be distinguished clearly in the images. Fixation also reduces the severity of motion artifacts related to the shrinkage of the sample.

Keywords: X-ray microtomography, formalin-fixed tissue sample, *in-vitro* study

DOI:10.34918/85110

Piotr WILCZEK[1,*]

Chapter 10. AN APPLICATION OF THE LOCAL BINARY PATTERN ALGORITHM AND ITS UNIFORM VARIANT TO IMPROVE THE RECURRENCE AND CROSS-RECURRENCE QUANTIFICATION ANALYSES OF THE PHARMACOLOGICALLY AND PHYSIOLOGICALLY IMPORTANT TIME SERIES

10.1. Introduction

Since its introduction in 1987 by J.-P. Eckmann et al. [2] and the development of several quantification approaches, the so-called *recurrence plots* (RPs) were widely used for the study of complex systems in a variety of scientific disciplines, as physiology, ecology, finance or earth sciences [10]. The RP and related methodologies are aimed to visualize and quantitatively assess the recurrences of nonlinear dynamical systems. For instance, if we consider a trajectory $\{\vec{x}_i\}_{i=1}^{n}$ of some dynamical system in its phase space, then it is straightforward to realize that the evolution of this system can be described by a time series whose components are the above vectors. The corresponding RP efficiently visualizes recurrences of the analyzed system and can be formally obtained from the following *recurrence matrix* (R) [10]:

$$R_{i,j}(\varepsilon) \coloneqq \Theta\left(\varepsilon - \left\|\vec{x}_i - \vec{y}_j\right\|_2\right) \tag{1}$$

for $i, j = 1, \ldots, n$ where n is the number of *measured time points*, ε is a *threshold distance*, $\Theta(\cdot)$ is the *Heaviside function* (i.e., $\Theta(\cdot) = 0$ if $x < 0$ and $\Theta(x) = 0$ otherwise) and $\|\cdot\|_2$ is the *Euclidean norm*. For ε-recurrent states, i.e., for states which are in an ε-neighbourhood, it is possible to introduce the subsequent condition [10]:

$$\vec{x}_i \approx \vec{x}_j \leftrightarrow R_{i,j} = 1. \tag{2}$$

---

[1] Computer Laboratory, Poznań, Poland.

[*] Corresponding author: piotr.wilczek.net@onet.pl.

Then, the RP structure is obtained by plotting a black dot at the coordinates (i, j) if $R_{i,j} = 1$ and a white dot if $R_{i,j} = 0$. Both axes of the RP are identified with time axes. Since $R_{i,i} = 1$ by convention, the RP has always a black main diagonal line, i.e., the line of identity (LOI). Moreover, the RP is symmetric by definition with respect to the main diagonal, i.e., $R_{i,j} = R_{j,i}$. Roughly speaking, the RP object juxtaposes the states of a nonlinear system at times i and j, i.e., if states are similar, then $R_{i,j} = 1$ and otherwise $R_{i,j} = 0$.

In turn, the *cross-recurrence plot* (CRP) can be understood as a bivariate extension of the RP and was proposed in order to analyze the dependencies between two different nonlinear dynamical systems by comparing their states [10]. Namely, if we consider two systems, each one represented by the trajectories $\vec{x}_i$ and $\vec{y}_j$ in the same phase space, then the *cross-recurrence matrix* (CR) is defined by the condition [10]:

$$CR_{i,j}^{\vec{x},\vec{y}}(\varepsilon) \coloneqq \Theta\left(\varepsilon - \left\|\vec{x}_i - \vec{y}_j\right\|_2\right) \tag{3}$$

for $i = 1, \dots, n$ and $j = 1, \dots, m$ where the length of the trajectories of both dynamical systems is not required to be identical and, consequently, the CR matrix is not necessarily square. Then, for ε-recurrent states, i.e., for states which are in an ε-neighbourhood, it is possible to introduce the subsequent condition [10]:

$$\vec{x}_i \approx \vec{y}_j \leftrightarrow CR_{i,j} = 1. \tag{4}$$

Finally, the CRP is obtained by plotting a black dot at the coordinates (i, j) if $CR_{i,j} = 1$ and a white dot if $CR_{i,j} = 0$. Thus, it can be easily observed that the CRP structure compares the states of two nonlinear systems at times i and j, i.e., if states are similar, then $CR_{i,j} = 1$ and otherwise $CR_{i,j} = 0$. In other words, the CRP object detects those time points when a state of the first system recurs to one of the second system. There is usually not a black main diagonal line on the CRP structure because the values of the main diagonal $CR_{i,i}$ are not necessarily equal to one.

In the present paper, we will show the usefulness of the *recurrence quantification analysis* (RQA) in investigating the pharmacological and physiological nonlinear time series. Moreover, we will propose to apply the *local binary pattern* (LBP) algorithm and its uniform variant (ULBP) to improve the performance of the RP and CRP methodologies. The paper is structured as follows: Section 3 details the RQA complexity descriptors, Section 4 presents the novel (U)LBP-based RQA complexity indices, Section 5 describes the datasets and computational procedure whereas Section 6 includes the results of the analysis of the exemplary time series. Finally, Section 7 concludes the paper.

10.2. The RQA complexity descriptors

A crucial parameter of the RP algorithm is the threshold distance ε [10]. According to J.P. Zbilut and coworkers, it is possible to choose ε with respect to the *recurrence rate* ($RR(\varepsilon)$) of the R matrix [21]. For the time series of the length n, this matrix-theoretical index is expressed by the formula [10]:

$$RR(\varepsilon) := \frac{1}{n^2} \sum_{i,j=1}^{n} R_{i,j}(\varepsilon). \tag{5}$$

It was suggested to select ε such that the $RR(\varepsilon)$ coefficient is approximately equal to 1% [21]. Moreover, the RQA developed several complexity descriptors in order to analyze nonlinear time series data and to extract meaningful statistics and other characteristics of the data [10]. All listed below RQA indices refer to the RP object obtained from a time series of the length n.

10.2.1. The complexity measures based on diagonal lines

The *histogram of diagonal lines of length* l ($P(\varepsilon, l)$) has the form [10]:

$$P(\varepsilon, l) := \sum_{i,j=1}^{n} \left(1 - R_{i-1,j-1}(\varepsilon)\right)\left(1 - R_{i+l,j+l}(\varepsilon)\right) \prod_{k=0}^{l-1} R_{i+k,j+k}(\varepsilon). \tag{6}$$

In the rest of the present article, the symbol ε will be omitted from the RQA measures for the sake of simplicity (i.e., $P(l) = P(\varepsilon, l)$). Then, the *determinism* (DET) of the RP structure is identified with the ratio of recurrence points that constitute diagonal structures (of at least length l_{min}) to all recurrence points [10], i.e.,

$$DET := \frac{\sum_{l=l_{min}}^{n} lP(l)}{\sum_{l=1}^{n} lP(l)}. \tag{7}$$

This descriptor is aimed to quantify the predictability of the dynamical system. In turn, the *average diagonal line length* (L_{mean}) of the RP structure, the *length of the longest diagonal line* (L_{max}) found in the RP and its *inverse* (i.e., the so-called *divergence – DIV*) are expressed by the subsequent conditions [10]:

$$L_{mean} \coloneqq \frac{\sum_{l=l_{min}}^{n} lP(l)}{\sum_{l=l_{min}}^{n} P(l)}, \tag{8}$$

$$L_{max} \coloneqq \max\left(\{l_i\}_{i=1}^{n_l}\right) \tag{9}$$

and by

$$DIV \coloneqq \frac{1}{L_{max}} \tag{10}$$

where $n_l \coloneqq \sum_{l \geq l_{min}} P(l)$ refers to the total number of diagonal lines, respectively. If in the computation of the L_{mean} index, we exclude the main diagonal line, then the resultant descriptor is denoted by the symbol L_{mean}^{-}. If we identify the *probability* of finding a diagonal line of the length l in the RP structure with the formula $p(l) \coloneqq \frac{P(l)}{n_l}$, then the *Shannon entropy of this distribution* ($ENTR$) is defined as [10]:

$$ENTR \coloneqq -\sum_{l=l_{min}}^{n} p(l)\ln p(l). \tag{11}$$

A further descriptor used to quantitatively assess the RP object is the *ratio* ($RATIO$) defined as the ratio between DET and RR. Formally, this index has the form [10]:

$$RATIO \coloneqq n^2 \frac{\sum_{l=l_{min}}^{n} lP(l)}{(\sum_{l=1}^{n} lP(l))^2}. \tag{12}$$

The RR measure computed for a certain line parallel to the LOI with the distance τ from the LOI is called the *τ-recurrence rate* ($\tau - RR$). The $\tau - RR$ for the diagonal lines with the distance τ from the LOI is given by the formula [10]:

$$RR_\tau \coloneqq \frac{1}{n-\tau}\sum_{i=1}^{n-\tau} R_{i,i+\tau} = \frac{1}{n-\tau}\sum_{l=1}^{n-\tau} lP_\tau(l) \tag{13}$$

where the symbol $P_\tau(l)$ refers to the number of diagonal lines of the length l on each diagonal $R_{i,i+\tau}$ parallel to the LOI. Namely, $\tau = 0$ corresponds to the main diagonal, $\tau > 0$ to diagonals above and $\tau < 0$ to diagonals below the LOI (i.e., $R_{i+|\tau|,i}$). Then, a linear regression coefficient over the $\tau - RR$ of the diagonals parallel to the LOI as a function of the distance between these diagonals and the LOI is called the *trend* ($TREND$) and is defined as [10]:

$$TREND := \frac{\sum_{\tau=1}^{\bar{n}} \left(\tau - \frac{\bar{n}}{2}\right) (RR_\tau - \langle RR_\tau \rangle)}{\sum_{\tau=1}^{\bar{n}} \left(\tau - \frac{\bar{n}}{2}\right)^2} \tag{14}$$

where $\bar{n} < n$ and the symbol $\langle \cdot \rangle$ denotes the average value.

10.2.2. The complexity measures based on vertical lines

The total number of vertical lines of the length v in the RP is calculated from the following histogram [10]:

$$P(v) := \sum_{i,j=1}^{n} (1 - R_{i,j})(1 - R_{i,j+v}) \prod_{k=0}^{v-1} R_{i,j+k}. \tag{15}$$

Then, the ratio between the recurrence points constituting vertical structures and the entire set of recurrence points is known as the *laminarity* (LAM) of the RP object. Formally, this descriptor is defined as [10]:

$$LAM := \frac{\sum_{v=v_{min}}^{n} vP(v)}{\sum_{v=1}^{n} vP(v)}. \tag{16}$$

Moreover, the *maximal length of vertical lines* (V_{max}) in the RP structure as well as the average length of vertical lines (V_{mean}) in the RP object are given by the conditions [10]:

$$V_{max} := max\left(\{v_l\}_{l=1}^{n_v}\right) \tag{17}$$

where n_v is the absolute number of vertical lines and

$$V_{mean} := \frac{\sum_{v=v_{min}}^{n} vP(v)}{\sum_{v=v_{min}}^{n} P(v)}, \tag{18}$$

respectively. Equivalently, the V_{mean} complexity descriptor is called the *trapping time* [10].

10.3. The novel (U)LBP-based RQA complexity descriptors

The recurrence and cross-recurrence matrices whose definitions include the threshold distance ε can be regarded as the *thresholded* recurrence and cross-recurrence two--dimensional arrays. Besides these thresholded objects, it is possible to define the *unthresholded recurrence matrix* (UR) as well as the *unthresholded cross-recurrence matrix* (UCR) [10]. Formally, for a nonlinear dynamical system, its UR matrix has the form [10]:

$$UR_{i,j} \coloneqq \left\| \vec{x}_i - \vec{y}_j \right\|_2 \tag{19}$$

for $i, j = 1, \dots, n$ where n is the number of measured time points. In turn, if we study two dynamical systems, each one represented by the trajectories $\vec{x}_i$ and $\vec{y}_j$ in the same phase space, then their UCR matrix is given by the condition [10]:

$$UCR_{i,j}^{\vec{x},\vec{y}} \coloneqq \left\| \vec{x}_i - \vec{y}_j \right\|_2 \tag{20}$$

for $i = 1, \dots, n$ and $j = 1, \dots, m$ where the length of both trajectories is not required to be identical and, therefore, the UCR array is not necessarily square. Moreover, the entries of both untresholded matrices can be transformed according to the following rules [3]:

$$\mathcal{UR}_{i,j} \coloneqq e^{-UR_{i,j}} \tag{21}$$

and

$$\mathcal{UCR}_{i,j}^{\vec{x},\vec{y}} \coloneqq e^{-UCR_{i,j}^{\vec{x},\vec{y}}}, \tag{22}$$

respectively. The resultant negatively exponentially transformed matrices are denoted by the calligraphic letters. Thus, for very similar states, the values of the entries of both matrices near one and for identical states they are exactly equal to one. In turn, for very dissimilar states, the values of the entries of both two-dimensional arrays approach zero [3]. Moreover, in the present paper, we propose to divide the values of the entries of both transformed matrices into ten equivalence classes. These classes are identified with the values 0.1, 0.2, …, 0.9, 1 and the resultant *coarse-grained* unthresholded recurrence and cross-recurrence matrices are denoted by the symbols $cg\mathcal{UR}$ and $cg\mathcal{UCR}$, respectively. In details, using the notation of the form $M \in \{\mathcal{UR}, \mathcal{UCR}\}$ and $cgM \in \{cg\mathcal{UR}, cg\mathcal{UCR}\}$, the equivalence classes, i.e., the entries of both coarse-grained matrices are defined by the conditions included in Table 1[2]. Next, we propose to run the *local binary pattern* (LBP) algorithm and/or its simplified counterpart on both unthresholded matrices and their coarse-grained counterparts in order to obtain the novel arrays whose entries are 8-*digit binary numbers* usually converted to *decimal numbers* for convenience. Recall that the LBP

---

[2] In Table 1, the symbol ∧ refers to the logical *conjunction*.

technique is a type of visual texture descriptor used for classification tasks in the realm of *computer vision* [1, 8, 11, 12]. Although the LBP methodology started in 1994 [12], the present contribution can be regarded as a first attempt to apply the LBP algorithm and its modified variant in the analysis of the nonlinear dynamical systems, for instance, in studying the dose-response time series data.

Table 1

The formal conditions defining the coarse-grained unthresholded matrices

ENTRY	CONDITION	ENTRY	CONDITION
$cgM_{i,j} = 0.1$	$M_{i,j} < 0.1$	$cgM_{i,j} = 0.6$	$M_{i,j} \geq 0.5 \wedge M_{i,j} < 0.6$
$cgM_{i,j} = 0.2$	$M_{i,j} \geq 0.1 \wedge M_{i,j} < 0.2$	$cgM_{i,j} = 0.7$	$M_{i,j} \geq 0.6 \wedge M_{i,j} < 0.7$
$cgM_{i,j} = 0.3$	$M_{i,j} \geq 0.2 \wedge M_{i,j} < 0.3$	$cgM_{i,j} = 0.8$	$M_{i,j} \geq 0.7 \wedge M_{i,j} < 0.8$
$cgM_{i,j} = 0.4$	$M_{i,j} \geq 0.3 \wedge M_{i,j} < 0.4$	$cgM_{i,j} = 0.9$	$M_{i,j} \geq 0.8 \wedge M_{i,j} < 0.9$
$cgM_{i,j} = 0.5$	$M_{i,j} \geq 0.4 \wedge M_{i,j} < 0.5$	$cgM_{i,j} = 1$	$M_{i,j} \geq 0.9$

More specifically, our paper proposes to scrutinize the recurrences of the studied dynamical system by extracting from its (coarse-grained) unthresholded recurrence and cross-recurrence matrices the (uniform) local binary patterns in the form of feature vectors. The LBP procedure can be run on *any* real matrices (for instance, on the $\mathcal{UR}, \mathcal{UCR}$ or on their coarse-grained analogues) whose entries are interpreted as *pixels*. A single instantiation of the LBP procedure is defined on 3×3 submatrix (of the input $n \times m$ two-dimensional real array) consisting the central pixel at the position (2,2) and eight neighboring pixels. Formally, this type of neighborhood is called the *Moore neighborhood*, i.e., all eight neighboring pixels are located at the *Chebyshev distance* equal to one from the central pixel. In the LBP algorithm, all neighboring pixels that have the values higher than or equal to the value of the central pixel are set to one otherwise they are set to zero. Then, starting from the pixel located at the cell (1,2) and moving clockwise to the pixel located at the cell (1,1), we obtain the 8-digit binary number. For convenience, the resultant binary number is converted to the corresponding decimal number. This decimal value is regarded as the LBP pattern of the central pixel. Consequently, if we repeat the same process for all 3×3 submatrices of the input $n \times m$ real matrix, then we obtain the histogram of all LBP values. Moreover, this histogram can be treated as the *feature vector*. From the fact that the number of 8-digit binary *words* is $2^8 = 256$, it follows that the dimensionality of the output feature vector is equal to 256. In turn, its length is equal to $nm - 2n - 2m + 4$. We will denote this numeric object by the symbol VB_{256}. Alternatively, the VB_{256} vector can be matricized into the $(n-2) \times (m-2)$ matrix denoted by the symbol B_{256}.

This matricization is performed column-wise, i.e., the first $(n-2)$ values of the VB_{256} object constitute the first column of the B_{256} matrix and the last $(m-2)$ values of the VB_{256} object form the last column of the B_{256} matrix. Some useful extension of the original LBP algorithm is the so-called *uniform local binary pattern* (ULBP) operator. This modified procedure can be used to reduce the length of the original feature vector. The conceptual idea of this modified procedure is motivated by the fact that some binary patterns are more frequent in texture images than other. The LBP is called *uniform* if it contains at most two 0-1 or 1-0 *transitions*. For instance, the word 0001000 includes two transitions and is treated as uniform, but the word 01010100 includes six transitions and, consequently, can not be regarded as uniform. In the computation of the ULBP histogram, the resultant histogram has a separate bin for every uniform pattern whereas all non-uniform patterns are assigned to one bin. It can be easily observed that among 256 8-digit binary words, it is possible to single out 58 uniform words. Consequently, by using the ULBP algorithm, the dimensionality of the resultant feature vector is reduced from 256 to 59. In this case, for the input $n \times m$ two-dimensional real array, the output 59-dimensional feature vector has also the length equal to $nm - 2n - 2m + 4$. In the ULBP approach, all uniform words are indexed by the decimal number from 0 to 57 whereas all non-uniform words are assigned to the decimal value equal to 58. Therefore, the obtained 59-dimensional vector is denoted by VB_{59} whereas its matricized form is denoted by B_{59}. In the remainder of the present article, the both types of feature vectors, i.e., VB_{59} and VB_{256} as well as their matricized forms, i.e., B_{59} and B_{256} derived from the arrays $\mathcal{UR}, \mathcal{UCR}$ are denoted by the symbols: $VB_{59}^{\mathcal{UR}}$, $VB_{59}^{\mathcal{UCR}}$, $VB_{256}^{\mathcal{UR}}$, $VB_{256}^{\mathcal{UCR}}$ and by $B_{59}^{\mathcal{UR}}$, $B_{59}^{\mathcal{UCR}}$, $B_{256}^{\mathcal{UR}}$, $B_{256}^{\mathcal{UCR}}$, respectively. In turn, the feature vectors and their matrix counterparts derived from the coarse-grained structures are denoted by the symbols: $VB_{59}^{cg\mathcal{UR}}$, $VB_{59}^{cg\mathcal{UCR}}$, $VB_{256}^{cg\mathcal{UR}}$, $VB_{256}^{cg\mathcal{UCR}}$ and by $B_{59}^{cg\mathcal{UR}}$, $B_{59}^{cg\mathcal{UCR}}$, $B_{256}^{cg\mathcal{UR}}$, $B_{256}^{cg\mathcal{UCR}}$, respectively.

Example 1. For the 6-element time series $ts = (2, 6, 4, 9, 0, 5)$, its UR matrix and its negatively exponentially transformed $\mathcal{UR}$ derivative have the forms

$$UR = \left|\begin{array}{ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 0 & 4 & 2 & 7 & 2 & 3 \\ 2 & 4 & 0 & 2 & 3 & 6 & 1 \\ 3 & 2 & 2 & 0 & 5 & 4 & 1 \\ 4 & 7 & 3 & 5 & 0 & 9 & 4 \\ 5 & 2 & 6 & 4 & 9 & 0 & 5 \\ 6 & 3 & 1 & 1 & 4 & 5 & 0 \end{array}\right|,$$

$$\mathcal{UR} = \left|\begin{array}{ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 1 & 0.018 & 0.135 & 0.001 & 0.135 & 0.05 \\ 2 & 0.018 & 1 & 0.135 & 0.05 & 0.002 & 0.368 \\ 3 & 0.135 & 0.135 & 1 & 0.007 & 0.018 & 0.368 \\ 4 & 0.001 & 0.05 & 0.007 & 1 & 0 & 0.018 \\ 5 & 0.135 & 0.002 & 0.018 & 0 & 1 & 0.007 \\ 6 & 0.05 & 0.368 & 0.368 & 0.018 & 0.007 & 1 \end{array}\right|,$$

respectively.

The entries of the coarse-grained counterpart of the $\mathcal{UR}$ array are given below

$$cg\mathcal{UR} = \left|\begin{array}{ccccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 1 & 0.1 & 0.2 & 0.1 & 0.2 & 0.1 \\ 2 & 0.1 & 1 & 0.2 & 0.1 & 0.1 & 0.4 \\ 3 & 0.2 & 0.2 & 1 & 0.1 & 0.1 & 0.4 \\ 4 & 0.1 & 0.1 & 0.1 & 1 & 0.1 & 0.1 \\ 5 & 0.2 & 0.1 & 0.1 & 0.1 & 1 & 0.1 \\ 6 & 0.1 & 0.4 & 0.4 & 0.1 & 0.1 & 1 \end{array}\right|.$$

The feature vectors obtained by the LBP and ULBP approaches have the forms

$$VB_{58}^{\mathcal{UR}} = \{58, 58, 58, 56, 58, 58, 58, 58, 58, 58, 58, 57, 56, 58, 57, 58\},$$

$$VB_{256}^{\mathcal{UR}} = \{17, 226, 197, 254, 142, 17, 235, 93, 71, 175, 17, 255, 254, 117, 255, 17\},$$

$$VB_{58}^{cg\mathcal{UR}} = \{58, 58, 57, 57, 58, 58, 57, 57, 57, 57, 58, 57, 57, 57, 57, 58\}$$

and

$$VB_{256}^{cg\mathcal{UR}} = \{17, 226, 255, 255, 142, 17, 255, 255, 255, 255, 17, 255, 255, 255, 255, 17\},$$

respectively. The corresponding matricized counterparts are given below

$$B_{58}^{\mathcal{UR}} = \left|\begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ 1 & 58 & 58 & 58 & 56 \\ 2 & 58 & 58 & 58 & 58 \\ 3 & 58 & 58 & 58 & 57 \\ 4 & 56 & 58 & 57 & 58 \end{array}\right|, B_{256}^{\mathcal{UR}} = \left|\begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ 1 & 17 & 142 & 71 & 254 \\ 2 & 226 & 17 & 175 & 117 \\ 3 & 197 & 235 & 17 & 255 \\ 4 & 254 & 93 & 255 & 17 \end{array}\right|$$

and

$$B_{58}^{cg\mathcal{UR}} = \left|\begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ 1 & 58 & 58 & 57 & 57 \\ 2 & 58 & 58 & 57 & 57 \\ 3 & 57 & 57 & 58 & 57 \\ 4 & 57 & 57 & 57 & 58 \end{array}\right|, B_{256}^{cg\mathcal{UR}} = \left|\begin{array}{ccccc} & 1 & 2 & 3 & 4 \\ 1 & 17 & 142 & 255 & 255 \\ 2 & 226 & 17 & 255 & 255 \\ 3 & 255 & 255 & 17 & 255 \\ 4 & 255 & 255 & 255 & 17 \end{array}\right|,$$

respectively.

For the unthresholded non-square cross-recurrence matrices and their coarse-grained derivatives the above algorithm is identical.

In the current paper, to improve the performance of the RQA methodology, we propose to introduce the several novel LBP-based and ULBP-based RQA descriptors to analyze nonlinear time series. These newly suggested complexity indices are classified as vector-based and matrix-based complexity measures. Namely, the vector-based metrics are extracted from the feature vectors obtained by the action of the (U)LBP algorithm on the arrays given by the formulae (21) and (22) as well as on their coarse-grained analogues. In turn, the matrix-based metrics are extracted from the matricized versions of these vectors. Accordingly, for any feature vector VB instantiated by the cases

$$VB \in \left\{VB_{58}^{\mathcal{UR}}, VB_{256}^{\mathcal{UR}}, VB_{58}^{cg\mathcal{UR}}, VB_{256}^{cg\mathcal{UR}}, VB_{58}^{\mathcal{UCR}}, VB_{256}^{\mathcal{UCR}}, VB_{58}^{cg\mathcal{UCR}}, VB_{256}^{cg\mathcal{UCR}}\right\} \tag{23}$$

and its matricized derivative B instantiated by the cases

$$B \in \left\{B_{58}^{\mathcal{UR}}, B_{256}^{\mathcal{UR}}, B_{58}^{cg\mathcal{UR}}, B_{256}^{cg\mathcal{UR}}, B_{58}^{\mathcal{UCR}}, B_{256}^{\mathcal{UCR}}, B_{58}^{cg\mathcal{UCR}}, B_{256}^{cg\mathcal{UCR}}\right\}, \tag{24}$$

we suggest to consider two types of complexity indices whose formulae are included in Subsections 4.1 and 4.2, respectively.

10.3.1. The vector-based complexity measures

For the feature vector whose elements are indexed by $i \in \{1, 2, \ldots, |VB|\}$, we will test the following complexity descriptors

10.3.1.1. The *arithmetical mean* ($\langle\cdot\rangle$), i.e.,

$$\langle VB\rangle := \frac{1}{|VB|}\sum_{i=1}^{|VB|} VB_i \tag{25}$$

where $\langle\cdot\rangle$ refers to the arithmetical mean and $|VB|$ is the length of the studied vector.

10.3.1.2. The *variance* (Var), i.e.,

$$Var(VB) := \frac{1}{|VB|}\sum_{i=1}^{|VB|}(VB_i - \langle VB\rangle)^2. \tag{26}$$

10.3.1.3. The *coefficient of variation* (CV) also known as the *relative standard deviation*, i.e.,

$$CV(VB) := \frac{\sqrt{Var(VB)}}{\langle VB \rangle}. \tag{27}$$

10.3.1.4. The *Gini index* (Gi), i.e.,

$$Gi(VB) := \frac{2\sum_{i=1}^{|VB|} iVB_i^*}{|VB|\sum_{i=1}^{|VB|} VB_i} - \frac{|VB|+1}{|VB|} \tag{28}$$

where VB_i^* denotes the vector of the VB_i values arranged in non-decreasing order, i.e., $VB_i \leq VB_{i+1}$.

10.3.1.5. The *structural information content* ($I(VB)$) also known as the *parametric entropy* [16, 19]. To present this index, the subsequent auxiliary notions must be introduced: First, we define the following quantity [16, 19]:

$$p(VB_i) := \frac{VB_i}{\sum_{j=1}^{|VB|} VB_j}. \tag{29}$$

In *Chemical Graph Theory*, the quantity expressed by the condition (29) is known as the *information functional* and is identified with a *positive* and *monotonous* function that captures structural information included in a *chemical graph* by defining the finite probability value for each graph node. In this case, the input feature vector is identified with some sequence of the vertex invariants [16, 19]. Next, from the reason that the following equation $p(VB_1) + p(VB_2) + \cdots + p\left(VB_{|VB|}\right) = 1$ is *a priori* valid, it follows that the vector $P(VB) = \left(p(VB_1), p(VB_2), \ldots, p\left(VB_{|VB|}\right)\right)$ constitutes a *finite probability distribution*. Its structural information content is expressed by the following formula [16, 19]:

$$I(VB) := -\sum_{i=1}^{|VB|} \frac{VB_i}{\sum_{j=1}^{|VB|} VB_j} \log_2\left(\frac{VB_i}{\sum_{j=1}^{|VB|} VB_j}\right). \tag{30}$$

10.3.1.6. The *mean information content* ($\bar{I}(VB)$) [16, 19]. First, we define the equivalence relation $\simeq$ on the analyzed vector given by the condition: $VB_i \simeq VB_j \leftrightarrow VB_i = VB_j$. Therefore, it is possible to obtain the partition of VB, where the resultant partitions are symbolized by $VB^1, VB^2, \dots, VB^h, \dots, VB^k$ where $k \leq |VB|$. In this context, the quantities $p_h = \frac{|VB^h|}{|VB|}$ for $h \in \{1,2, \dots, k\}$ can be regarded as probabilities for each singled out partition VB^h. Moreover, it is apparent that $0 \leq p_h \leq 1$ as well as $\sum_{h=1}^{k} p_h = 1$. Accordingly, the vector $P(VB) = (p_1, p_2, \dots, p_h, \dots, p_k)$ can be also treated as a *finite probability distribution* over the VB object. Its mean information content is given by the expression [16, 19]:

$$\bar{I}(VB) \coloneqq -\sum_{h=1}^{k} p_h \log_2 p_h. \tag{31}$$

Thus, for any feature vector VB which can be partitioned into k many disjoint subsets of the cardinalities equal to $|VB^h|$ for $h \in \{1,2, \dots, k\}$, we obtain the following information-theoretical complexity measure [16, 19]:

$$\bar{I}(VB) \coloneqq -\sum_{h=1}^{k} \frac{|VB^h|}{|VB|} \log_2 \left(\frac{|VB^h|}{|VB|} \right). \tag{32}$$

Note that the structural and mean information content complexity indices defined on the graph centrality measures were successfully used to quantify the intricacy of the complex networks with respect to these vertex invariants [19].

10.3.2. The matrix-based complexity measures

For the matrix B whose rows and columns are indexed by $i \in \{1, 2, \dots, n-2\}$ and by $j \in \{1, 2, \dots, m-2\}$, respectively, we will test the following matrix-based complexity descriptors

10.3.2.1. The *Frobenius norm* ($\|B\|_F$), i.e.,

$$\|B\|_F \coloneqq \sqrt{tr(B^T B)} \tag{33}$$

where $tr(\cdot)$ refers to the trace function whereas the subscript ^{T} denotes the matrix transposition.

10.3.2.2. The *maximum norm* ($\|B\|_{max}$), i.e.,

$$\|B\|_{max} \coloneqq \max_{i,j} |B_{i.j}|. \tag{34}$$

10.3.2.3. The *maximum absolute column sum norm* ($\|B\|_1$), i.e.,

$$\|B\|_1 := \max_j \left(\sum_{i=1}^{n-2} \left| B_{i,j} \right| \right). \tag{35}$$

10.3.2.4. The *maximum absolute row sum norm* ($\|B\|_\infty$), i.e.,

$$\|B\|_\infty := \max_i \left(\sum_{j=1}^{m-2} \left| B_{i,j} \right| \right). \tag{36}$$

10.3.2.5. The *spectral norm* ($\|B\|_2$), i.e.,

$$\|B\|_2 := \sigma_{max}(B) \tag{37}$$

where σ_{max} is the largest singular value of B.

10.3.2.6. The *rank* ($rk(B)$), i.e.,

$$rk(B) := \text{the number of non} - \text{zero singular values.} \tag{38}$$

The performance of the above descriptors will be quantitatively assessed in the time series classification task experiments presented in Section 6.

10.4. The datasets and computational methodology

In the pharmacological time series analysis, the use of time-dependent measurements of the cell proliferation can relate drug-responses to cell phenotypes. The HTS007 dataset downloaded from https://demo.thunor.net includes a panel of eight breast cancer cell lines treated with 27 anticancer drugs at several different concentrations. The cell proliferation was quantified over 5 days [9]. All time series contained in the above dataset consist of 24 measured time points and they have two class labels: control and non-control. In our computational experiments, we selected eight dose-response time series corresponding to the following anticancer drugs: *abemaciclib*, *alpelisib*, *bleomycin*, *cediranib*, *dasatinib*, *etoposide*, *everolimus* and *ipatasertib* whose efficiency was tested on the BT20 cell line. In sum, our numerical experiments were performed on 393 time-dependent cell proliferation measurements. Moreover, for a comparative purpose, we tested our improved RQA methodology on the

ECG200 dataset composed of 200 time series. This dataset was downloaded from https://www.timeseriesclassification.com. Each time series in this dataset contains 96 measured time points. All time series contained in the ECG200 dataset have also two class labels: normal heartbeat and myocardial infarction. In the numerical classification experiments, we perform the 10-fold cross validation using *1-nearest neighbors classifier* (1-NN classifier). In the traditional RQA experiments, we followed the proposal of J.P. Zbilut and coworkers and we set the threshold distance ε such that the RR coefficient is approximately equal to 1% (cf. Table 2) [21]. Moreover, the RQA parameters [10], i.e., l_{min}, v_{min}, the *embedding dimension* and the *time lag* were set to two, two, one and to zero, respectively. All computations were performed in the R programming language [4, 6, 7, 13–15, 18].

10.5. Results and discussion

The LPB and ULPB algorithms were run on the matrices (21) and (22). The matrix (21) was obtained from the raw time series whereas the matrix (22) was obtained from the raw time series and its first differenced variant. In our numerical classification experiments, we try to predict the class labels of the time series contained in the HTS007 and ECG200 datasets. The results of the performance of the 1-NN classifier, i.e., the number of correctly classified time series are recorded in percentage (scaled to the range from zero to one). In the tables below, the names of time series are identified with the names of tested anticancer drugs. In all cases, the best results are in bold.

Table 2

The RQA results for five time series

INDEX	ABEMACICLIB	ALPELISIB	BLEOMYCIN	CEDIRANIB	DASATINIB
RR	**0.784**	0.674	0.608	0.625	0.562
$RATIO$	0.765	**0.696**	0.549	0.688	0.646
DET	0.647	0.435	0.471	0.604	0.562
DIV	0.689	0.674	**0.647**	0.688	0.667
L_{max}	0.745	0.674	**0.647**	0.708	0.667
L_{mean}	0.686	0.652	0.451	0.562	0.583
L_{mean}^{-}	0.6	0.522	0.471	0.667	0.604
$ENTR$	0.725	0.565	0.608	0.688	**0.771**
$TREND$	0.667	0.543	**0.647**	**0.729**	0.542
LAM	0.667	**0.696**	0.588	0.667	0.5
V_{max}	0.725	**0.696**	0.627	0.708	**0.771**
V_{mean}	0.706	0.587	0.549	0.688	0.604

Table 3

The RQA results for four time series

INDEX	ETOPOSIDE	EVEROLIMUS	IPATASERTIB	ECG
RR	0.64	0.66	**0.612**	0.575
$RATIO$	0.64	0.62	**0.612**	0.56
DET	0.48	0.6	0.51	0.595
DIV	**0.72**	**0.72**	0.592	0.64
L_{max}	0.7	0.7	0.551	0.615
L_{mean}	0.54	0.54	0.429	0.585
$\bar{L}_{mean}$	0.64	0.66	0.469	**0.67**
$ENTR$	0.6	0.6	0.51	0.59
$TREND$	0.6	0.62	0.571	0.6
LAM	0.58	0.62	0.571	0.59
V_{max}	0.66	0.66	**0.612**	0.66
V_{mean}	**0.72**	**0.72**	0.551	0.63

Table 4

The (U)LBP-based RQA results for Abemaciclib using the vector-based descriptors

Index	$VB_{58}^{\mathcal{UR}}$	$VB_{256}^{\mathcal{UR}}$	$VB_{58}^{cg\mathcal{UR}}$	$VB_{256}^{cg\mathcal{UR}}$	$VB_{58}^{\mathcal{UCR}}$	$VB_{256}^{\mathcal{UCR}}$	$VB_{58}^{cg\mathcal{UCR}}$	$VB_{256}^{cg\mathcal{UCR}}$
$\langle VB\rangle$	0.667	0.608	0.667	0.745	0.765	0.608	0.608	0.667
$Var(VB)$	0.569	0.745	0.608	0.804	0.706	0.667	0.412	0.706
$CV(VB)$	0.471	0.647	0.627	**0.843**	0.667	0.706	0.569	0.784
$Gi(VB)$	0.608	0.608	0.686	0.745	0.647	0.706	0.627	0.667
$I(VB)$	0.451	0.627	0.706	0.745	0.706	0.608	0.471	0.725
$\bar{I}(VB)$	0.706	0.706	0.608	0.824	0.647	0.765	0.667	0.392

Table 5

The (U)LBP-based RQA results for Abemaciclib using the matrix-based descriptors

Index	$B_{58}^{\mathcal{UR}}$	$B_{256}^{\mathcal{UR}}$	$B_{58}^{cg\mathcal{UR}}$	$B_{256}^{cg\mathcal{UR}}$	$B_{58}^{\mathcal{UCR}}$	$B_{256}^{\mathcal{UCR}}$	$B_{58}^{cg\mathcal{UCR}}$	$B_{256}^{cg\mathcal{UCR}}$
$\lVert B\rVert_F$	0.647	0.706	0.706	0.784	0.49	0.686	0.608	0.706
$\lVert B\rVert_2$	0.588	0.569	0.647	0.784	0.824	0.549	0.49	0.686
$\lVert B\rVert_{max}$	0.608	0.824	0.608	0.608	0.608	0.765	0.608	0.608
$\lVert B\rVert_\infty$	0.627	0.549	0.608	0.686	0.725	0.667	0.608	0.647
$\lVert B\rVert_1$	0.569	0.745	0.471	0.549	0.667	0.686	0.608	0.647
$rk(B)$	0.608	0.608	0.608	0.608	0.725	0.745	0.49	0.49

Table 6

The (U)LBP-based RQA results for Alpelisib using the vector-based descriptors

Index	$VB_{58}^{\mathcal{UR}}$	$VB_{256}^{\mathcal{UR}}$	$VB_{58}^{cg\mathcal{UR}}$	$VB_{256}^{cg\mathcal{UR}}$	$VB_{58}^{\mathcal{UCR}}$	$VB_{256}^{\mathcal{UCR}}$	$VB_{58}^{cg\mathcal{UCR}}$	$VB_{256}^{cg\mathcal{UCR}}$
$\langle VB\rangle$	0.565	0.522	0.543	0.5	0.609	0.587	**0.739**	0.5
$Var(VB)$	0.478	0.565	0.587	0.565	0.457	0.652	0.543	0.674
$CV(VB)$	0.522	0.478	0.587	0.63	0.609	0.587	0.565	0.717
$Gi(VB)$	0.391	0.457	0.63	0.522	0.587	0.5	0.522	0.5
$I(VB)$	0.543	0.522	0.587	0.587	0.5	0.5	0.5	0.63
$\bar{I}(VB)$	0.630	0.652	0.478	0.565	0.565	0.587	0.543	0.565

Table 7

The (U)LBP-based RQA results for Alpelisib using the matrix-based descriptors

Index	$B_{58}^{\mathcal{UR}}$	$B_{256}^{\mathcal{UR}}$	$B_{58}^{cg\mathcal{UR}}$	$B_{256}^{cg\mathcal{UR}}$	$B_{58}^{\mathcal{UCR}}$	$B_{256}^{\mathcal{UCR}}$	$B_{58}^{cg\mathcal{UCR}}$	$B_{256}^{cg\mathcal{UCR}}$
$\|B\|_F$	0.587	0.478	0.717	0.587	0.522	0.587	0.63	0.543
$\|B\|_2$	0.63	0.565	0.5	0.609	0.543	**0.739**	0.609	0.543
$\|B\|_{max}$	0.674	0.717	0.674	0.674	0.674	0.522	0.674	0.674
$\|B\|_\infty$	0.609	0.609	0.522	0.652	0.63	0.587	0.674	0.674
$\|B\|_1$	0.652	0.63	0.543	0.609	0.543	0.37	0.609	0.609
$rk(B)$	0.674	0.674	0.674	0.674	0.609	0.63	0.609	0.609

Table 8

The (U)LBP-based RQA results for Bleomycin using the vector-based descriptors

Index	$VB_{58}^{\mathcal{UR}}$	$VB_{256}^{\mathcal{UR}}$	$VB_{58}^{cg\mathcal{UR}}$	$VB_{256}^{cg\mathcal{UR}}$	$VB_{58}^{\mathcal{UCR}}$	$VB_{256}^{\mathcal{UCR}}$	$VB_{58}^{cg\mathcal{UCR}}$	$VB_{256}^{cg\mathcal{UCR}}$
$\langle VB \rangle$	0.451	0.49	0.549	0.51	0.608	0.529	0.627	0.725
$Var(VB)$	0.627	0.588	0.392	**0.765**	0.588	0.471	0.627	0.667
$CV(VB)$	0.49	0.529	0.608	0.608	0.667	0.529	0.451	0.706
$Gi(VB)$	0.51	0.608	0.647	0.627	0.608	0.51	0.549	0.608
$I(VB)$	0.549	0.549	0.627	0.431	0.549	0.529	0.451	0.549
$\bar{I}(VB)$	0.529	0.529	0.529	0.549	0.333	0.471	0.431	0.471

Table 9

The (U)LBP-based RQA results for Bleomycin using the matrix-based descriptors

Index	$B_{58}^{\mathcal{UR}}$	$B_{256}^{\mathcal{UR}}$	$B_{58}^{cg\mathcal{UR}}$	$B_{256}^{cg\mathcal{UR}}$	$B_{58}^{\mathcal{UCR}}$	$B_{256}^{\mathcal{UCR}}$	$B_{58}^{cg\mathcal{UCR}}$	$B_{256}^{cg\mathcal{UCR}}$
$\|B\|_F$	0.549	0.51	0.569	0.647	0.529	0.51	0.647	0.608
$\|B\|_2$	0.627	0.647	0.627	0.667	0.627	0.451	0.627	0.529
$\|B\|_{max}$	0.608	0.647	0.608	0.608	0.608	0.608	0.608	0.608
$\|B\|_\infty$	0.51	0.588	0.588	0.569	0.588	0.667	0.608	0.647
$\|B\|_1$	0.608	0.471	0.49	0.588	0.529	0.451	0.431	0.588
$rk(B)$	0.608	0.608	0.608	0.608	0.725	0.686	0.51	0.451

Table 10

The (U)LBP-based RQA results for Cediranib using the vector-based descriptors

Index	$VB_{58}^{\mathcal{UR}}$	$VB_{256}^{\mathcal{UR}}$	$VB_{58}^{cg\mathcal{UR}}$	$VB_{256}^{cg\mathcal{UR}}$	$VB_{58}^{\mathcal{UCR}}$	$VB_{256}^{\mathcal{UCR}}$	$VB_{58}^{cg\mathcal{UCR}}$	$VB_{256}^{cg\mathcal{UCR}}$
$\langle VB \rangle$	0.562	0.604	0.604	0.521	0.646	0.583	0.646	0.604
$Var(VB)$	0.729	**0.792**	0.542	0.729	0.583	0.5	0.625	**0.792**
$CV(VB)$	0.688	0.646	0.688	0.646	0.729	0.583	0.604	0.729
$Gi(VB)$	0.771	0.604	0.792	0.562	0.562	0.542	0.479	0.625
$I(VB)$	0.646	0.646	0.75	0.542	0.729	0.542	0.583	0.562
$\bar{I}(VB)$	0.562	0.542	0.667	0.5	0.5	0.562	0.688	0.458

Table 11

The (U)LBP-based RQA results for Cediranib using the matrix-based descriptors

Index	$B_{58}^{\mathcal{UR}}$	$B_{256}^{\mathcal{UR}}$	$B_{58}^{cg\mathcal{UR}}$	$B_{256}^{cg\mathcal{UR}}$	$B_{58}^{\mathcal{UCR}}$	$B_{256}^{\mathcal{UCR}}$	$B_{58}^{cg\mathcal{UCR}}$	$B_{256}^{cg\mathcal{UCR}}$
$\|B\|_F$	0.604	0.625	0.625	0.646	0.5	0.458	0.646	0.688
$\|B\|_2$	0.646	0.708	0.667	0.667	0.625	0.5	0.438	0.583
$\|B\|_{max}$	0.646	0.75	0.646	0.646	0.646	0.688	0.646	0.646
$\|B\|_\infty$	0.708	0.604	0.583	0.75	0.458	0.625	0.688	0.708
$\|B\|_1$	0.604	0.562	0.438	0.688	0.625	0.542	0.542	0.562
$rk(B)$	0.646	0.646	0.646	0.646	0.667	0.604	0.75	0.729

Table 12

The (U)LBP-based RQA results for Dasatinib using the vector-based descriptors

Index	$VB_{58}^{\mathcal{UR}}$	$VB_{256}^{\mathcal{UR}}$	$VB_{58}^{cg\mathcal{UR}}$	$VB_{256}^{cg\mathcal{UR}}$	$VB_{58}^{\mathcal{UCR}}$	$VB_{256}^{\mathcal{UCR}}$	$VB_{58}^{cg\mathcal{UCR}}$	$VB_{256}^{cg\mathcal{UCR}}$
$\langle VB\rangle$	0.604	0.688	0.646	0.458	0.729	0.667	0.562	0.542
$Var(VB)$	0.729	0.708	0.5	0.625	0.562	0.604	0.417	0.667
$CV(VB)$	0.562	0.688	0.604	0.562	0.604	0.667	0.438	0.75
$Gi(VB)$	0.646	0.562	0.479	0.5	0.521	0.562	0.438	0.479
$I(VB)$	0.667	0.688	0.667	0.583	0.646	0.521	0.521	0.583
$\bar{I}(VB)$	0.562	0.625	0.604	0.646	0.521	0.625	0.625	0.562

Table 13

The (U)LBP-based RQA results for Dasatinib using the matrix-based descriptors

Index	$B_{58}^{\mathcal{UR}}$	$B_{256}^{\mathcal{UR}}$	$B_{58}^{cg\mathcal{UR}}$	$B_{256}^{cg\mathcal{UR}}$	$B_{58}^{\mathcal{UCR}}$	$B_{256}^{\mathcal{UCR}}$	$B_{58}^{cg\mathcal{UCR}}$	$B_{256}^{cg\mathcal{UCR}}$
$\|B\|_F$	0.625	0.562	0.583	0.583	0.5	0.625	0.562	0.5
$\|B\|_2$	0.562	0.5	0.542	0.542	0.562	0.542	0.396	0.583
$\|B\|_{max}$	0.646	0.771	0.646	0.646	0.646	0.729	0.646	0.646
$\|B\|_\infty$	0.604	0.583	0.625	0.604	**0.854**	0.562	0.604	0.729
$\|B\|_1$	0.604	0.542	0.625	0.625	0.417	0.667	0.562	0.542
$rk(B)$	0.646	0.646	0.646	0.646	0.625	0.625	0.625	0.625

Table 14

The (U)LBP-based RQA results for Etoposide using the vector-based descriptors

Index	$VB_{58}^{\mathcal{UR}}$	$VB_{256}^{\mathcal{UR}}$	$VB_{58}^{cg\mathcal{UR}}$	$VB_{256}^{cg\mathcal{UR}}$	$VB_{58}^{\mathcal{UCR}}$	$VB_{256}^{\mathcal{UCR}}$	$VB_{58}^{cg\mathcal{UCR}}$	$VB_{256}^{cg\mathcal{UCR}}$
$\langle VB\rangle$	0.64	0.5	0.44	0.56	0.6	0.56	0.64	0.54
$Var(VB)$	0.62	0.66	0.56	0.56	0.52	0.52	0.46	0.58
$CV(VB)$	0.62	0.54	0.6	0.56	0.5	0.68	0.52	0.56
$Gi(VB)$	0.78	0.7	0.54	0.5	0.6	0.58	0.5	0.48
$I(VB)$	0.62	0.52	0.68	0.56	0.5	0.54	0.38	0.66
$\bar{I}(VB)$	0.42	0.4	0.64	0.58	0.52	0.7	0.54	0.48

Table 15

The (U)LBP-based RQA results for Etoposide using the matrix-based descriptors

Index	$B_{58}^{\mathcal{UR}}$	$B_{256}^{\mathcal{UR}}$	$B_{58}^{cg\mathcal{UR}}$	$B_{256}^{cg\mathcal{UR}}$	$B_{58}^{\mathcal{UCR}}$	$B_{256}^{\mathcal{UCR}}$	$B_{58}^{cg\mathcal{UCR}}$	$B_{256}^{cg\mathcal{UCR}}$
$\|B\|_F$	0.64	0.5	0.52	0.56	0.4	0.6	0.5	0.62
$\|B\|_2$	0.6	0.5	0.5	0.54	0.7	0.62	0.34	0.64
$\|B\|_{max}$	0.62	0.7	0.62	0.62	0.62	0.44	0.62	0.62
$\|B\|_\infty$	0.66	0.42	0.58	0.56	0.68	0.62	0.62	0.62
$\|B\|_1$	0.6	0.68	0.44	0.48	0.46	0.62	0.58	0.5
$rk(B)$	0.62	0.62	0.64	0.64	0.78	**0.8**	0.52	0.48

Table 16

The (U)LBP-based RQA results for Everolimus using the vector-based descriptors

Index	$VB_{58}^{\mathcal{UR}}$	$VB_{256}^{\mathcal{UR}}$	$VB_{58}^{cg\mathcal{UR}}$	$VB_{256}^{cg\mathcal{UR}}$	$VB_{58}^{\mathcal{UCR}}$	$VB_{256}^{\mathcal{UCR}}$	$VB_{58}^{cg\mathcal{UCR}}$	$VB_{256}^{cg\mathcal{UCR}}$
$\langle VB\rangle$	0.66	0.5	0.42	0.56	0.6	0.56	0.62	0.54
$Var(VB)$	0.62	0.66	0.56	0.56	0.52	0.52	0.46	0.58
$CV(VB)$	0.62	0.54	0.6	0.56	0.5	0.68	0.52	0.56
$Gi(VB)$	0.78	0.7	0.54	0.5	0.6	0.58	0.5	0.48
$I(VB)$	0.62	0.52	0.68	0.56	0.5	0.54	0.38	0.66
$\bar{I}(VB)$	0.42	0.4	0.64	0.58	0.52	0.7	0.54	0.48

Table 17

The (U)LBP-based RQA results for Everolimus using the matrix-based descriptors

Index	$B_{58}^{\mathcal{UR}}$	$B_{256}^{\mathcal{UR}}$	$B_{58}^{cg\mathcal{UR}}$	$B_{256}^{cg\mathcal{UR}}$	$B_{58}^{\mathcal{UCR}}$	$B_{256}^{\mathcal{UCR}}$	$B_{58}^{cg\mathcal{UCR}}$	$B_{256}^{cg\mathcal{UCR}}$
$\|B\|_F$	0.64	0.5	0.52	0.56	0.4	0.6	0.5	0.62
$\|B\|_2$	0.6	0.5	0.5	0.54	0.7	0.62	0.34	0.64
$\|B\|_{max}$	0.62	0.7	0.62	0.62	0.62	0.44	0.62	0.62
$\|B\|_\infty$	0.62	0.38	0.54	0.5	0.66	0.6	0.6	0.62
$\|B\|_1$	0.6	0.72	0.54	0.48	0.42	0.64	0.56	0.54
$rk(B)$	0.62	0.62	0.64	0.64	0.8	**0.82**	0.56	0.52

Table 18

The (U)LBP-based RQA results for Ipatasertib using the vector-based descriptors

Index	$VB_{58}^{\mathcal{UR}}$	$VB_{256}^{\mathcal{UR}}$	$VB_{58}^{cg\mathcal{UR}}$	$VB_{256}^{cg\mathcal{UR}}$	$VB_{58}^{\mathcal{UCR}}$	$VB_{256}^{\mathcal{UCR}}$	$VB_{58}^{cg\mathcal{UCR}}$	$VB_{256}^{cg\mathcal{UCR}}$
$\langle VB\rangle$	0.449	0.367	0.592	0.592	0.612	0.633	0.714	0.51
$Var(VB)$	0.592	0.531	0.408	0.633	0.571	0.653	0.449	0.612
$CV(VB)$	0.571	0.429	0.571	0.612	0.347	0.531	0.469	0.592
$Gi(VB)$	0.612	0.49	0.449	0.612	0.592	0.571	0.571	0.551
$I(VB)$	0.551	0.429	0.449	0.49	0.51	0.49	0.388	0.571
$\bar{I}(VB)$	0.551	0.551	0.612	0.571	0.429	0.531	0.469	0.449

Table 19

The (U)LBP-based RQA results for Ipatasertib using the matrix-based descriptors

Index	$B_{58}^{\mathcal{UR}}$	$B_{256}^{\mathcal{UR}}$	$B_{58}^{cg\mathcal{UR}}$	$B_{256}^{cg\mathcal{UR}}$	$B_{58}^{\mathcal{UCR}}$	$B_{256}^{\mathcal{UCR}}$	$B_{58}^{cg\mathcal{UCR}}$	$B_{256}^{cg\mathcal{UCR}}$
$\|B\|_F$	0.429	0.408	0.633	0.653	0.49	0.714	0.531	0.714
$\|B\|_2$	0.531	0.429	0.551	0.449	0.449	0.531	0.429	0.571
$\|B\|_{max}$	0.633	**0.735**	0.633	0.633	0.633	0.673	0.633	0.633
$\|B\|_\infty$	0.551	0.551	0.612	0.551	0.551	0.653	0.592	0.633
$\|B\|_1$	0.633	0.612	0.49	0.551	0.449	0.449	0.673	0.653
$rk(B)$	0.633	0.633	0.633	0.633	0.531	0.551	0.571	0.49

Table 20

The (U)LBP-based RQA results for ECG200 using the vector-type descriptors

Index	$VB_{58}^{\mathcal{UR}}$	$VB_{256}^{\mathcal{UR}}$	$VB_{58}^{cg\mathcal{UR}}$	$VB_{256}^{cg\mathcal{UR}}$	$VB_{58}^{\mathcal{UCR}}$	$VB_{256}^{\mathcal{UCR}}$	$VB_{58}^{cg\mathcal{UCR}}$	$VB_{256}^{cg\mathcal{UCR}}$
$\langle VB \rangle$	0.565	0.59	0.61	0.645	0.61	0.52	0.54	0.61
$Var(VB)$	0.605	0.615	0.61	0.66	**0.74**	0.635	0.63	0.605
$CV(VB)$	0.62	0.64	0.6	0.61	0.6	0.58	0.62	0.64
$Gi(VB)$	0.69	0.57	0.58	0.635	0.58	0.61	0.6	0.73
$I(VB)$	0.655	0.675	0.59	0.595	0.61	0.635	0.585	0.64
$\bar{I}(VB)$	0.655	0.6	0.59	0.625	0.62	0.55	0.54	0.585

Table 21

The (U)LBP-based RQA results for ECG200 using the matrix-type descriptors

Index	$B_{58}^{\mathcal{UR}}$	$B_{256}^{\mathcal{UR}}$	$B_{58}^{cg\mathcal{UR}}$	$B_{256}^{cg\mathcal{UR}}$	$B_{58}^{\mathcal{UCR}}$	$B_{256}^{\mathcal{UCR}}$	$B_{58}^{cg\mathcal{UCR}}$	$B_{256}^{cg\mathcal{UCR}}$
$\|B\|_F$	0.62	0.51	0.55	0.595	0.675	0.585	0.515	0.605
$\|B\|_2$	0.61	0.57	0.59	0.61	0.645	0.535	0.62	0.645
$\|B\|_{max}$	0.665	0.665	0.665	0.665	0.665	0.665	0.665	0.665
$\|B\|_\infty$	0.72	0.59	0.665	0.56	0.645	0.58	0.605	0.61
$\|B\|_1$	0.6	0.615	0.54	0.62	0.585	0.625	0.535	0.535
$rk(B)$	0.66	0.66	0.66	0.66	0.64	0.67	0.68	0.68

Table 22

The parameters of the traditional RQA algorithm and the improvement attained by the novel (U)LBP descriptors

Time series	Radius	Mean RR coefficient	Improvement
Abemaciclib	0.07	0.1	6.7 %
Alpelisib	0.13	0.1	5.82 %
Bleomycin	0.13	0.11	15.42 %
Cediranib	0.11	0.1	7.95 %
Dasatinib	0.12	0.1	9.72 %
Etoposide	0.11	0.1	10 %
Everolimus	0.11	0.1	12.2 %
Ipatasertib	0.13	0.11	16.73 %
ECG200	0.11	0.11	9.46 %

From the above results, it follows that the newly proposed ULBP-based and LBP-based complexity descriptors are more or significantly more efficient in the classification tasks of pharmacokinetical and physiological data than their original counterparts. This regularity is observed for the relatively short dose-response time series as well as for the longer ECG time series. Specifically, in our numerical experiments, the improvement achieved by the newly suggested (U)LBP-based indices ranges from 5.82% to 16.73%.

In the second measurement, we perform the *principal component analysis* (PCA) [5] on the ECG200 dataset in order to visualize the mutual relationships between the tested complexity descriptors. In Fig. 6.1 and 6.2, the analyzed indices are labeled as follows: 1 – RR, 2 – $RATIO$, 3 – DET, 4 – DIV, 5 – L_{max}, 6 – L_{mean}, 7 – $\bar{L}_{mean}$, 8 – $ENTR$, 9 – $TREND$, 10 – LAM, 11 – V_{max}, 12 – V_{mean}, 13 – $\langle VB\rangle$, 14 – $Var(VB)$, 15 – $CV(VB)$, 16 – $Gi(VB)$, 17 – $I(VB)$, 18 – $\bar{I}(VB)$, 19 – $\|B\|_F$, 20 – $\|B\|_2$, 21 – $\|B\|_{max}$, 22 – $\|B\|_\infty$, 23 – $\|B\|_1$, 24 – $rk(B)$ where the instantiations of the VB vector and the B matrix are listed by the conditions (23) and (24), respectively. Moreover, in both .s, the complexity descriptors derived from the coarse-grained derivative arrays are labeled as cgULBP or as cgLBP.

From eight PCA projections presented in Fig. 1 and 2, it follows that all standard RQA complexity measures are concentrated almost in one point. Also, the PCA projections from Fig. 1 demonstrate that four matrix-based descriptors, i.e., $\|B\|_F$, $\|B\|_2$, $\|B\|_\infty$ and $\|B\|_1$ as well as one vector-based index, i.e., $Var(VB)$ carry very specific information. Moreover, from four two-dimensional projections presented in Fig. 2, it can be seen that two vector-based central tendency metrics, i.e., $\langle VB\rangle$ and $Var(VB)$ as well as two vector-based information-theoretical metrics, i.e., $Gi(VB)$ and $I(VB)$ together with one matrix-based coefficient, i.e., $\|B\|_2$ do not duplicate the structural information contained in the original RQA metrics.

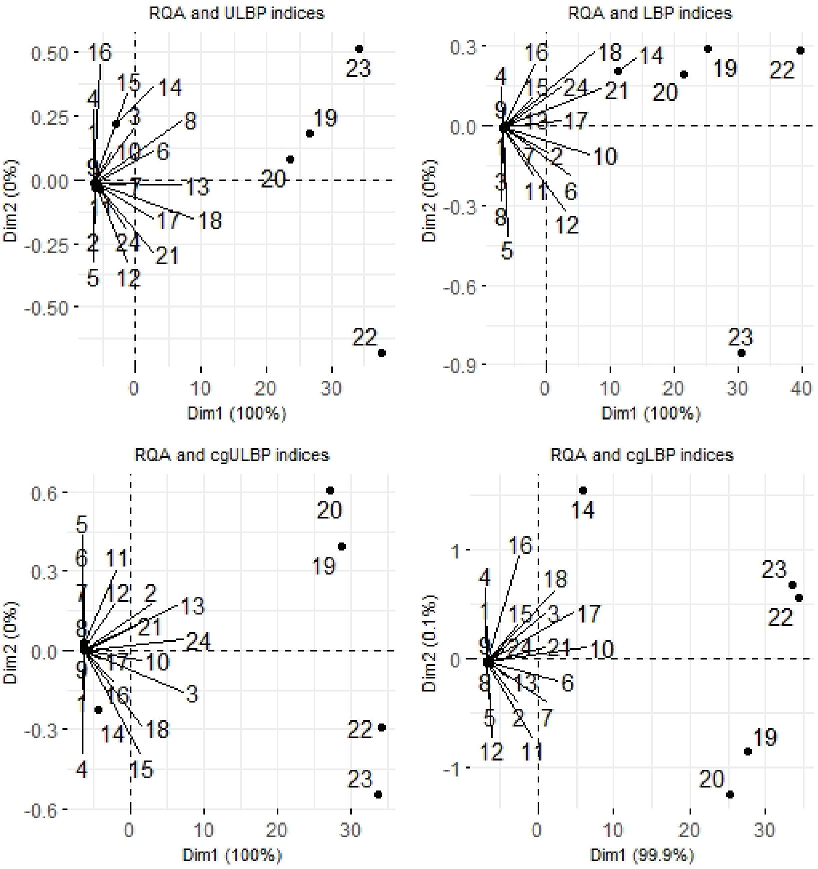


Fig. 1. The PCA projection results of the novel (U)LBP-based descriptors derived from the matrix (21)

Rys. 1. Wyniki analizy głównych składowych nowych (U)LBP współczynników otrzymanych z macierzy nr 21

Accordingly, it can be concluded that the collection of the newly suggested (U)LBP-based RQA complexity descriptors is overwhelmingly more informative than the set of the original RQA indices. In our opinion, the above phenomena can (to some extent) elucidate the enhanced performance of the novel (U)LBP complexity descriptors in comparison with the standard RQA metrics.

10.6. Conclusions

Recent pharmacological and biochemical technologies have permitted scholars to collect a large number of pharmacokinetical data in the medical domain. For this reason, the development of reliable methods for automated analysis of pharmacologically and physiologically important time series extracted from medical databases is a very important research topic. Moreover, in the recent years, several authors tried to use the LBP methodology to study medical images [11, 17, 20]. The present contribution proposes to apply the local binary pattern algorithm as well as its uniform variant to analyze the physiologically and pharmacologically significant time series. Our preliminary results demonstrated that the LBP and ULBP methodologies applied to the unthresholded recurrence and cross-recurrence matrices are very promising as sensitive metrics enabling to characterize the structural patterns in the medical data under consideration.

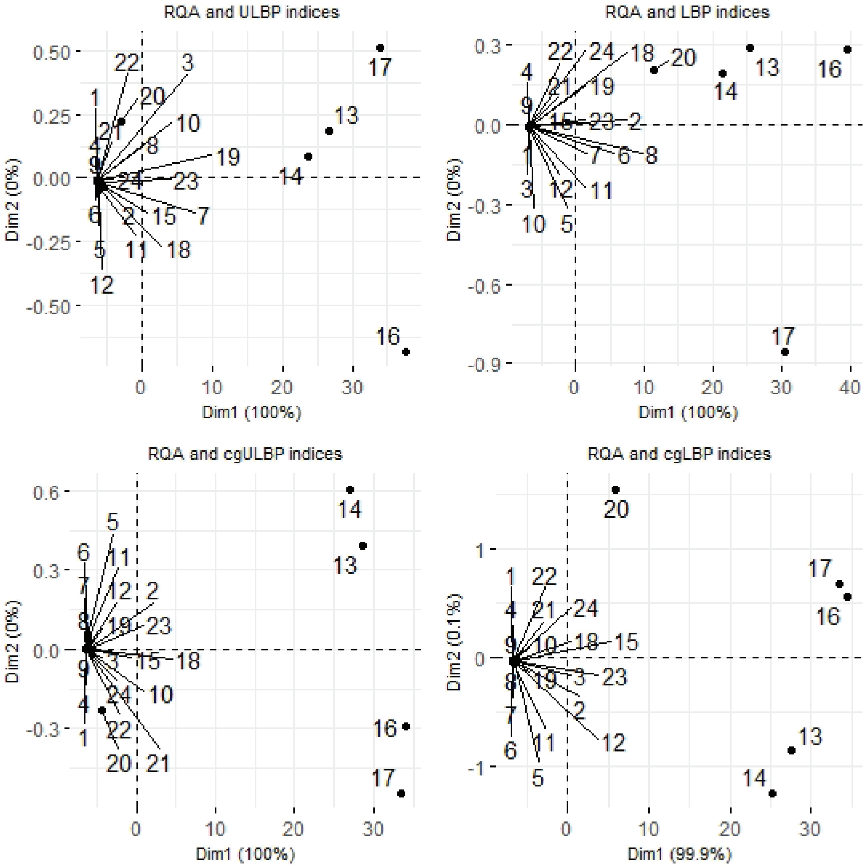


Fig. 2. The PCA projection results of the novel (U)LBP-based descriptors derived from the matrix (22)

Rys. 2. Wyniki analizy głównych składowych nowych (U)LBP współczynników otrzymanych z macierzy nr 22

Therefore, it can be conjectured that not only medical images but also medical numerical data can be analyzed by the (U)LBP techniques.

Bibliography

1. S. Brahnam, L.C. Jain, L. Nanni, A. Lumini: Local Binary Patterns: New Variants and Applications, Springer – Verlag, Berlin, Heidelberg 2014.
2. J.-P. Eckmann, S.O. Kamphorst, D. Ruelle: Recurrence plots of dynamical systems, Europhys. Lett. (1987) **5**: 973–977.

3. D. Eroglu, T.K. DM. Peron, N. Marwan, F.A. Rodrigues, L. da F. Costa, M. Sebek, I.Z. Kiss, J. Kurths: Entropy of weighted recurrence plots, Phys. Rev. E (2014) **90**: 042919.
4. C.A. Garcia, nonlinearTseries: Nonlinear Time Series Analysis. R package version 0.2.11. (2021), https://CRAN.R-project.org/package=nonlinearTseries.
5. R.A. Johnson, D.W. Wichern: Applied Multivariate Statistical Analysis, Pearson Education, Inc., Upper Saddle River 2007.
6. A. Kassambara, ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0 (2020), https://CRAN.R-project.org/package=ggpubr
7. A. Kassambara, F. Mundt, factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.7. (2020), https://CRAN. R- project. org/package=factoextra
8. E.J. Leavline, D.A.A.G. Singh, P. Maheswari: Local binary pattern family descriptors for texture classification, IJIGSP (2018) **10**: 40–45.
9. A.L.R. Lubbock, L.A. Harris, V. Quaranta, D.R. Tyson, C.F. Lopez, Thunor: visualization and analysis of high-throughput dose-response datasets, Nucleic Acid Res. (2021) **49**: W633-W640.
10. N. Marwan, M.C. Romano, M. Thiel, J. Kurths: Recurrence plots for the analysis of complex systems, Phys. Rep. (2007) **438**: 237–329.
11. L. Nanni, A. Lumini, S. Brahnam: Local binary patterns variants as texture descriptors for medical image analysis, Artif. Intell. Med. (2010) **49**: 117–125.
12. T. Ojala, M. Pietikäinen, D. Harwood: Performance evaluation of texture measures with classification based on Kullback discrimination of distributions, Proceedings of the 12th IAPR International Conference on Pattern Recognition vol. 1 (1994): 582–585.
13. R Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2020), https://www.R-project.org
14. K. Slowikowski, ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.8.2. (2020), https://CRAN.R-project.org/package=ggrepel
15. J. Sugiyama, K. Kobayashi, wvtool: Image Tools for Automated Wood Identification. R package version 1.0. (2016), https://CRAN.R-project.org/package=wvtool
16. R. Todeschini, V. Consonni, *Molecular Descriptors for Chemoinformatics, Volume I & II*, Wiley – VCH Verlag GmbH & Co. KGaA, Weinheim 2009.

17. O.A. Vătămanu, M. Ionescu, G.-I. Mihalaş: Analysis and classification of ultrasound medical images using the Local Binary Pattern operator, Stud. Health Technol. Inform. (2013) **190**:175–178.
18. H. Wickham, ggplot2: Elegant Graphics for Data Analysis, Springer – Verlag, New York 2016, https://ggplot2.tidyverse.org
19. P. Wilczek: Novel centrality measures and distance-related topological indices in network data mining, Silesian J. Pure Appl. Math. (2017) **7**, 21–63.
20. X. Xianchuan, Q. Zhang: Medical image retrieval using local binary patterns with image Euclidean distance, 2009 International Conference on Information Engineering and ComputerScience (2009): 1–4.
21. J.P. Zbilut, J.-M. Zaldívar-Comenges, F. Strozzi: Recurrence quantification based Liapunov exponents for monitoring divergence in experimental data, Phys. Lett. A (2002) **297** (3–4): 173–181.

AN APPLICATION OF THE LOCAL BINARY PATTERN ALGORITHM AND ITS UNIFORM VARIANT TO IMPROVE THE RECURRENCE AND CROSS-RECURRENCE QUANTIFICATION ANALYSES OF THE PHARMACOLOGICALLY AND PHYSIOLOGICALLY IMPORTANT TIME SERIES

Abstract

The *recurrence* and *cross-recurrence matrices* are recognized as a very efficient tool allowing to visualize the recurrences of nonlinear dynamical systems. The complexity of the structural patterns detected in these matrices is quantified by the RQA descriptors. The purpose of the present work was to introduce several novel RQA complexity descriptors based on the *local binary pattern* (LBP) operator and its *uniform* version (ULBP). These operators label the pixels of an image by thresholding the neighborhood of each pixel and report the results as binary numbers. In our approach, we consider the (*coarse-grained*) *unthresholded* recurrence and cross-recurrence matrices as images on which the LBP and ULBP operators can be defined. The usefulness of the newly suggested RQA complexity descriptors in the *time series classification tasks* is tested on the selected 393 time-dependent cell proliferation measurements extracted from the (publicly available) HTS007 dataset as well as on one (also publicly available) ECG dataset containing 200 time series. In all exemplary cases, our results showed that, in comparison to the traditional RQA indices, the classification accuracy attained by the novel (U)LBP-based RQA descriptors is higher or significantly higher. Moreover, by using the *multivariate technique* (PCA), we demonstrated that some of the newly proposed complexity descriptors carry very specific structural information which can not be obtained by means of the original RQA procedure. Accordingly, it can be conjectured that, in the realm of pharmacokinetics and physiology, the novel (U)LPB-based RQA complexity indices will be regarded as a valuable statistical tool in analyzing nonlinear time series.

Keywords: nonlinear time series, pharmacokinetics, RQA methodology, LBP and ULBP algorithms

DOI:10.34918/85111

Jarosław GIL[1,*], Krzysztof SZYMICZEK[1], Andrzej POLAŃSKI[1]

Chapter 11. IMPROVING GILLESPIE SIMULATION ALGORITHM FOR FITNESS IN CLONAL EVOLUTION

11.1. Introduction

Neoplastic transformations in human tissues are consequences of accumulating somatic, clonal mutations. In the ongoing research on cancers, observations on occurrences of somatic mutations are collected and then their roles in neoplasm are explained in biological and physiological terms. A strong impulse in studies on somatic mutations in cancers are provided by large, experimental projects of DNA and RNA sequencing of cancer tissues leading to creation of large databases of cancer mutations, such as TCGA project, TCGA database [1] and COSMIC database [2]. These researches led to significant advance in understanding cancer development as well as improving tools for diagnosis and therapy.

Large volumes of data and its detailedness encourages elaborating mathematical models, which would correspond to scenarios of cancer initiation and development. Mathematical modelling of tumor growth is based on probabilistic description of events seen in the neoplastic processes, cellular replications and deaths and occurrences of somatic mutations. Mathematical models most often used are Markov birth - death processes, branching processes or multitype branching processes [3–5].

Variety of possible, potentially complicated laws for probability distributions of events in mathematical models of neoplastic transformations motivate for developing stochastic simulations algorithms. Numerous papers devoted to simulation systems of evolution of cancer cell populations already appeared in the literature, e.g., [6–8]. The basic approach for simulation of events occurring in neoplastic processes is by using

---

[1] Silesian University of Technology, Poland.

[*] Corresponding author: jaroslaw.gil@polsl.pl.

Gillespie algorithm [9], i.e., the algorithm of successive updates of processes state vector and reactions / events propensities on the basis of simulated times of reactions / events. Application of Gillespie's algorithm to clonal evolution of populations was described e.g., in [4] and [10]. In [4] a very simple version of Gillespie's algorithm for simulating clonal evolution of cancer cells was used, with the state vector containing two components. However, it enabled simulating evolutionary effects of mildly deleterious passenger mutations leading to shrinking population size versus strongly advantageous driver mutations causing selective sweeps. In [10] a high-resolution simulation tool was presented with high dimensional state vector and mutations with different effects on fitness was presented. It enabled observing various scenarios of cancer clonal evolution but required substantial amount of computational power.

In this study we present Gillespie simulation algorithm for generating evolutionary events in developing cancer cells populations, with ability of simulating growth of cancer cells population detailed enough for observing moving wave of fitness of cancer cells. Mutations in our model have equal, mildly advantageous fitness effects. We pay attention to efficiency of simulation, aiming to achieving populations of sizes of millions of cells on desktop computer. We compare efficiency of two algorithms, one with the state vector containing components corresponding to each of the cells of the population, and another with smaller state vector built of bins corresponding to groups of cells with equal number of mutations.

11.2. Model description

The idea of the Gillespie algorithm provides methodology to simulate population evolution analyzing it cell by cell. For small population sizes that algorithm is very good but for large cell number it causes long simulation time. Improvement can make simulation faster with keeping results of calculations the same or similar.

11.2.1. Possible events in clonal evolution

Genetic forces behind clonal evolution are replications, mutations, selection and genetic drift. Mainly can be highlighted cell death, division and mutation. In Fig. 1 were presented phenomena simulated by Gillespie algorithm.

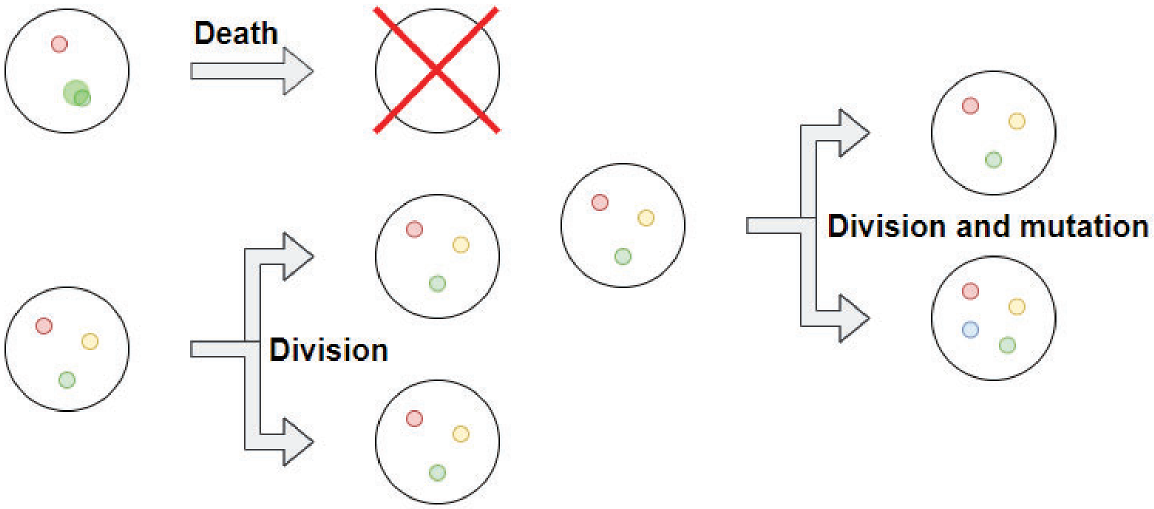


Fig. 1. Events analyzed in model
Rys. 1. Zdarzenia analizowane w modelu

Cell death is dependent on population capacity. If number of cells is greater than population capacity death probability coefficient should be larger. Equation for intensity of cell death process has the following form:

$$\mu_D = \frac{N}{K} \tag{1}$$

where N denotes population size and K stands for environment capacity.

Cell mutation can occur while cell is dividing. Cell division is dependent on fitness coefficient which increase while cell is mutating. Mutation can cast positive, neutral or negative effect on fitness coefficient. In clonal evolution process cell fitness should be mostly increasing to provide fast population growth. For simulation purposes mutation effect is assumed as positive value. Intensity of cellular divisions/births is different for each cell and is given by

$$\mu_B(l) = (1+s)^l \tag{2}$$

where s is the fitness effect of a single mutation and l is the number of mutations accumulated in the cell. Probability of occurrence of a mutation is denoted by p so the intensity of mutations process is

$$\mu_p(l) = p(1+s)^l. \tag{3}$$

11.2.2. Gillespie algorithm

The first approach involves simulating evolution process cell by cell treating cell division or death as two reactions and using first reaction or next reaction version of Gillespie's algorithm [11]. After analysis, the time value is subtracted from all other cells and for updated cell new death or division time is generated as random variable with exponential distribution. Event kind depends on which time variable is the smallest. For death cell is simply erased from population but while division cell mutation probability is checked. If cell mutates clone with one more mutation is added to population. Fig. 2 contains block diagram of that algorithm.

Finding the smallest time variable in population needs to compare all cells to each other. In worst case computational complexity of algorithm is equal to O((2n)!). In every loop one variable is compared to every other time variable in whole population composed of n cells. For small population this complexity is not a problem but for large initial size simulation time will be very long.

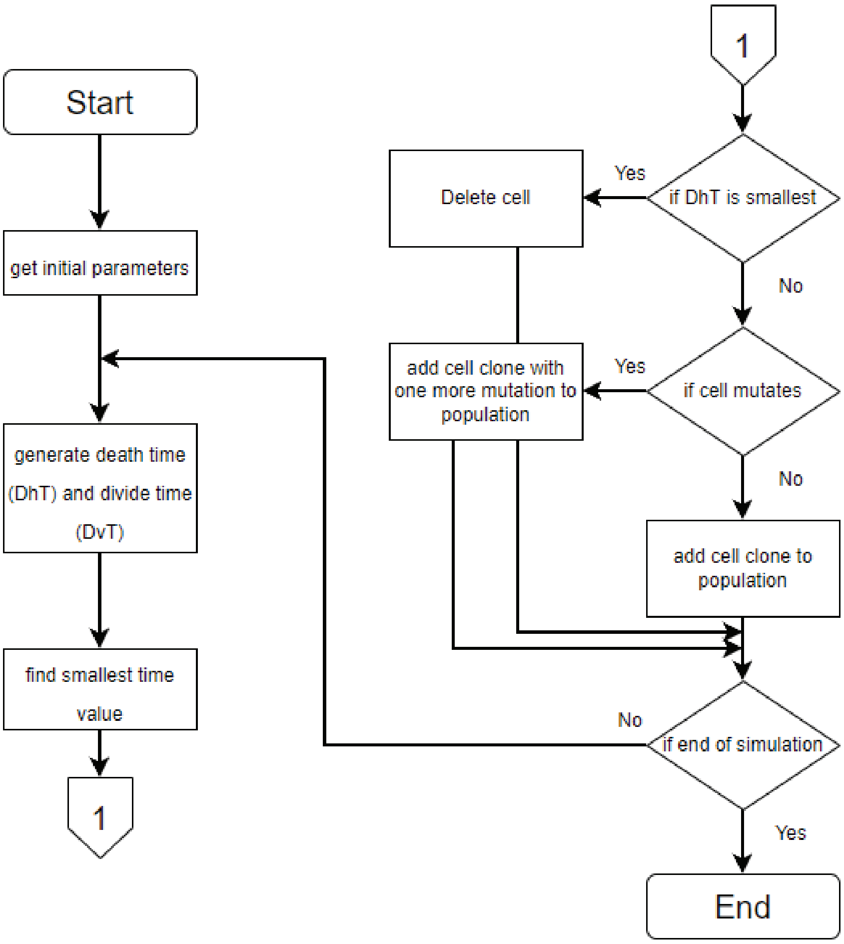


Fig. 2. Original Gillespie algorithm block diagram
Rys. 2. Schemat blokowy podstawowego algorytmu Gillespiego

11.2.3. Tau leap algorithm

To simplify algorithm complexity in one simulation cycle can be analyzed more cells. Specifying tau value – time step, limits number of iterations to population size. Fig. 3 contains block schema for tau loop algorithm. After time generation for all cells two comparisons are made – if death time or divide time is smaller than tau. First phenomena kind is determined by smaller number for one cell. Algorithm idea is the same as that of the original approach.

To update all cells analyzed in one cycle it is needed to compare both times to tau. For one cell two comparisons are done so the algorithm complexity can be described as O(2n). That mean simulation time is linear dependent on population size and one cycle time is smaller than in original approach.

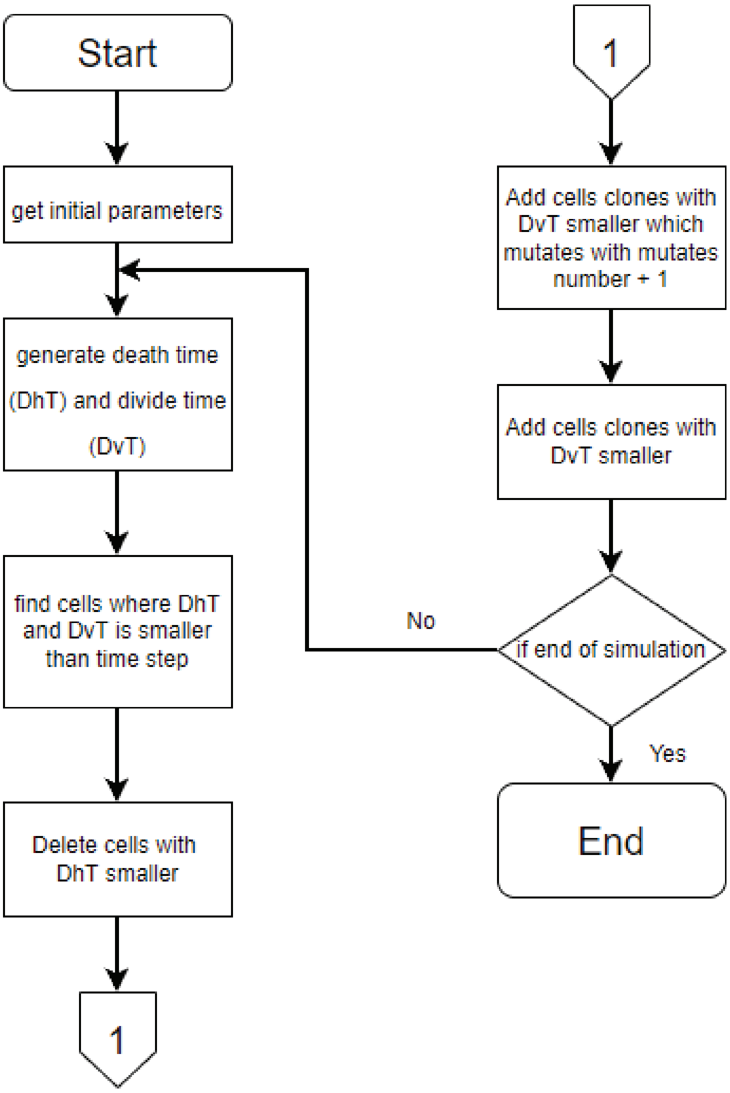


Fig. 3. "Tau loop" algorithm block diagram
Rys. 3. Schemat blokowy algorytmu „tau loop"

11.2.4. Binned Gillespie algorithm

Still lowering original Gillespie algorithm cause loss of data accuracy. More assumptions are needed to take which can provide false simulation data. To simplify Gillespie algorithm, we propose its binned version. For large initial population basic algorithm is very slow cause of iteration through whole population. Every cell can be characterized by mutation number which can provide method for cell grouping. In one cycle than would be much less groups than cells in population.

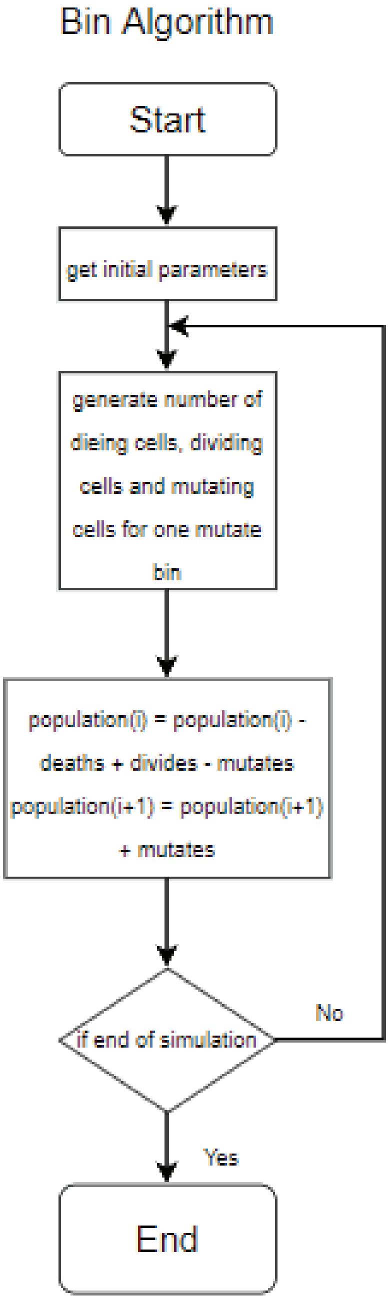


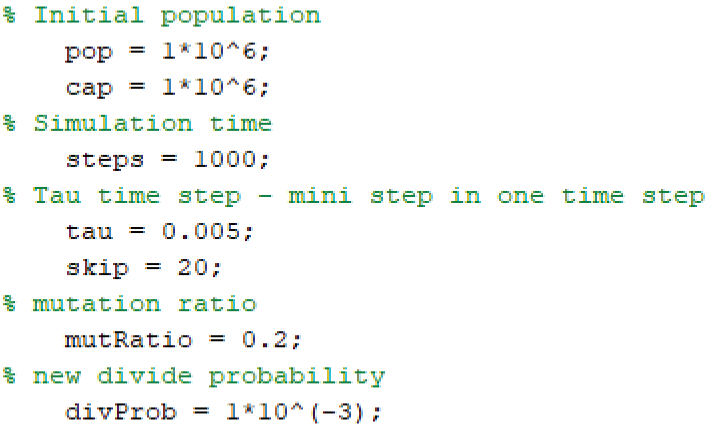
Fig. 4. Binned algorithm block diagram
Rys. 4. Schemat blokowy grupowanego algorytmu

Fig. 4 contains block schema for binned Gillespie algorithm. In each simulation cycle all mutation groups are updated by generating random probabilities based on population parameters. Death probability depends on population size and population capacity, division probability depends on fitness factor of each group. For cell mutation constant probability is assumed – part of dividing cells is mutating.

The complexity of that approach is equal to O(n) where n is interpreted as number of bins. For large population size number of mutation groups is much smaller then number of cells so simulation time is also smaller than cell-based algorithm.

11.2.5. Simulation parameters

To properly simulate population evolution few parameters are taken in consideration. Fig. 5 contains model initial parameters which are base for calculating events probability.



```
% Initial population
    pop = 1*10^6;
    cap = 1*10^6;
% Simulation time
    steps = 1000;
% Tau time step - mini step in one time step
    tau = 0.005;
    skip = 20;
% mutation ratio
    mutRatio = 0.2;
% new divide probability
    divProb = 1*10^(-3);
```

Fig. 5. Model initial parameters
Rys. 5. Początkowe parametry modelu

Death probability is calculated based on population size and population capacity. Divide probability depends on bin fitness. Mutation probability is assumed as constant value describing mutation-division ratio. In each simulation cycle only, events for cells with death/division time smaller than tau are occurring. In large population tau value can be interpreted as probability that event for cell will occurs. In simulation every calculated parameter is multiplied by tau. That describes how many cells should be updated.

11.3. Results and discussion

Introduced algorithm modification reduce algorithm complexity and simulation time. To analyze impact on data accuracy few experiments were made.

11.3.1. Algorithms comparison

To prove that binned version of Gillespie algorithm result accuracy does not differ from original and tau loop version comparison of results after 10 cycles is presented. Experiment was performed multiple times, for all attempts results were as described. Fig. 6 shows result of "tau loop" algorithm and binned algorithm. Simulation parameters assumed in that experiment are shown on Fig. 5.

Simulation time for one cycle is eight times greater for the "tau loop" approach. The complexity of binned version is better. Distribution of mutated cells are very similar. Can be assumed the differences between both results are neglectable and could occurs because of randomness in coefficient generation.

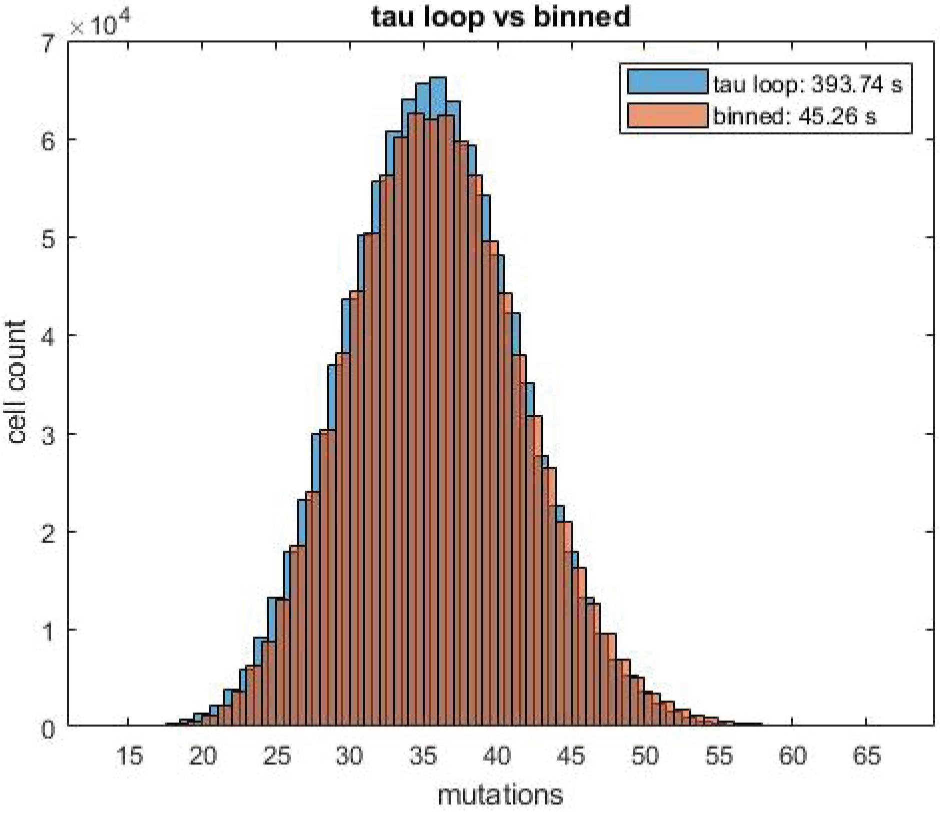


Fig. 6. "Tau loop" and binned algorithm comparison after 10 cycles of simulation
Rys. 6. Porównanie wyników algorytmów "tau loop" i zbinowanego po 10 cyklach symulacji

11.3.2. Mutation wave

Accumulation of somatic mutations in cancer clone can be seen a traveling wave of increasing numbers of mutations in cancer cells [12]. The mutation wave obtained in our simulations is presented on Fig. 6 for “tau loop” algorithm and binned algorithm. Clonal evolution can be described as chaotic process with causes very fast population growth and cell mutations. To analyze that phenomena experiments with simulation parameters were made. Fig. 7, Fig. 9 and Fig. 11 shows mutation wave speed while capacity, mutation/division ratio and fitness modifier were change.

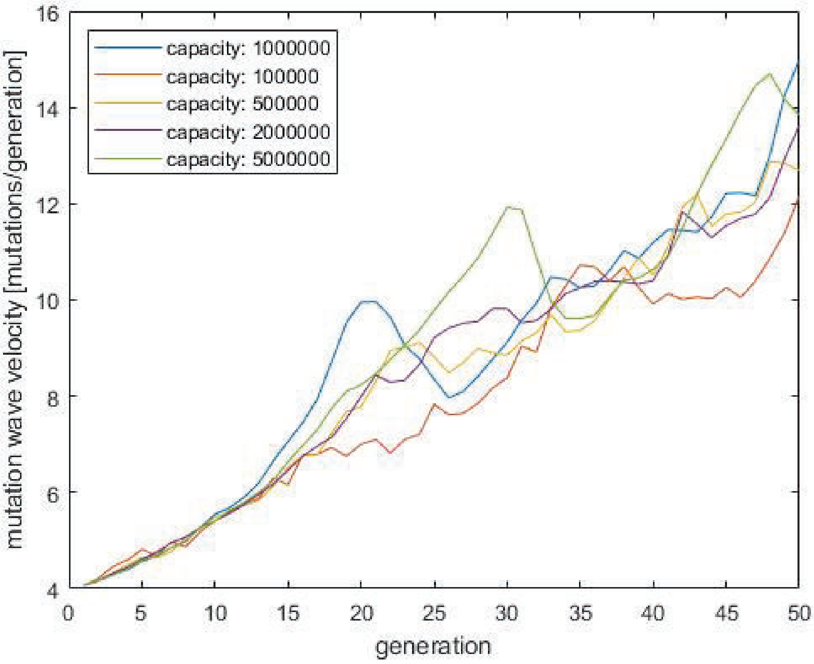


Fig. 7. Mutation wave speed versus population capacity
Rys. 7. Prędkość fali mutacji względem pojemności populacji

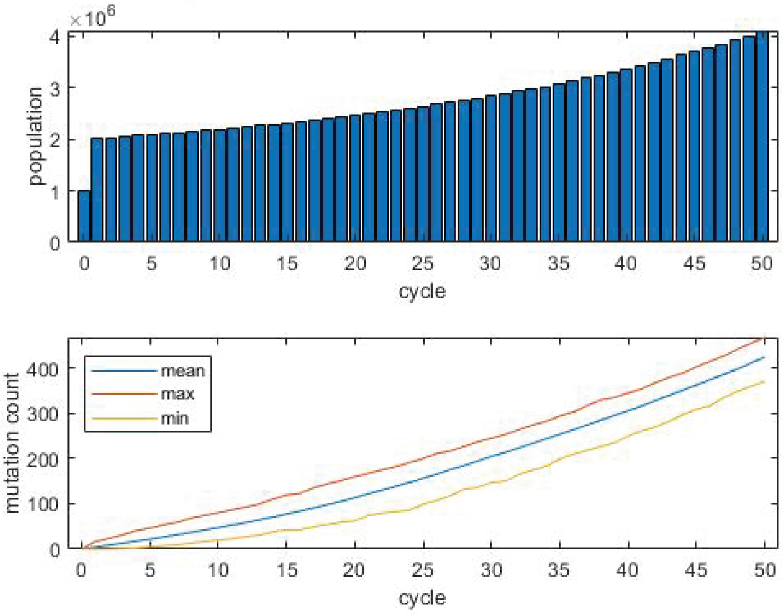


Fig. 8. Population statistics across 50 generations for capacity equal 2000000
Rys. 8. Statystyka populacji przez 50 generacji dla pojemności 2000000

Population capacity seems to have no impact on mutation wave. Independently on its value mutation wave velocity rise with number of generations. When cells population mutates, fitness factor changes in positive way. The cells are mutating and dividing more spontaneously providing wave velocity rise and population growth. Fig. 8 shows population statistics for capacity two times larger than initial population. The number of cells at the very beginning doubles and still rises. Mutation number also rises – at simulation beginning slowly then faster.

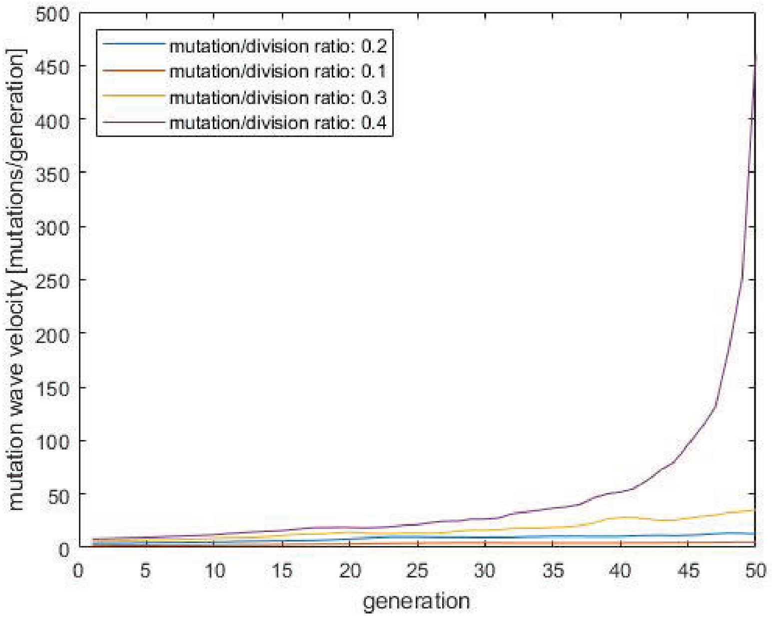


Fig. 9. Mutation wave velocity versus mutation/division ratio
Rys. 9. Prędkość fali mutacji względem współczynnika mutacji

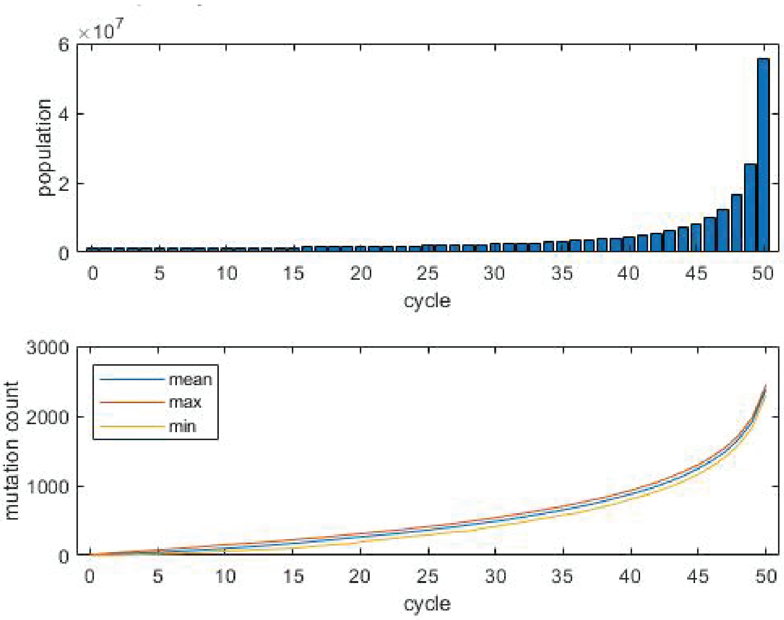


Fig. 10. Population statistics across 50 generations for mutation/division ratio equal 0.4
Rys. 10. Statystyka populacji przez 50 generacji dla współczynnika mutacji 0.4

Mutation-division ratio can be interpreted as chance for cell to mutate while dividing. Its value has the highest impact on mutation probability what is shown on Fig. 9. For small ratio values wave velocity rise slowly. Small rose of ratio causes enormous change in wave speed and population growth. High velocity of mutation wave causes many mutations in cells what is shown on Fig. 10. Also, population size is growing rapidly.

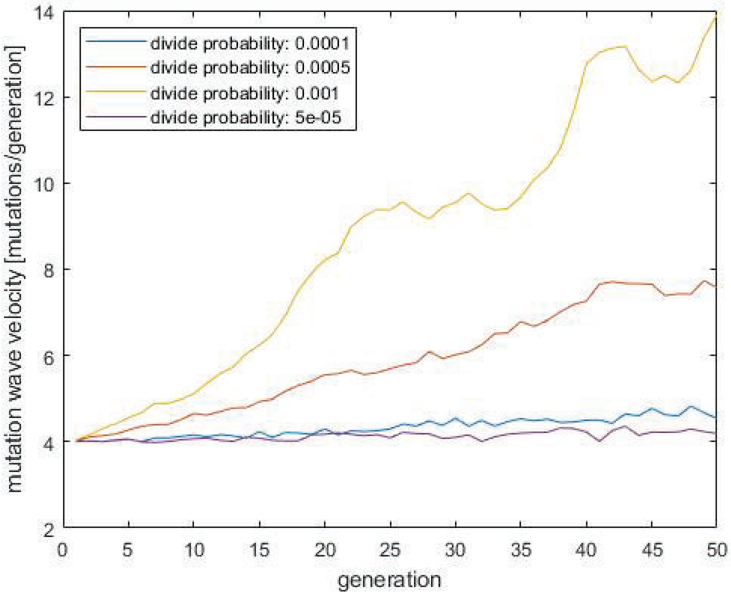


Fig. 11. Mutation wave velocity versus fitness factor modifier
Rys. 11. Prędkość fali mutacji względem modyfikatora współczynnika dopasowania

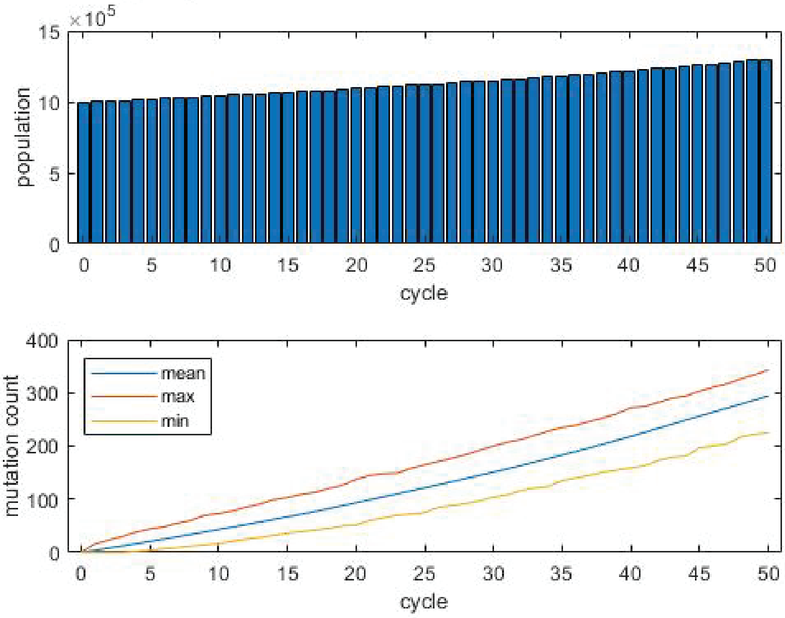


Fig. 12. Population statistics across 50 generations for fitness modifier equals 0.0005
Rys. 12. Statystyka populacji przez 50 generacji dla modyfikatora dopasowania 0.0005

In Fig. 11 is presented mutation wave velocity while changing division probability modifier – fitness factor modifier. Every mutation has impact on cell division probability. In experiment was assumed only positive effect. When its value rise, wave velocity also rises. For small values wave velocity seems to stay constant. Population size, as have been shown on Fig. 12, rises slowly for small modifier value. Mutation number changes nearly linear.

11.3.3. Fitness wave

The moving mutation wave can also be interpreted as the wave of fitness moving in the population of cancer cells. Fitness factor provide information about population adaptation. If its value is higher cell division probability also is higher. Fig. 13, Fig. 14 and Fig. 15 contains fitness waves of few generations dependent on capacity, mutation/division ratio and fitness factor modifier. Data present on these figures complements information from 0.

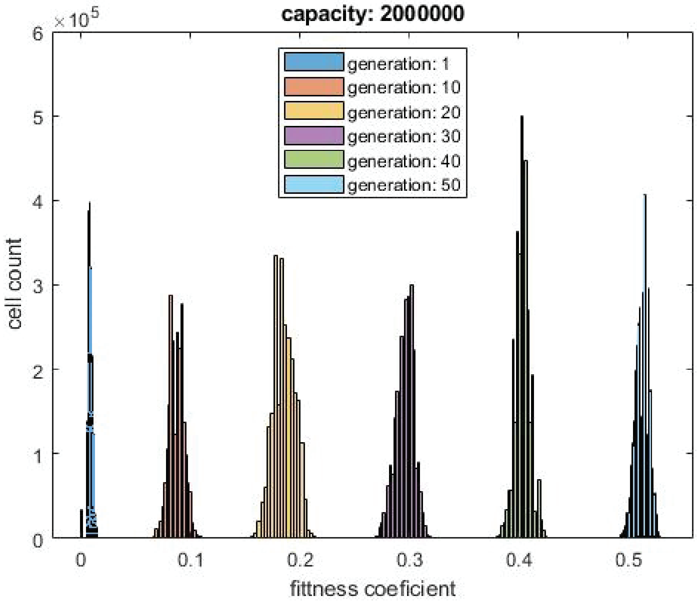


Fig. 13. Fitness wave for population capacity 2000000
Rys. 13. Fala dopasowania dla pojemności populacji 2000000

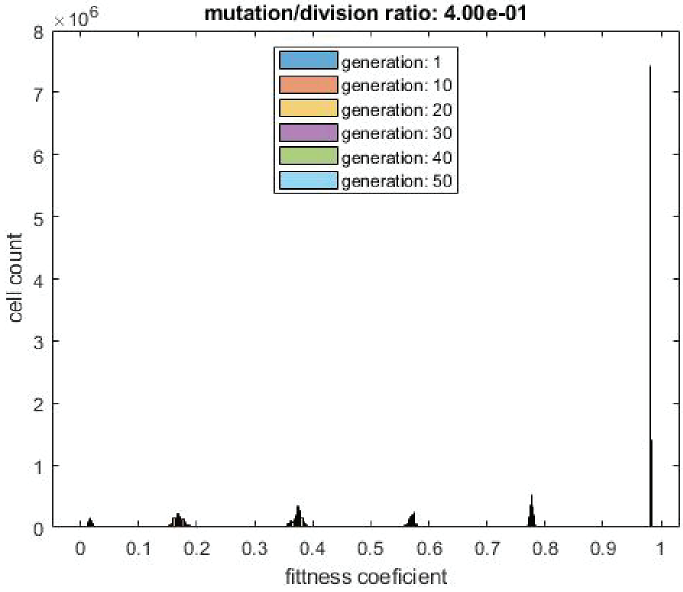


Fig. 14. Fitness wave for mutation/division ratio 0.4
Rys. 14. Fala dopasowania dla współczynnika mutacji 0.4

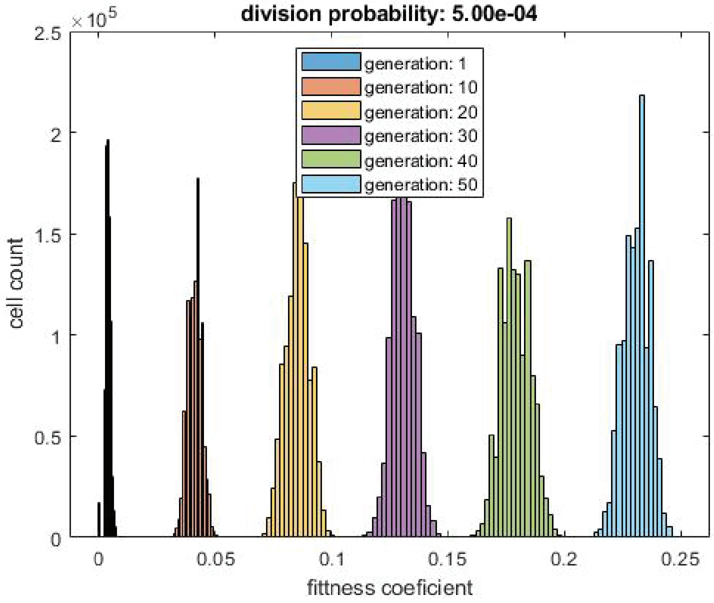


Fig. 15. Fitness wave for fitness modifier 0.0005
Rys. 15. Fala dopasowania dla modyfikatora dopasowania 0.0005

Moving fitness wave provide information about population adaptation so also about population growth. If fitness wave is steady, population evolution should be slow and random. For positive movement of fitness wave can be observed growing cell number and also mutation number.

11.4. Conclusions

Studying clonal evolution of tumor cells reveals changing dynamics of the size of tumor as well as of numbers of somatic mutations in cancer cells. These dynamics is a derivative of acquisition of somatic mutations in the cells and is related with mutations wave moving forward. In principle, the occurring mutations may alter cell fitness/adaptation in different ways: if the point mutation occurs at a gene causing the cell to divide or survive more likely, this mutation gives an advantage for the cell and the underlaying gene. Likewise, the mutation can occur at gene with little to none effect on cell fitness/adaptation or it can cause deterioration of evolutionary fitness of the cancer cell carrying it.

In this study we analyzed simulation scenarios of evolution of cancer clones with each somatic mutation causing small increase of fitness of cancer cells. We have elaborated and implemented two versions of Gillespie’s algorithm and we have pursued several computational/simulation experiments.

We have observed phenomena in genomic clonal evolution of cellular populations described in the literature, population growth in response to increasing adaptation of cells and traveling wave of advantageous mutations and fitness in the cancer cells population.

Acknowledgements

This publication was supported by the Department of Graphics, Computer Vision and Digital Systems, under research project for young scientists (Rau6, 2022), Silesian University of Technology (Gliwice, Poland).

Bibliography

1. Liu Jianfang, et al.: An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 173.2 (2018): 400–416.
2. Bamford Sally, et al.: The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. British journal of cancer 91.2 (2004): 355–358.
3. Bozic I., et al.: Accumulation of driver and passenger mutations during tumor progression, PNAS, Vol. 107, no. 43, p. 18545–18550
4. McFarland C.D., et al.: Tug-of-war between driver and passenger mutations in cancer and other adaptive processes, PNAS, Vol. 111, no. 42.
5. McDonald Thomas O., Marek Kimmel: Multitype infinite-allele branching processes in continuous time. Journal of Applied Probability, 54.2 (2017): 550–568.
6. Hoban Sean, Giorgio Bertorelle, Oscar E. Gaggiotti: Computer simulations: tools for population and evolutionary genetics. Nature Reviews Genetics, 13.2 (2012): 110–122.
7. Miller Christopher A., et al.: Visualizing tumor evolution with the fishplot package for R., BMC genomics, 17.1 (2016): 1–3.
8. McDonald Thomas O., Franziska Michor: SIApopr: a computational method to simulate evolutionary branching trees for analysis of tumor clonal evolution. Bioinformatics 33.14 (2017): 2221–2223.
9. Gillespie D.: Approximate accelerated stochastic simulation of chemically reacting systems. Journal of Chemical Physics, Volume 115, Number 4, p. 1716–1733.
10. Szymiczek K., et al.: CCES: cancer clonal evolution simulation program, in Advances in Intelligent Systems and Computing vol 659 2194-5357, 5th International Conference on Man-Machine Interactions, ICMMI, 2017, Cracow, Poland, Springer International Publishing, p.: 172–181.

11. Anderson David F.: A modified next reaction method for simulating chemical systems with time dependent propensities and delays. The Journal of chemical physics 127.21 (2007): 214107.
12. Neher Richard A.: Genetic draft, selective interference, and population genetics of rapid adaptation. Annual review of Ecology, evolution, and Systematics, 44 (2013): 195–215.

IMPROVING GILLESPIE SIMULATION ALGORITHM FOR FITNESS IN CLONAL EVOLUTION

Abstract

Haploidal clonal evolution governs adaptation dynamics of many populations to environmental conditions. Importantly, clonal evolution stands behind growth of cancer cells populations in human tumors. Genetical forces behind haploidal evolution are replications, mutations, selection and genetic drift. Due to the lack of recombination in the process of haploidal evolution of a population one observes formation of population clones, i.e., subpopulations of identical/similar genetic profiles.

Important area of studying evolution of clones in haploid evolution is mathematical modelling. Due to nonlinearity, interference of several forces and large scale of models mathematical modelling is often supported by computer simulations. In this study we present a simulation system for modelling clonal evolution of haploidal populations based on Gillespie scenario of generating evolutionary events. Due to large cellular/bacterial population we propose modifications of the algorithm based on binning subgroups of cells with equal number of mutations and generating distributions of times of cellular divisions, deaths and mutations in subgroups. We demonstrate results of simulations and improvements in efficiency of modelling due to introduced mutations.

Keywords: clonal evolution, mutation waves, numerical model, stochastic simulation, Gillespie algorithm

DOI:10.34918/85112

Katarzyna SIERADZKA[1,*], Joanna POLAŃSKA[1]

Chapter 12. FEATURE SELECTION METHODS FOR CLASSIFICATION PURPOSES

12.1. Introduction

Nowadays, the development of the needs to generate larger and larger data sets is undoubtedly related to the development of technology and equipment that generates this data. It would seem that the more data we collect in a research problem, the more precise the description of the studied phenomena becomes. However, it often turns out that instead of generating more and more accurate and useful information in the analysis of high-dimensional data, we receive redundant and highly distorted information. Without a careful selection of these data, we expose our research to unwanted use of more and more computational resources and the time necessary to analyze even unnecessary and without practical information features. The problem of large amounts of data for analysis occurs with automated text and language analysis. As described by Justin Grimmer et al. [1] in the analysis of a large number of texts, it is not possible to manually read all the articles related to the problem. It is connected not only with a physical lack of time, but also with limited finances. Therefore, there is a need for automatic text analysis, with the use of complex algorithms dedicated to this problem. This phenomenon of data redundancy very often occurs also in the problem of generating data from single-cell sequencing experiments, where as a rule we get a lot of complex information about gene expression. This data is made up of tens of thousands of cells and hundreds or even thousands of features. The need for in-depth analysis and the use of automated methods to search for patterns relevant to the research problem for these data is of key scientific importance [2]. The use of decision support algorithms is therefore of great importance in healthcare, where artificial

---

[1] Department of Engineering and Data Exploratory Analysis, Silesian University of Technology, Gliwice, Poland.

[*] Corresponding author: katarzyna.sieradzka@polsl.pl.

intelligence analyzing real data is able to significantly assist healthcare workers in making decisions, and what is more, such systems can also warn about detected irregularities [3].

However, cooperation with self-learning artificial intelligence algorithms requires careful consideration of the analyzed features. The lack of control over these factors may lead to a significant extension of the waiting time for learning the predictive model, disturbances in the learning process and, consequently, to making wrong decisions [2]. The natural process of generating high-dimensional data is the formation of irrelevant and redundant features. Keeping control over them is therefore particularly important from the point of view of the analysis and computational costs. Due to the high significance and the need to eliminate features, undoubtedly adversely affecting the processes of machine learning and attempts to solve the research problem. In 1997 the first studies on feature selection (FS) were described, based on large data sets [4–5]. Since then, many FS techniques have been developed, and their development continues to this day. Particular workload is focused on the field of machine learning, where the appropriate selection and analysis of features describing the domain, classifying and recognizing significant patterns is the key to obtaining practically useful and interpretable results. The development of these techniques is of particular importance nowadays, when analyzing often tens of thousands of features at the same time [6]. Removal of worthless features enables a significant reduction in the costs and computational resources necessary to conduct an appropriate study. An alternative solution is also the introduction of ensemble feature selection (EFS) algorithms, i.e. a combination of several single FS techniques, which enables the integration of the advantages of many FS techniques while eliminating their disadvantages [6].

12.2. Overviev of irrevelant and redundant features removal techniques

A very frequently used technique in the initial stages of high-dimensional data analysis is dimensional reduction. Dimensionality reduction refers to the reduction of the number of features in a data set while maintaining information relevant from the point of view of the research problem. This technique can be divided into feature extraction and feature selection. The concept of feature extraction refers to the generation of completely new features, which are a combination of original features present in the input data set. This allows you to limit the dimensionality of the data by selecting only those newly created features that will explain the problem to the greatest extent. However, it is very difficult to associate the newly created features with the input features later. The new features are in no way physically interpretable, as was the case with the original features. Feature extraction techniques are very well developed in a machine learning environment and include Principle Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Canonical Correlation Analysis (CCA), among others [7]. On the other hand, feature selection techniques are based on selecting an appropriate number of features without transforming them. The problematic aspect in the case of FS techniques is the selection of this appropriate number of features. With the use of new methods of generating a very large amount of data, we obtain much more information that is important to us, but it should also be remembered that redundant and irrelevant features also arise to a much greater extent. The selected features should constitute a proportionally small part of the set of input features, and at the same time retain as much information relevant as possible from the point of view of the research problem. Feature selection as compared to feature extraction maintains the readability and interpretability of the data, because no transformations combining many features into one are used here. This is an undoubted advantage of the FS techniques, because the obtained sets of selected features have a physically interpretable meaning, which is particularly important when looking for, for example, expression patterns or genes influencing specific diseases [7]. Therefore, further considerations in this manuscript regarding dimensional reduction techniques will focus on feature selection techniques. Among them, we can distinguish three main types: wrappers, filters, and embedded techniques.

12.2.1. Wrappers

Wrapper is a method of selecting the optimal set of features by a learning algorithm. The choice of the classifier is arbitrary, but it is a method that requires a large amount of computational expenditure, and thus a large amount of time, due to the need to run the classifier repeatedly based on different feature subsets [6]. Due to the fact that computing the objective function is a computationally demanding task, wrappers are not a perfect solution for data with a high complexity of features [2]. In a very general approach, this method consists in using the quality of the prediction provided for the selected classifier to determine the usefulness of the analyzed subsets of features. The applied feature search strategies can be divided into a range of searching strategies e.g. hill-climbing , best-first, branch-and-bound, or genetic algorithms [8]. However, the most promising in terms of robust against overfitting, and at the same time the most computationally complex, are greedy search strategies [7]. They are divided into two techniques of features selection: forward selection and backward elimination. Forward selection is the gradual addition of features to a subset of features that ensure the achievement of better and better classification results. In this type of feature selection, the subset of features begins at the empty subset of features and grows with each successive step, until the point where adding more features does not significantly affect the classification process. The opposite effect is seen in the case of backward elimination. In this case, in the following steps, from the full set of features, those with the least promising effect on the quality of the classifier are removed from the feature subset. In the case of the two methods described above, i.e. forward selection and backward elimination, one can often find different opinions about which method gives the best results. Which method we use can of course be defined by the purpose of the research being carried out. The use of forward selection allows us to choose from the set of features the one feature that best allows the separation of variables – the forward selection selects the best differentiating feature in the first step, adding the next one in the next step, which in cooperation with the first one gives better results. On the other hand, backward elimination, due to the fact that it starts on the full set of features, rejects the one feature that contributes the least to the interaction of the remaining features in the set [8]. The earlier dimensionality reduction could seem to be a solution that allows to significantly reduce the time of FS procedures using wrappers methods. However, the selection of a significant, in terms of a complex and difficult research problem, subset of features on a previously limited subset may adversely affect the results of this method. The features that were previously eliminated could potentially be included in a subset selected with the use of wrappers methods, significantly improving the quality parameters of the classifier. In this case, the dimensionality of the data can be reduced using simple linear data transforms such as PCA or LDA, as well as more sophisticated ones, such as the Fourier transform, while (as mentioned in the introduction) it deprives the possibility of a physical interpretation of a selected subset of features [8].

The problem of selecting features is not only based on the selection of the smallest possible subset of features that will enable the achievement of average classification results. Particularly in areas where machine learning is intertwined and cooperates with genetics or health care, one of the most important aspects is data interpretability, finding disease patterns, connections of individual features, signal pathways, and recreating a certain biological history that stands behind the actual state of health. However, in the case of data with very large dimensions, and we have to deal with such data very often, the feature selection path should be very seriously considered.

As Isabelle Guyon et al. writes [8], this path is based on using a linear predictor of choice, and consists in the first step of preselecting features using simple statistical methods such as correlation and ranking of features, and in the second step using forward selection or backward elimination methods. Choosing such a path with several thousand features may significantly accelerate the computational processes, and does not have to deteriorate the results for the selected subset of features in a statistically significant way.

12.2.2. Filters

The second FS method are filters. Filters as a method of FS is not based on classification, as was the case with wrappers [9]. Therefore, they allow to perform the necessary, to select a subset of the features, analysis in a shorter time, so they can be utilized in the case of high-dimensionality datasets [6]. These methods are quite simple in terms of computational complexity and generally consist of two basic steps. In the first step, features are ranked using an appropriate criterion. This step can be performed both on individual features and on sets of features. In the next step, features with the highest ranks are selected for a subset of features, on the basis of which further analysis and inference are possible [7]. Many algorithms have been developed that play an important role in selecting filters as FS methods. Moreover, these methods can be divided into unsupervised and supervised [2]. Among the unsupervised methods, we can distinguish: term-variance (TV) criterion [10], Laplacian score (LS) [11], and Fisher score. TV score sorts the features according to their variance in the given samples. LS use the local geometric structure to assess the significance of a feature [2]. In the Fisher score, individual features are treated separately, so it is not able to detect the relationship between the individual features. Thus, it is unable to detect redundant features [7]. In order to solve this problem, Gu et al. [12] proposed a generalized Fisher score in which features are selected together and it is possible to maximize the lower bound of traditional Fisher score. Another approach to using filters in FS methods is the mutual based information method. This method is based on counting the information gain between successive features and labels. The feature is significant if it has a high information gain. In this method, as in the traditional Fisher score approach, features are selected in univariate way and it is not possible to determine redundant features. The second group of filters are supervised methods such as Relief [13]. It is an algorithm based on the measure of significance of features, the value of which is related to how well the values of this feature distinguish instances of the same and other classes. This method is very useful in the analysis of data of high complexity, but it does not allow the removal of redundant features [2]. In the ReliefF method, extended by the multiclass problem, compared to the Relief, which was used to handle two-class problems, the features are selected in such a way as to separate the individual randomly selected, in the subsequent stages of the algorithm's operation, elements of the data set coming from different classes [13]. In this case, the quality of the features is calculated on the basis of how well their values are able to distinguish between individual instances located close to each other [14]. Another, also well known, FS algorithm is the fast correlation-based filter (FCBF) [15]. This method is based on the significance of features, but cannot be successfully applied to high-dimensional data [2]. For this reason, a new solution was introduced, such as the minimum redundancy maximum relevancy (mrMR) criterion [16]. It focuses both on redundancy between features and the importance of each feature. As Artur J. Ferreira et al. [2] writes, many algorithms do not work well with high-dimensional data. They proposed a new filter-based procedure dedicated to the large-dimension data approach. The first proposed algorithm uses the dispersion measure to determine the relevance of each feature, and then sorts them in descending order to preserve the selected number of features with the highest value of the measure used. The second algorithm, however, when more than one feature turns out to be redundant, is responsible for selecting only one of them, making comparisons of subsequent features. These comparisons, for the sake of saving computing time, are performed only for the top relevant features. As the authors write, this method removes the most redundant features from a subset of the most relevant features. What is worth emphasizing, both algorithms can work in unsupervised or supervised mode.

12.2.3. Embedded methods

Embedded methods are another method used in FS procedures. Their main goal is to derive the best results from the learning process based on a subset of features. For this reason, these are methods that combine the previously described wrappers and filters [6]. The feature selection algorithm is integrated as part of the learning algorithm. A learning algorithm takes advantage of its own features selection process and performs FS and classification at the same time. Embedded methods combine the advantages of wrappers, i.e. they take into account interactions with the selected classification and filter model, i.e. they are much less computationally demanding than wrappers [7, 17–19]. Currently, there are many variants of embedded methods for the FS problem. These are, for example, C4.5 programs [20] to create decision trees, random forest [21], as well as algorithms based on multinomial logistic regression with its variants [22]. The random forest methodology is based on the combination of many decision trees. Each individual tree is created on the basis of a random sample selected from the input data set. For each tree, in each node, successive divisions occur on the basis of randomly selected traits, which are used as candidate traits to make the best division. In the course of subsequent divisions, the statistics of the significance of features are calculated, among others [23]. Other methods are based on regularization models which by minimizing the matching errors in a short time make it possible to set some feature coefficients to very small or even exact zero values [24, 25]. Methods that penalize features that do not affect the model's performance are typically those that work with linear classifiers (for example, SVM [26]), such as LASSO [24]. LASSO is the Least Absolute Shrinkage and Selection Operator method utilized in linear models, but can be applied in another statistical models including tree-based and generalized regression models [27]. LASSO uses a regularization procedure to reduce the value of the regression coefficients. In the process of selecting features after regularization, variables with non-zero values are selected for the model. In this case, the lambda parameter is of great importance, as it is used to control the strength of penalties, and thus affects the number of zero regression coefficients [28]. Despite the many advantages of embedded methods, it should be noted that we are not able to calculate the significance value of features for all types of machine learning algorithms, for example for nearest neighbors method [19].

12.3. Classification problem

The selection of traits mainly affects the training phase of the classification process. The FS process can be completely independent of the learning process (filters), but it can also be built into an algorithm that iteratively evaluates the significance of individual selected features (wrappers) [7]. The feature selection is of key importance, because its task and assumption is to select a subset of features that in the best possible way discriminates against observations belonging to different classes. Thus, the significance of a given trait can be described by its ability to distinguish between classes. When trying to define the concept of classification itself, it can be concluded that it is a problem of assigning an unknown observation to a specific class, based on the training set for which the belonging of observations to particular classes is known. In the learning process, the algorithm uses information about selected features and belonging of individual observations to specific classes to build the function of assigning observations to classes. Having this function, in the prediction process, the classifier assigns new observations to membership classes based on the previously collected information. A very important element related to the classification problem is the validation set, which should not participate in any of the stages of teaching the classifier. This set should be randomly selected before the information is introduced into the learning process. This will ensure the independence of this set and enable a real assessment of the quality of the final classifier performance. An additional advancement in the learning process is the drawing from the rest of the training set dataset and repeating the training procedure on randomly selected subsets. This procedure can be essential for very small sample sizes. The assessment of what proportion of the input data should be the test set and which part of the training set is unclear. In the case of small sets, it is possible to apply leave-one-our cross-validation, which consists in drawing one observation from the set and then testing it. The rest of the set is for training. This method, due to a very large generalization in the selected observation, is burdened with a very large evaluation error [8, 29]. In the case of classifiers, we can of course use both supervised and unsupervised methods. The supervised selection of features determines their significance based on information about the class affiliation of observations, but for the learning process to be successful, there is often a need to provide a large number of observations with assigned affiliation labels, which is time-consuming. On the other hand, unsupervised selection works, of course, based on observations without assigned labels, but in this case the assessment of the significance of the features is not straightforward and is difficult [7].

In general, there are four basic and necessary steps to consider [17]: feature selection, feature evaluation, stop criterion, and validation. In the first step, using an appropriate method, a subset of features is selected, which in the next step is assessed according to a specific criterion. The selection of significant features is completed when the assumed stop criterion is reached. The selected subset is evaluated against the validation set.

12.4. Proposed method

The process of selecting a subset of features, that will be utilized to build a final model for heterogeneous data set, proposed in this manuscript is a two-step process. The first step is to use the wrapper method, more specifically the forward selection.

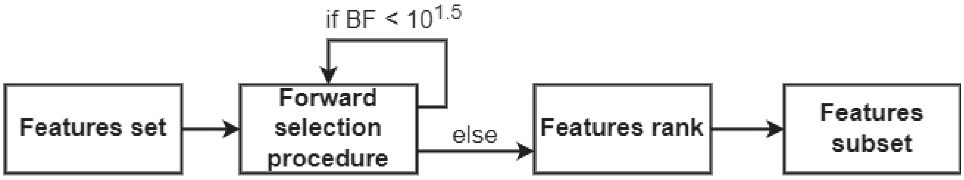


Fig. 1. General outline of presented method for features selection problem
Rys. 1. Ogólny zarys zaprezentowanej metody selekcji cech

The second stage is the use of the filter methodology and the introduction of the feature ranking measure on the subset previously pre-selected by the forward selection algorithm. A general diagram showing the individual steps of this method is presented in Fig. 1. In the forward selection procedure, Bayes Factor (BF) was used as a measure informing about the profitability of keeping the model with a higher degree of complexity (more features included in the proposed model). The BF value was calculated from (1).

$$BF = e^{(LogLikelihood_i - LogLikelihood_{i-1})} \tag{1}$$

where:
i means the level of models' complication.

In the initial set subjected to the FS procedure, there were 406 features. The initial data consisted of 4,214 items, among which two classes were distinguished. In the control class, there were 2,252 elements, and in the class subjected to the research procedure, hereinafter referred to as the procedural class, 1,962 elements. A very important step, which was carried out in the initial phase of the analysis, was to determine the elements representing the validation set. These elements were randomly selected from both classes, maintaining the dependence of 20% of the number of elements in these classes. The remaining part of the input data created a set which was subjected to the procedure presented in Fig 1. The numbers of elements broken down into classes and generated sets are presented in Table 1.

Table 1

Number of elements in both randomly generated sets

	FS procedure	Validation
Control class	1800	452
Procedural class	1571	391
Overall	3371	843

12.4.1. Forward selection procedure

The selection of a subset of features consists in running the model building, logistic regression (LR) methods based, algorithm 50 times. In the initial phase of the procedure, the training and test set are drawn. Importantly, the test set is balanced, each time the algorithm is run, in terms of the number of elements from the control and procedural class. 1012 elements were randomly selected for the test set (506 elements from each of the two classes). The remaining part of the set subjected to the feature selection, i.e. 2359 elements (1294 from the control class and 1065 from the procedural class), was the training set. This set was then subjected to the feature selection stage using forward selection methods. In the first stage of model building, single-element models were created taking into account each of the available features. The values of the N parameters of the generated models were calculated until a sufficiently small difference was obtained between successive likelihood value estimates for the model with a given set of parameters. All N univariate models were then compared using the Bayesian Information Criterion (BIC) measure described by (2). The best model, with the lowest BIC value, was recorded along with the estimated log-likelihood value. In the next step, we started building models composed of two features.

$$BIC = N_{parameters} \times \ln(N_{cells}) - 2 \times \mathrm{LL} \tag{2}$$

where:

LL is the log-likelihood function.

For this purpose, each of the remaining N-1 features was attached to the feature selected in the previous step, which resulted in the generation of N-1 two-element models. The procedure for selecting the best two-factor model remained the same as for the one-factor models. The BIC values of each of the generated models were compared and the best one was selected. The BF measure was used to determine whether the previously selected one-factor model or the two-factor model calculated in the current iteration will be kept. A threshold indicating strong evidence in favor of a more complex model was selected as the value determining the next stages of the procedure (the exact value of the BF threshold is shown in Fig. 2). If a two-factor model is selected, the three-factor model building procedure is started and the procedure is the same as that described above. If a model with a less complexity is selected, the procedure of adding further features to the model is completed. In a critical case, the procedure can also be completed when all available features in the input set are used.

After receiving 50 models generated in this way, testing is carried out with the use of randomly selected, in subsequent runs of the algorithm, balanced test sets. The entirety of the described procedures leading from the input set of features to model testing is presented in Fig. 2.

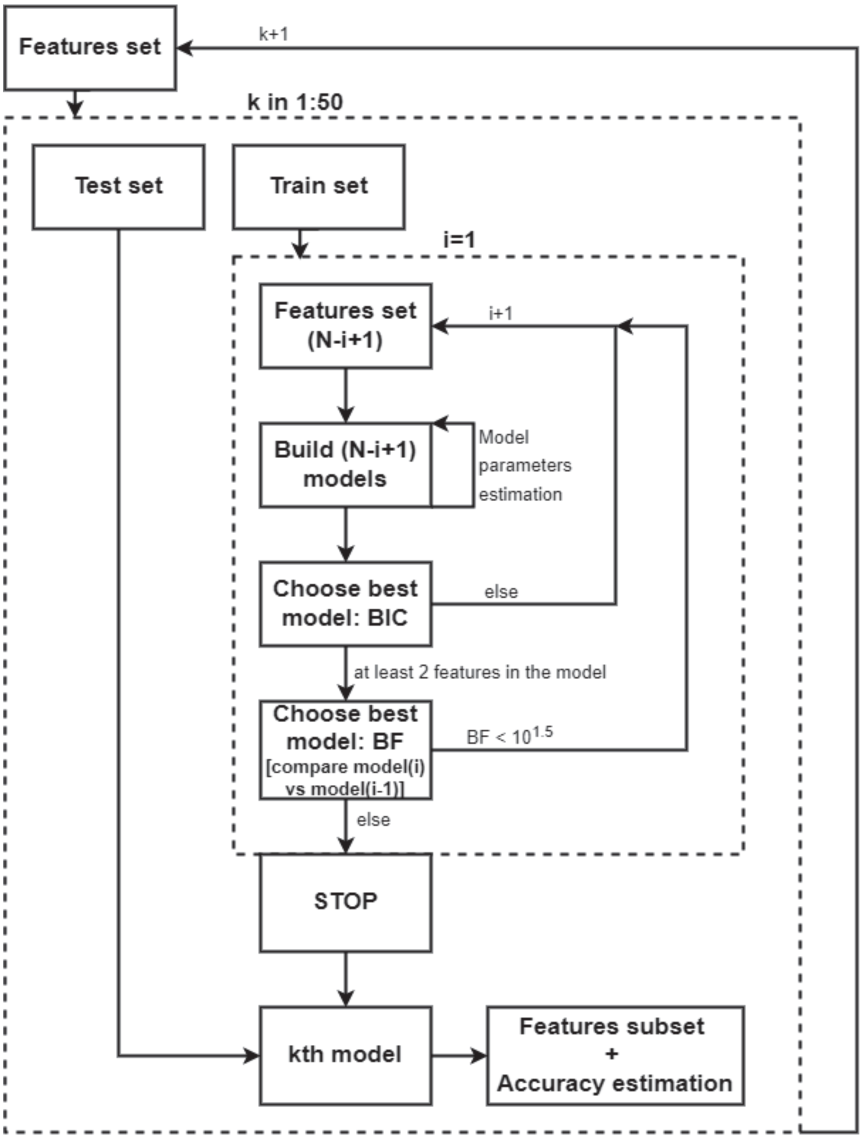


Fig. 2. Forward feature selection procedure. Where: *k* is the algorithm run ID, *N* is the total number of features, and *i* is the number of genes in model

Rys. 2. Procedura selekcji cech w przód. Gdzie: k to identyfikator uruchomienia algorytmu, N to całkowita liczba cech, i to liczba genów w modelu

12.4.2. Features rank

In the second step of FS, the filters method was used, based on the ranking of features. The ranking was created on the basis of 50 models generated with the use of logistic regression methods. The FeatureRank measure from (3) is based on 3 key elements related to individual features across all generated models: classification quality for a given feature related to the test set, number of features in the longest model, and the position of a feature in the model. Each feature that appeared at least once in one of the 50 models was taken into account in the ranking. The enumerated FeatureRank values were then normalized to a range of 0 to 1.

$$FeatureRank = \sum_{j=1}^{N} \frac{accuracy_{test_j} \times (k - i_j + 1)}{k} \tag{3}$$

where:
N is the models' ID, k is the number of features in the longest model, i is specified features' position in the model, and $accuracy_{test}$ is the estimated accuracy value for testing set for specified model.

With the use of the ranking of features, there has also been a need to determine the appropriate number of features that are important. For this purpose, the cut-off threshold was set before the features that showed a decrease in the differences in the estimated FeatureRank values. In other words, the features were rejected which not only had very low values of the measure assigned, but also the differences between these values for the following features were insignificant.

12.5. Feature selection results

Table 2

Estimated accuracy values based on testing set

Model ID	Testing accuracy [%]	Number of features	Model ID	Testing accuracy [%]	Number of features
1	91.25	31	**26**	91.87	31
2	92.27	33	**27**	92.20	31
3	90.32	27	**28**	89.57	32
4	92.99	30	**29**	90.81	37
5	90.35	26	**30**	90.43	33

cont. table 2

Model ID	Testing accuracy [%]	Number of features	Model ID	Testing accuracy [%]	Number of features
6	91.31	34	**31**	90.63	22
7	90.24	31	**32**	91.65	29
8	91.51	30	**33**	**89.37**	25
9	92.61	28	**34**	91.22	23
10	**93.59**	37	**35**	89.95	21
11	90.13	21	**36**	89.88	34
12	91.32	28	**37**	91.95	28
13	91.09	24	**38**	90.43	31
14	91.99	30	**39**	91.38	24
15	91.24	23	**40**	91.22	28
16	90.52	29	**41**	91.60	36
17	91.54	30	**42**	90.91	31
18	91.61	29	**43**	91.12	36
19	92.31	39	**44**	91.40	22
20	93.58	32	**45**	93.43	33
21	91.57	39	**46**	91.13	28
22	89.97	29	**47**	92.69	28
23	92.43	36	**48**	90.82	33
24	91.04	26	**49**	91.91	35
25	91.91	27	**50**	90.24	29

After using the logistic regression-based classifier on the input set of 406 features the average value of the classification quality, for the test sets for 50 models, was obtained at the level of 91.33% with the 95% confidence interval (91.05÷91.61). The minimum value for quality is 89.37% and the maximum is 93.59%. Table 2 describes the obtained classification qualities for the test sets for all the created models. Importantly, in this case, the quality of the classification was not counted as the weighted quality of the classification due to the fact that the test set was balanced in each subsequent draw in terms of the number of elements from the control class and the procedural class. Additionally, there was also estimated the 95% confidence interval for the mean number of features included in specified models which equals (28.51÷31.05).

After receiving a set of 50 models with information on selected features, the FeatureRank values were calculated for each feature that appeared in any of the models at least once. Calculations of this measure were performed for all models simultaneously. 159 features entered this stage of feature selection, based on filter methods. Each of them was assigned a FeatureRank value ranging from 0 to 1, and then the features were ordered in descending order of significance measure. The ranking of the features is presented in Fig. 3 below.

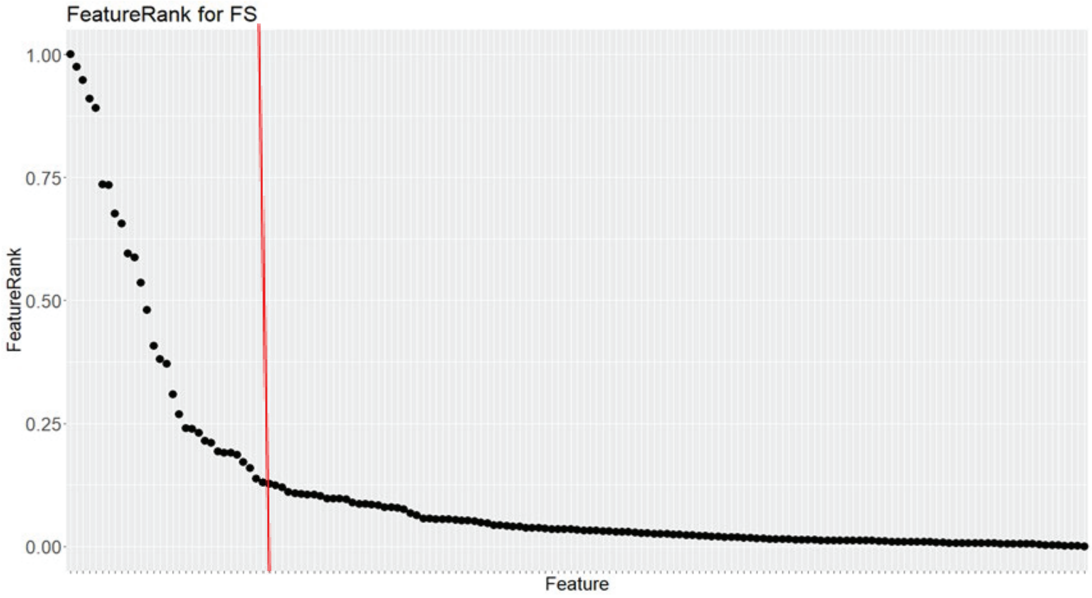


Fig. 3. Ranked FeatureRank metric with marked threshold for the number of selected features
Rys. 3. Miara FeatureRank z oznaczonym progiem odcięcia dla liczby wyselekcjonowanych cech

Importantly, the first feature in the ranking (the highest FeatureRank value) obtained the value of the measure equal to 1. This means that this feature always appeared in the first place in each of the 50 created models, i.e. the implemented algorithm each time, regardless of the selected training set, considered this feature as the most important, enabling the best division of the two classes in the set. Figure 3 also shows the cut-off point for the number of features selected for the model. Below this threshold, the features are characterized by insignificant differences in the value of the determined FeatureRank significance measure.

By first using the wrapper method with the use of a classifier based on logistic regression, and then the filter method based on the entered significance measure, the dimensionality of the data set was significantly reduced from 406 input features to 29 selected features.

Table 3

Parameter values for final model

Feature ID	**Intercept**	**1**	**2**	**3**	**4**
Parameter	-2.4687	0.7876	0.2519	-0.2097	0.7381
	5	**6**	**7**	**8**	**9**
	1.2674	0.7112	0.4770	-0.6276	-0.2917
	10	**11**	**12**	**13**	**14**
	0.8514	0.3900	-0.2118	0.3395	0.8966
	15	**16**	**17**	**18**	**19**
	-0.1732	-0.6903	-0.3346	-0.4697	-0.2658
	20	**21**	**22**	**23**	**24**
	-0.1970	-0.1940	-0.0599	-0.4010	-0.3079
	25	**26**	**27**	**28**	**29**
	0.2947	-0.2597	-0.2527	-0.4715	-0.1366

In the next step, the quality of selected features was assessed in a given research problem, i.e. the ability to separate elements from the control class and procedural class. For this purpose, parameter values were determined for individual features included in the final model, maintaining the order of features resulting from the obtained FeatureRank metric values. The values of the estimated parameters are shown in Table 3.

In order to determine the threshold value of the classification probability, the Youden index method was used. This method finds a trade-off between the sensitivity and specificity of the classification, and the computation is stepwise for each possible value of the classification probability. For this purpose, the ROC curve was determined and the value of the threshold classification probability was estimated as the point farthest from the diagonal of the plot. The new probability threshold value for the classification, determined with the use of the Youden index, was 0.7047.

In the last step of the study, the quality of the constructed classifier based on logistic regression methods was determined, based on the validation set. It is worth emphasizing that this set was not used at any stage of constructing the classifier and was completely randomly selected from the set of input data.

Table 4

Classification quality metrics based on the validation set

Quality metric	Result
PPV	0.9385
NPV	0.9147
Sensitivity	0.8977
Specificity	0.9491
Weighted accuracy	0.9253

The following measures of classification quality were used: Positive Predictive Value (PPV), Negative Predictive Value (NPV), Sensitivity, Specificity, and the weighted quality of the classification, which takes into account the disproportions in the representation of individual classes Table 4 shows the results related to the classification quality of the validation set.

12.6. Discussion

The application of the proposed approach, taking into account the use of the forward selection technique as the leading method, requires large computational resources and is quite time-consuming, especially when considering large-scale data sets. Undoubtedly, a very big benefit resulting from the use of a time-consuming and computationally complicated methodology of forward selection on a full set of features is its potential to reveal often hidden relationships between successive features. By using filters, as the first FS method, we are able to significantly limit the input set of features to the most important ones, in other words allowing to capture the greatest differentiation between the studied groups. However, by using such a scheme, we can remove the hidden connections between the features in the first stage. It is particularly important when building classifiers to distinguish various biological complexities, in very heterogeneous data sets, related to gene expression, e.g. healthy tissue and cancer tissue, control cells and irradiated cells, distinguishing types of cancers, etc. In very complex biological problems, it is important to capture any differences, but also the connections and cooperation of individual features and their common, cumulative ability to evaluate a given phenomenon. Hence, the applied approach based on the use of wrapper techniques in the first phase of feature selection, which is much more time-consuming and computationally complicated than filters, allowed for the capture of many hidden inter-features dependencies. The further methodology, i.e. the use of a simple metric to determine the significance of individual selected features, allowed for a further reduction in dimensionality and the selection of only the most essential features. The proposed method is not only able to capture what is seemingly invisible in the analyzed sets of features, but also allows you to select features that separate classes of elements present in the given classes with very satisfactory results, which was confirmed in this work. What is more, using the proposed integration methodology, we are able to build the final model that can be successfully used to classify elements from highly heterogeneous data sets.

12.7. Conclusions

The presented results concerning the classification quality measures for 29 selected out of 406 features unambiguously allow to state that the selected features allow, with satisfactory results, to separate the elements from the control and procedural groups. The constructed classifier is characterized by a very high specificity, i.e. the ability to correctly classify elements from the control group, and a slightly lower, but still high value of sensitivity, i.e. the ability to correctly classify elements from the procedural group. The very good results of the classification are clearly evidenced by the high value of the weighted quality at a level above 92.5%, while the 95% confidence interval for mean weighted accuracy value for 50 build models was equal (91.05÷91.61).

Referring to the applied FS methodology, it allowed for a significant reduction in the dimensionality of the data. The first stage of FS, i.e. the application of the method belonging to the wrapper group - a classifier using the forward selection procedure, allowed for the identification of 159 important features from the point of view of the classification problem. Already this stage allowed for a significant reduction in the number of features that were subjected to the second stage of FS, i.e. the use of the methodology from the group of filters. Describing each of the 159 features, using a measure based on the previously calculated values of the classification quality on the test sets and the order of attaching subsequent features to individual models, allowed for the selection of features that have a significant impact on the quality of the classification. At this stage, there was another dimensionality reduction from 159 to

29 features that were incorporated into the final model. What is also worth emphasizing, the final model for heterogeneous data set classification, was built from 29 features, while the 95% confidence interval for the average value of the number of features over the 50 built models was (28.51÷31.05).

The presented method, integrating the methodology of wrappers and filters, allows for a significant reduction in dimensionality (the number of features), while maintaining a very high quality of classification, in relation to a very heterogeneous set of data.

Acknowledgments

This work has been supported by European Union under the European Social Fund grant AIDA – POWR.03.02.00-00-I029, and by the Silesian University of Technology Grant BK-02/070/BK22/0033.

Bibliography

1. J. Grimmer, B.M. Stewart, Text as data: The promise and pitfalls of automatic content analysis methods for political texts., Political analysis 21.3, pp. 267–297, (2013).
2. J. Ferreira, M.A.T. Figueiredo: Efficient feature selection filters for high-dimensional data., Pattern recognition letters 33.13, pp. 1794–1804, (2012).
3. R. Spencer, F. Thabtah, N. Abdelhamid, M. Thompson: Exploring feature selection and classification methods for predicting heart disease, Digital health 6, p. 2055207620914777, (2020).
4. L. Blum, P. Langley: Selection of relevant features and examples in machine learning, Artificial intelligence 97.1–2, pp. 245–271, (1997).
5. R. Kohavi, G.H. John: Wrappers for feature subset selection, Artificial intelligence 97.1–2, pp. 273–324, (1997).
6. M. Mera-Gaona, D.M. López, R. Vargas-Canas, U. Neumann: Framework for the Ensemble of Feature Selection Methods, Applied Sciences 11.17, p. 8122, (2021).
7. J. Tang, S. Alelyani, H. Liu: Feature selection for classification: A review, Data classification: Algorithms and applications, p. 37, (2014).
8. I. Guyon, A. Elisseeff, An introduction to variable and feature selection., Journal of machine learning research, pp. 1157–1182, (2003).

9. H. Liu, H. Motoda (ed.): Computational methods of feature selection, CRC Press, (2007).
10. L. Liu, J. Kang, J. Yu, Z. Wang: A comparative study on unsupervised feature selection methods for text clustering, in 2005 International Conference on Natural Language Processing and Knowledge Engineering. IEEE,(2005).
11. X. He, D. Cai, P. Niyogi: Laplacian score for feature selection, Advances in neural information processing systems 18, (2005).
12. Q. Gu, Z. Li, J. Han: Generalized fisher score for feature selection, arXiv preprint arXiv:1202.3725, (2012).
13. K. Kira, L.A. Rendell: A practical approach to feature selection, Machine learning proceedings 1992. Morgan Kaufmann, pp. 249–256, (1992).
14. M. Robnik-Šikonja, I. Kononenko: Theoretical and empirical analysis of ReliefF and RReliefF, Machine learning 53.1 , pp. 23–69, (2003).
15. L. Yu, H. Liu: Feature selection for high-dimensional data: A fast correlation-based filter solution, in Proceedings of the 20th international conference on machine learning (ICML-03), (2003).
16. H. Peng, C. Ding: Minimum redundancy and maximum relevance feature selection and recent advances in cancer classification, Feature Selection for Data Mining 52, (2005).
17. H. Liu, L. Yu: Toward integrating feature selection algorithms for classification and clustering, IEEE Transactions on knowledge and data engineering 17.4, pp. 491–502, (2005).
18. S. Ma, J. Huang: Penalized feature selection and classification in bioinformatics, Briefings in bioinformatics 9.5 , pp. 392–403, (2008).
19. Y. Saeys, I. Inza, L. Pedo: A review of feature selection techniques in bioinformatics, bioinformatics 23.19, pp. 2507–2517, (2007).
20. J.R. Quinlan, C4. 5: programs for machine learning, Elsevier, (2014).
21. M. Sandri, P. Zuccolotto: Variable selection using random forests, Springer, in Data analysis, classification and the forward search., Berlin, Heidelberg, (2006), 263–270.
22. G. Cawley, N. Talbot, M. Girolami: Sparse multinomial logistic regression via bayesian l1 regularisation, in Advances in neural information processing systems 19, (2006).
23. C. Nguyen, Y. Wang, H.N. Nguyen: Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic, Journal of Biomedical Science and Engineering, (2013).

24. S. Ma, J. Huang: Penalized feature selection and classification in bioinformatics, Briefings in bioinformatics 9.5 , pp. 392–403, (2008).
25. A. Jović, K. Brkić, N. Bogunović: A review of feature selection methods with applications 2015, in 38th international convention on information and communication technology, electronics and microelectronics (MIPRO), Ieee, (2015).
26. I. Guyon, J. Weston, S. Barnhill, V. Vapnik: Gene selection for cancer classification using support vector machines, Machine learning 46.1, pp. 389–422, (2002).
27. R. Tibshirani: Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58.1, pp. 267–288, (1996).
28. V. Fonti, E. Belitser: Feature selection using lasso, VU Amsterdam research paper in business analytics 30 , pp. 1–25, (2017).
29. V.N. Vapnik, A.Y. Chervonenkis: Necessary and sufficient conditions for the uniform convergence of means to their expectations, Theory of Probability & Its Applications 26.3, pp. 532–553, (1982).
30. R.-C. Chen, C. Dewi, S.-W. Huang, R.E. Caraka: Selecting critical features for data classification based on machine learning methods, Journal of Big Data 7.1, pp. 1–26, (2020).
31. R. Quinlan, in C4.5: Programs for Machine Learning. Morgan Kaufmann, (1993).
32. J.R. Quinlan: Induction of decision trees. Machine learning, 1(1), (1986), p. 81–106.

FEATURE SELECTION METHODS FOR CLASSIFICATION PURPOSES

Abstract

Feature selection methods are nowadays more and more developed. Modern and very accurate techniques that allow for the generation of very extensive data sets are also becoming more popular. Text analysis sections and the new, promising single-cell sequencing technique are specific areas with the privilege of high-dimensional data. A very large amount of expenditure on the continuous improvement of feature selection methods is widely appreciated by scientists and analysts, especially now. Methods combining two popular techniques of feature selection – wrappers and filters – are becoming more and more popular. The method proposed by us, combining the effectiveness of wrappers techniques and the speed of filters techniques, allows you to choose the features important from the classification point of view with great efficiency. These features are capable of carrying significant information about the differences between elements coming from different classes. What is more, our method also allows us to capture many invisible, without complicated analyzes, relationships between the analyzed features. The effectiveness of the proposed methodology is supported by a very high quality of heterogeneous data set classification at the level above 92.5%, as well as very satisfactory sensitivity and specificity metrics.

Keywords: feature selection, logistic regression, machine learning, heterogeneous data sets classification

DOI:10.34918/85113

Mateusz KANIA[1,*], Andrzej POLAŃSKI[1]

Chapter 13. UNSUPERVISED CLUSTERING FOR DETECTION OF GENE EXPRESSION PATTERNS IN HUMAN CANCERS

13.1. Introduction

Cancer evolution is a complex dynamical process of uncontrolled growth of tissues/cells with dysregulated signalling, metabolism, and replication mechanisms. It is caused by somatic alterations/mutations of DNA, which can cumulate during mitotic replication of cells. Some studies report that somatic mutation might influence gene expression levels. Our study provides reliable, highly statistically significant support for gene expression pattern occurrence in human cancer. Our analysis is based on the Cancer Genome Atlas (TCGA) (Genomic Data Commons) database.

Our data is focused on gene expression levels in cancer since the matter of gene expression itself is still an extensively researched topic. We are studying the hypothesis that gene expression profiles would allow us to distinguish between different types of cancer. Another goal is to decide which unsupervised clustering algorithms will perform the best in the given task. The criterion performance in various metrics is calculated using clustering results and the ground truth.

13.2. Methods

For the experiment, we used the data from cBioportal, a portal for cancer genomics data. It is related to TCGA (The Cancer Genomic Atlas) being an interactive resource for the exploration of multivariate cancer genomic data. Moreover, cBioPortal

---

[1] Department of Computer Graphics, Vision and Digital Systems, Silesian University of Technology, Gliwice, Poland.

* Corresponding author: mateusz.kania@polsl.pl.

provides open access to molecular profiles and clinical attributes of different cancer genomic studies.

The resources of cBioPortal contain but are not limited to DNA methylation data, mRNA and microRNA expression or phosphoprotein level data (RPPA). We have used mRNA (messenger RNA) expression data for our analysis. mRNAs are the product of DNA transcription. The central role of messenger RNA is to function as a template for translation. During this process, mRNA sequences are first translated to amino acids, which then build functional proteins. Increased or decreased mRNA levels might be related to various diseases, including cancer. We wanted to determine if this kind of data contains enough information to distinguish between different types of cancer. The data we have used consisted of the median expression level of RNA sequencing data. We parsed the data to the format presented in Table 2. In rows, we gathered different cancer types mentioned in Table 1.

Table 1

Types of cancers

Cancer type
Stomach adenocarcinoma
Glioblastoma multiforme
Lung squamous cell carcinoma
Lung adenocarcinoma
Breast invasive carcinoma
Ovarian serous cystadenocarcinoma
Brain lower grade glioma
Thyroid carcinoma
Prostate adenocarcinoma
Pancreatic adenocarcinoma

Each row contains a subject, and each column describes the case using gene expression information. Next, we mixed gene expression information from all selected cancer types in possible combinations without repetitions. We created 50 mixtures of 2, 3, 4, 5 and 6 components. Each mixture was created ten times, containing different types of cancer. In addition, each set consisted of almost 20 000 features, and the number of observations ranged between 500-4000. Since the data is categorical and data points belong to real numbers, we assumed that this type of data might be described by a mixture of gaussian distributions.

Table 2

Input data – simplified example

	Gene 1	Gene 1	Gene 2
Cancer 1	15.4	15.4	5.2
Cancer 1	11.67	10.6	2.57
Cancer 2	18.67	12.64	1.8

13.2.1. Pipeline

The very first step in the analysis was data preprocessing. The data was parsed to fixed matrix format N × M where N indicates the number of patients and M number of genes, our variables of interest. Since we had to analyze more than 20 000 features, we used the decomposition method based on a Gaussian mixture of variances. According to the model, the features with the highest variance were left in the dataset, and the rest was treated as noise. In our comparison, we used four unsupervised algorithms: k-means, hierarchical agglomerative clustering, fuzzy c-means, and Gaussian EM. The listed algorithms are based on distance metrics, while the last uses a mixture of Gaussian distributions. To measure algorithms efficiency, we combined a few approaches. One was to prepare a binomial test for each algorithm. The other one was calculating Adjusted Rand Index, Simple Matching Coefficient, its weighted version and Averaged Jaccard Index. We applied the Hungarian algorithm, for the SMC index, which allowed us to assign classification results to their respective clusters.

13.2.2. Hierarchical clustering

Hierarchical clustering (HC) is called such because of the way it creates the clusters. HC results are a series of partitions with a visible hierarchy resembling tree branches. In the analogy, each branch is a cluster. The bigger branch consists of many smaller branches; at the end, it becomes a trunk, that incorporates all the data. However, there are two ways to cluster the data. We can start from a single point up to the whole data set or the complete data and successively build clusters up to one point. It is called agglomerative and divisive clustering. In our analysis, we are using the agglomerative method [3].

Hierarchical Agglomerative Clustering (HAC) is the most popular way to cluster the data. We first must decide upon the distance and linkage methods we will use during the analysis. The distance method explains how the distance or similarity between points will be measured. Some commonly known measures are Euclidean, Manhattan, Minkowski and others. The second choice is the linkage method. It determines how data points will be grouped in consecutive clusters. The few examples here will be single, complete or average linkage. The hierarchical clustering model allows us to choose any number of clusters without the need to repeat the calculations. It is a unique attribute of HC, not present in other unsupervised algorithms. Initially, we need to determine the metric that describes a relation between the data points. To do that, we can use similarity (e.g., Jaccard index) or distance (e.g., Euclidean) measure. The choice depends on scientific questions and the data itself. Next is the choice of linkage method. It will determine the way of how the data will be clustered together. It has a heavy impact on the results.

Last but not least thing to do is choose the number of clusters. We can do it in two different ways. The first one is to choose the number of groups exactly. Another way is to cut the branches at a specific tree height. The tree's height depends on the largest distance or similarity between two clusters in the data. We can also base our choice of clusters on this metric [2].

$$d_E(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$d_E(X,Y) = d_E(X,Y)$$

where:
d_E – Euclidean distance
X, Y – probability distributions
x_i, y_i – realizations of X and Y
As for the linkage method, we used Ward's method that is based on minimizing error sum of squares (ESS).

$$\mathrm{ESS}\ (X) = \sum_{i=1}^{N_X} \left| x_i - \frac{1}{N_X} \sum_{j=1}^{N_X} x_j \right|^2$$

$$d(X,Y) = ESS(XY) - [ESS(X) + ESS(Y)]$$

13.2.3. k-means

K-means algorithm is an iterative, distance-based algorithm. It is easy on resources, quick and computationally effective. Because of the low usage of memory, k-means is suitable for clustering huge data sets. It is a significant advantage over hierarchical clustering methods. Apart from their informative tree structure, they are computationally heavy.

Moreover, a k-means algorithm might be used to initialize other algorithms, for example, those based on Expectation-Maximization. From the mathematical perspective, k-means is similar to the normal mixture model. Estimation of parameters is done by the maximum likelihood method. The primary idea behind the k-means is that observations are gathered around artificially introduced centres, called centroids. Centroid can be treated as a mean generalization, a geometric centre of a convex object. In general, the distance between centres and observations should be minimal. Data points closest to the particular centre are part of its cluster [3]. The initial number of centroids is equivalent to the number of clusters k in the data. The number of initial groups is required to start the algorithm. There are different ways to choose the number of clusters beforehand, but we can also use expert knowledge or assumptions. The algorithm stops in a few cases. The most desirable one is the occurrence of convergence. For example, the creation of clusters with the highest similarity of points within a given cluster and the lowest between different ones. In the commonly used Hartigan-Wang algorithm, the stop criterion is based on minimizing the total sum of variance within clusters (WCSS). It is given by the formula [5].

$$WCSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left\| x_{ij} - c_i \right\|^2$$

There are a few k-means algorithms: Lloyd, Forgy, MacQueen and the one already mentioned, Hartigan-Wong. The last one is the default k-means algorithm in the R software, used in the study.

13.2.4. Multivariable Gaussian mixture clustering (GaussEM)

Fitting the multivariable Gaussian mixture model to data can be done using the Expectation-Maximization algorithm was (Dempster, Laird and Rubin [1]). It is commonly used in a situation when the observations can be viewed as incomplete.

Some examples of cases when it is used are: missing data, truncated distributions, censored or grouped data [7]. The usual requirement to start EM for Gaussian mixture is to provide a k number of clusters. Knowing the number of subgroups in the data, we can initialize parameters in the next step. Initialization During initialization, we need to create a first guess of the parameters. In the case of Gaussian mixture, we need to initialize mixing proportions (α), mean (p), and variance σ 2 for each mixture component k ∈ {1..K}. Mixing proportions indicates how much of the mixture space belongs to K. Depending on the number of K, we need to provide equal number of α_K.

$$\alpha_1 \quad \alpha_2 \quad \alpha_3 \quad \cdots \quad \alpha_K$$

Assume that we have a mixture where k = 3. In that case, we need to create three α parameters. We can use uniform distribution $\alpha_K \sim U(0.1, 1)$ to obtain initial alphas. We advise to keeping the interval within [0.1, 1] because shallow values might cause over dominance of larger α during the estimation step. After choosing values from a uniform distribution, they should be standardized.

$$\hat{\alpha}_K = \frac{\alpha_1 + \alpha_2 + \cdots + \alpha_k}{\sum_{k=1}^{K} \alpha_k} \text{ and } \sum_{k=1}^{K} \hat{\alpha}_k = 1$$

The Gaussian probability density function is expressed as:

$$f_{k,m}(x) = \frac{1}{\sqrt{2\pi}\sigma_{k,m}} \exp\left[-\frac{\left(x - \mu_{k,m}\right)^2}{2\sigma_{k,m}^2}\right]$$

The E-step utilizes Bayes Theorem. Likelihood of data, given model is multiplied by prior value, alpha. Alpha is treated as a mixing proportion value.

$$p(k \mid x_n^1, \dots, x_n^M, p^{\text{old}}) = \frac{\alpha_k^{\text{old}} \prod_{m=1}^{M} f_{k,m}(x_n^m, p^{\text{old}})}{\sum_{\kappa=1}^{K} \alpha_\kappa^{\text{old}} \prod_{m=1}^{M} f_{\kappa,m}(x_n^m, p^{\text{old}})}$$

p is equal to the vector of required parameters, p = [$\mu_{1,1}$, . . . , $\mu_{1,M}$, . . . , $\mu_{K,1}$, . . .,$\mu_{K,M}$, $\sigma_{1,1}$, . . . , $\sigma_{1,M}$, . . . , $\sigma_{K,1}$, . . . , $\sigma_{K,M}$]

The M-step is used to update parameters α, μ and σ^2, according to the presented equations.

$$\alpha_k^{\text{new}} = \frac{\sum_{n=1}^{N} p\left(k \mid x_n^1, \dots, x_n^M, p^{\text{old}}\right)}{N}$$

$$\mu_{k,m}^{\text{new}} = \frac{\sum_{n=1}^{N} x_n^m p\left(k \mid x_n^1, \dots, x_n^M, p^{\text{old}}\right)}{\sum_{n=1}^{N} p(k \mid x_n^1, \dots, x_n^M, p^{\text{old}})}, \; k = 1,2, \dots, K$$

$$\left(\sigma_{k,m}^{\text{new}}\right)^2 = \frac{\sum_{n=1}^{N} \left(x_n - \mu_k^{\text{new}}\right)^2 p\left(k \mid x_n^1, \dots, x_n^M, p^{\text{old}}\right)}{\sum_{n=1}^{N} p\left(k \mid x_n^1, \dots, x_n^M, p^{\text{old}}\right)} k = 1,2,\dots,K, m = 1,\dots,M$$

The algorithm finishes its iterations when the absolute difference between old and new parameters is less than 1e-6.

13.2.5. Adjusted Rand Index

To determine efficiency of algorithms, we compared them using the Adjusted Rand Index (ARI) index. We have implemented the version that was proposed in [13].

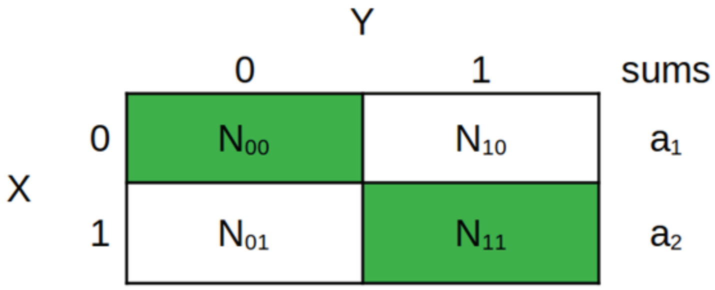
$$ARI = \frac{\sum_{ij} \binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

where:

where:	Y_1	Y_2	$\cdots$	Y_s	sums
X_1	n_{11}	n_{12}	$\cdots$	n_{1s}	a_1
X_2	n_{21}	n_{22}	$\cdots$	n_{2s}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
X_r	n_{r1}	n_{r2}	$\cdots$	n_{rs}	a_r
sums	b_1	b_2	$\cdots$	b_s	

13.2.6. Simple Matching Coefficient and its weighted version

To calculate Simple Matching Coefficient (SMC) and Weighted Simple Matching Coefficient (WSMC), we have used the Hungarian algorithm. However, instead of finding a minimal value for each row/column, we were looking for a maximal value. In this way, the maximal value was considered a true positive. The Simple Matching Coefficient is a straightforward metric. We divide the sum of true positives by the number of all values. The downside of this solution is that it does not impose any weights [10]. The green colour in Fig. 1 indicates matching groups.



		Y		
		0	1	sums
X	0	N_{00}	N_{10}	a_1
	1	N_{01}	N_{11}	a_2

Fig. 1. Cross-table of clustering results
Rys. 1. Tabela krzyżowa wyników grupowania

Simple Matching Coefficient equation:

$$\text{SMC} = \frac{N_{00} + N_{11}}{N_{00} + N_{11} + N_{01} + N_{10}}$$

To mitigate the problem with unequal groups, we can use the WSMC metric that addresses this issue:

$$\text{WSMC} = \frac{N_{00}}{a_1} + \frac{N_{11}}{a_2}$$

13.2.7. Averaged Jaccard index

The Jaccard index is a popular metric [4]. Here we are using its simple variation to take two and more clusters.

$$\text{J} = \left(\frac{N_{00}}{N_{11} + N_{01} + N_{10}} + \frac{N_{11}}{N_{00} + N_{01} + N_{10}}\right)/2$$

With the increased number of clusters, we have more terms in brackets and the denominator changes accordingly.

13.2.8. Binomial test

We used a binomial test to test hypotheses about correctly classifying cancer. It was possible since the classification result might have two outcomes: success or failure. The probability p was scaled accordingly to the number of clusters with $p = 1/k$.

13.3. Results and discussion

To show the efficacy of algorithms from diverse perspectives, we present results using various index values.

We start with the classical one, the Adjusted Rand Index.

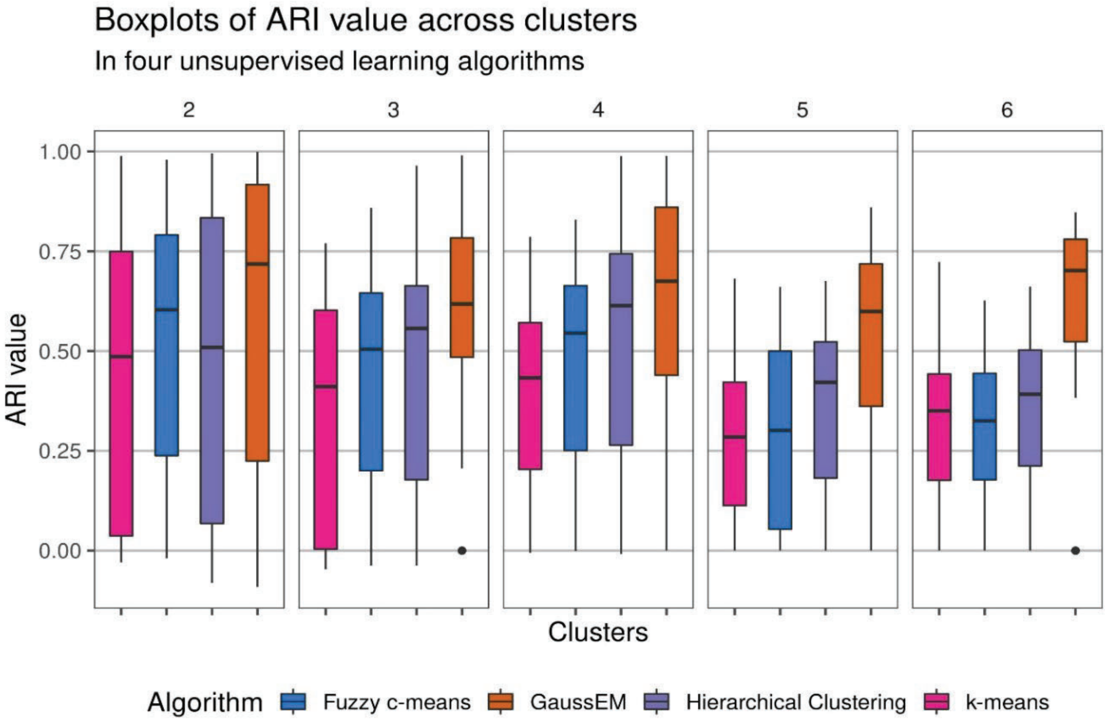


Fig. 2. Performance of four algorithms in various number of clusters
Rys. 2. Wydajność czterech algorytmów przy zmiennej ilości grup

In Fig. 2, in the case of HC, k-means and fuzzy c-means, we observe that the ARI index decreases with the increased number of groups. This trend is barely observed in the case of the GaussEM algorithm. Here, the median value oscillates between 0.74 and 0.60.

We can use violin plots to make the results more compact (Fig. 3). In addition to the values, violin plots show their density. The first three algorithms are denser. In the case of Gauss EM, the values are more concentrated in the upper parts of value indexes.

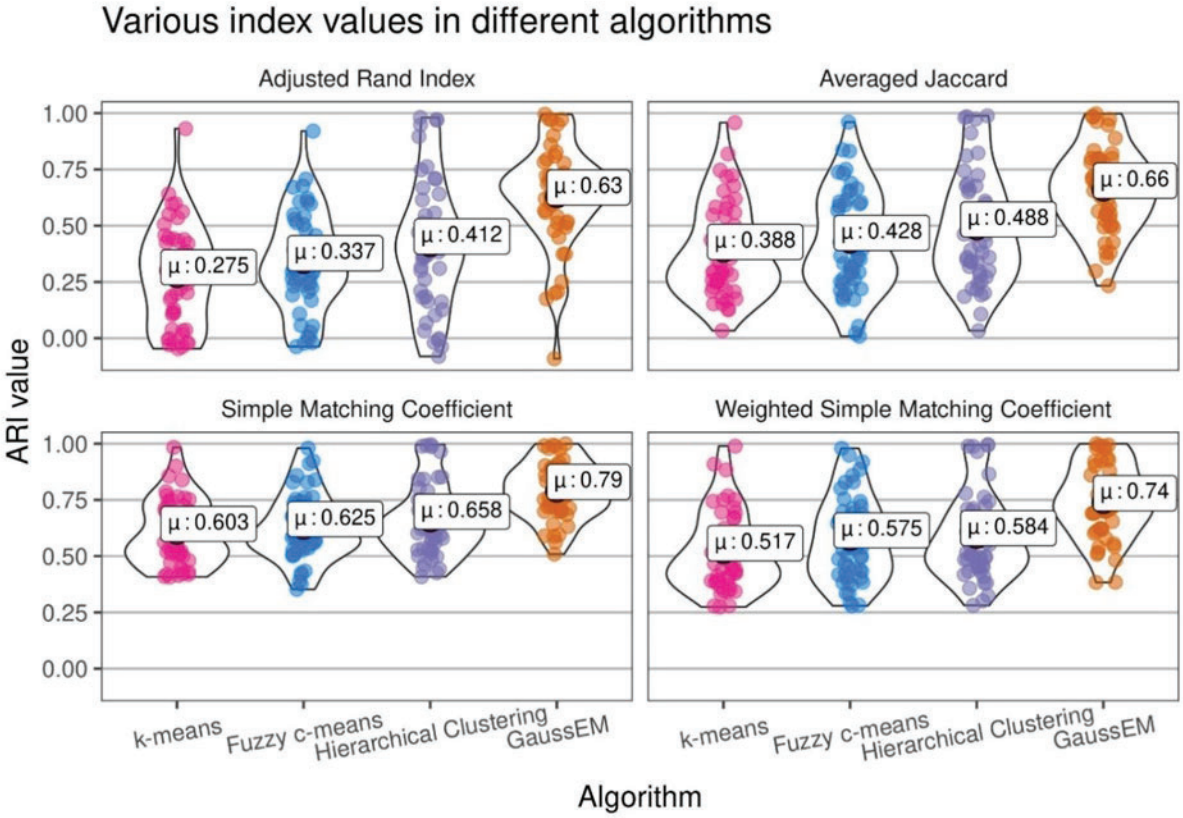


Fig. 3. Violin plots showing different quality measures
Rys. 3. Wykresy skrzypcowe pokazujące różne miary jakości

Fig. 4 shows results from the binomial test. The p-values are scaled to the power of 1/60 to make the results visible. We can see that majority of the values are below the p = 0.05. What is worth noticing is that even after raising results to the power of 1/60, the median value of GaussEM remains close to zero.

We compared the performance of four different unsupervised clustering methods that were based on distance metrics and maximum likelihood. For the comparison, we used the gene expression data from the TCGA portal. To assess the efficacy of the algorithms, we used various metrics, like Adjusted Rand Index, Jaccard, Simple Matching Coefficient, Weighted Simple Matching Coefficient, and binomial test. All of the compared algorithms showed statistically significant results. HAC performed as second best. Although computationally heavy, it is not very useful for big data in its original form.

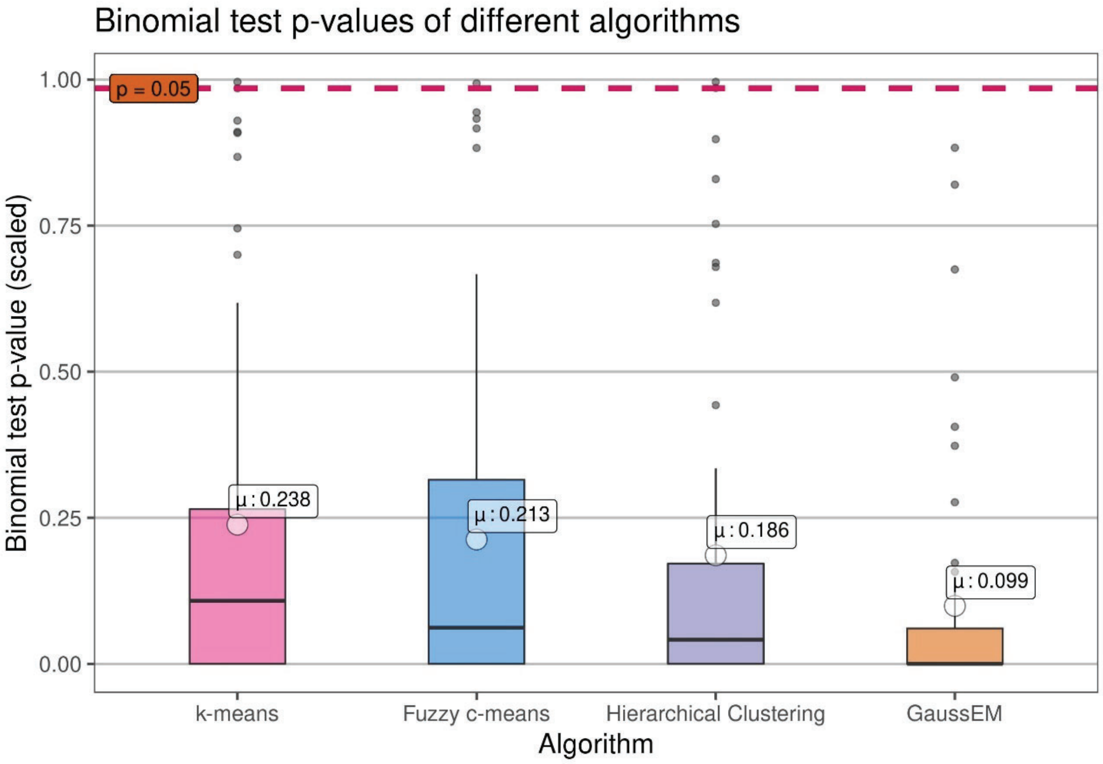


Fig. 4. Scores from the binomial test, raised to the power of 1/60
Rys. 4. Wyniki testu dwumianowego, podniesione do potęgi 1/60

Fuzzy c-means, a bridge between distance and maximum likelihood-based methods, performed better than k-means. Perhaps results might be improved using different parameters, e.g. fuzzyfied value. K-means scored the lowest both in the binomial test and other metrics. Although its simplicity, and low computational power, the requirement makes it still valuable, whether it is data exploration or starting point for the Expectation-Maximization methods. Finally, the multivariable Gaussian mixture with Expectation-Maximization algorithm obtained the highest score in all the presented metrics. It indicates that GaussEM is a good approach to finding gene expression patterns in human cancers.

Acknowledgment

This work is supported by the European Social Fund grant no. POWR.03.02.00-00-I029.

Bibliography

1. Arthur P. Dempster, Nan M. Laird, Donald B. Rubin: Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.
2. Jiarui Ding, Melissa K. McConechy, Hugo M. Horlings, Gavin Ha, Fong Chun Chan, Tyler Funnell, Sarah C Mullaly, Jueri Reimand, Ali Bashashati, Gary D. Bader, et al.: Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nature communications*, 6(1):1–13, 2015.
3. Paolo Giordani, Maria Brigida Ferraro, Francesca Martella: An Introduction to Clustering with R. *Springer*, 2020.
4. Lieve Hamers et al.: Similarity measures in scientometric research: The jaccard index versus salton's cosine formula. *Information Processing and Management*, 25(3):315–18, 1989.
5. Peilin Jia, Zhongming Zhao: Impacts of somatic mutations on gene expression: an association perspective. *Briefings in bioinformatics*, 18(3):413–425, 2017.
6. Leonard Kaufman, Peter J. Rousseeuw: Finding groups in data: an introduction to cluster analysis, volume 344. *John Wiley & Sons*, 2009.
7. Geofrey J. McLachlan, Thriyambakam Krishnan: The EM algorithm and extensions, volume 382. *John Wiley & Sons*, 2007.
8. Laurence Morissette, Sylvain Chartier: The k-means clustering technique: General considerations and implementation in mathematica. *Tutorials in Quantitative Methods for Psychology*, 9(1):15–24, 2013.
9. Joanna Polańska, Piotr Widłak, Joanna Rzeszowska-Wolny, Marek Kimmel, Andrzej Polański: Gaussian mixture decomposition of time-course DNA microarray data. In Mathematical Modeling of Biological Systems, Volume I, pages 351–359. *Springer*, 2007.
10. Jorge M. Santos, Mark Embrechts: On the use of the adjusted rand index as a metric for evaluating supervised classification. In International conference on artificial neural networks, pages 175–184. *Springer*, 2009.
11. Robert R. Sokal: A statistical method for evaluating systematic relationships. Univ. Kansas, *Sci. Bull*., 38:1409–1438, 1958.
12. Christopher Wilks, Melissa S. Cline, Erich Weiler, Mark Diehkans, Brian Craft, Christy Martin, Daniel Murphy, Howdy Pierce, John Black, Donavan Nelson, et al.: The cancer genomics hub (cghub): overcoming cancer through the power of torrential data. *Database*, 2014.
13. Lawrence Hubert and Phipps Arabie, Comparing partitions. *Journal of classification 2*(1):193–218, 1985.

UNSUPERVISED CLUSTERING FOR DETECTION OF GENE EXPRESSION PATTERNS IN HUMAN CANCERS

Abstract

In this study, we compare four unsupervised algorithms in the gene expression data of different human cancers. We based our analysis on openly available data from Cancer Genome Atlas (TCGA) (Genomic Data Commons) database. We tested two properties. The first is if there is a clear pattern in the gene expression data. The other was to select the algorithm which performs the best. Our results suggest that an expression pattern exists in different types of human cancer. As for the algorithm, the EM algorithm based on multivariate Gaussian mixtures showed the most promising performance.

Keywords: unsupervised clustering, GMM, k-means, HC, cancer, gene expression, EM

DOI:10.34918/85114

Maksymilian JĘDRZEJOWSKI[1], Ewa MAJCHRZAK[1,*]

Chapter 14. NUMERICAL ANALYSIS OF SKIN TUMOR FREEZING USING DUAL-PHASE LAG MODEL

14.1. Introduction

Low temperature can lead to cell necrosis. When the temperature is low enough to freeze water, ice crystals appear in extracellular spaces. Hence, a hyperosmotic extracellular environment is created, which draws water from cells, leading to the shrinkage of cells and destruction of their membranes [1]. As the temperature decreases, ice crystals appear within the cells. During thawing recrystallisation occurs, that is, ice crystals fuse and form larger crystals. When the ice melts, a hypotonic extracellular environment is created, and water flows back into cells. Their volume increases, leading to additional cell damage. Cryobiological research proved that any part of the freeze-thaw cycle may be injurious. Skin tumor (e.g. melanoma) is treated with cylindrical cryoprobe, which is applied directly to its surface (Fig. 1).

Rapid cooling is more destructive, so the cooling rate should be as high as possible [1]. Experimental data vary from 22°C / min to 50°C / min to even 220°C / min [1, 2]. However, in vitro research reports significant cell damage in a cooling ratio of as low as 1°C / min [3]. It should be noted that rapid cooling appears only near the cryoprobe, and the cooling rate is reduced as the distance from the probe increases. Taking these variances into account, the cooling rate has a low impact on injury production [1].

In cancer treatment, there is a difference between lethal and surely lethal tissue temperature. The lethal tissue temperature can be achieved in the range of –15°C to –20°C [1, 4, 5]. The cell survival rate should be minimalised; therefore, the surely

---

[1] Department of Computational Mechanics and Engineering, Faculty of Mechanical Engineering, Silesian University of Technology, Gliwice, Poland.
* Corresponding author: ewa.majchrzak@polsl.pl.

lethal tissue temperature can be established below –40°C [1, 4, 5]. In some experiments, the lethal dose is defined as low as –60°C with the requirement of repetitive freeze-thaw cycles [6]. However, the temperature –40°C nowadays is defined as a lethal dose in a variety of research, connecting this temperature with the physics of water (the crystal growth is highest in the range from 0°C to –40°C [1]).

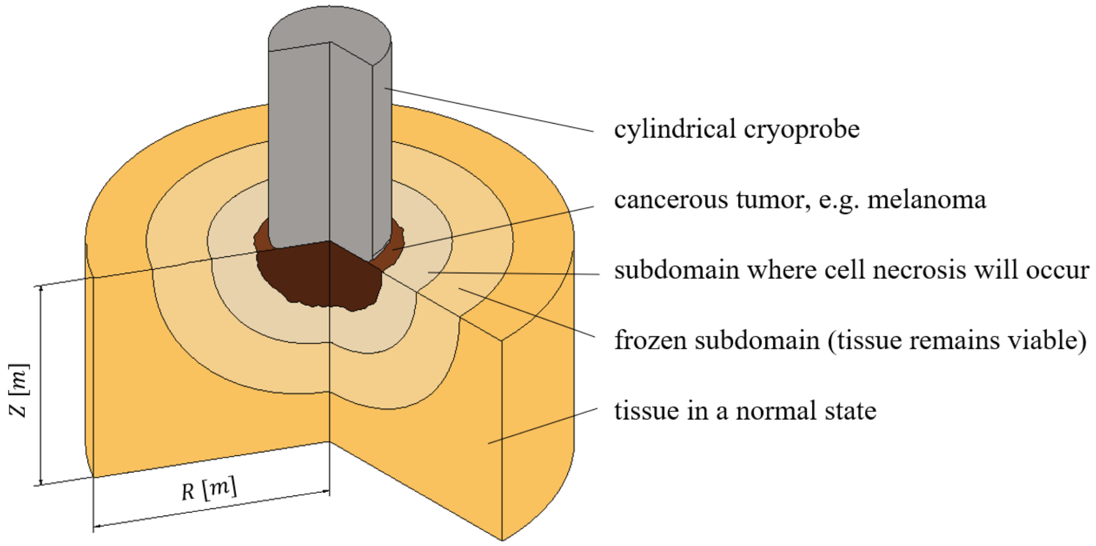


Fig. 1. Considered domain
Rys. 1. Rozważany obszar

The thawing rate is a key destructive factor. Cell damage increases greatly with prolonged thawing and the thawing rate should be as low as 1°C / min [7]. Rapid thawing increases cell survival ratio very highly. In fact, slow thawing is more important in cell destruction than rapid cooling [1, 7, 8].

Repetition of the freeze-thaw cycle increases previously frozen volume and intensifies cell necrosis. For example, in the treatment of facial basal cell carcinomas, the cure rate for the double freeze-thaw cycle is 95.3% compared to only 79.4% for the single freeze-thaw cycle [9]. The effect of repeated freeze-thaw cycles has a greater impact on cure rates when the freeze temperature is relatively warm, e.g. –20°C.

Mathematical modelling methods are extremely helpful in planning cryosurgical procedures. They allow for the analysis of many variants of freezing with different cooling rates, number of freeze-thaw cycles, and different durations of the process. In this paper, the freezing process is described by the dual-phase lag equation supplemented with appropriate boundary and initial conditions. The problem is solved using the finite difference method both in the explicit and implicit schemes. Using an in-house computer program, the temperature distribution in the domain analysed is calculated, the subdomain of the frozen region is determined, and the effectiveness of freezing the tumor is estimated.

14.2. Mathematical model

In 1995 Tzou [10] proposed the introduction of two time lags in Fourier law, representing a macroscopic lag (delayed response) between the temperature gradient and the heat flux as a result of microstructural effects, governing the following formula

$$\boldsymbol{q}\left(\mathbf{x}, t+\tau_q\right) = -\lambda \nabla T\left(\mathbf{x}, t+\tau_T\right) \tag{1}$$

where $\boldsymbol{q}$ is the heat flux vector, $\mathbf{x} = \{r, z\}$ denotes geometrical coordinates, t is the time, λ is thermal conductivity, τ_q and τ_T are the phase lags (called relaxation and thermalization time respectively).

The following first-order (linear) approximation can be obtained with the use of Taylor series expansions

$$\boldsymbol{q}(\mathbf{x}, t) + \tau_q \frac{\partial \boldsymbol{q}(\mathbf{x}, t)}{\partial t} \cong -\lambda \left[\nabla T(\mathbf{x}, t) + \tau_T \frac{\partial \nabla T(\mathbf{x}, t)}{\partial t}\right] \tag{2}$$

Both phase lags are assumed to be small, so nonlinear orders are negligible [10].

Introducing formula (2) to the well known macroscopic energy equation one obtains [10, 11]

$$c\left[\frac{\partial T(\mathbf{x}, t)}{\partial t} + \tau_q \frac{\partial^2 T(\mathbf{x}, t)}{\partial t^2}\right] = \nabla[\lambda \nabla T(\mathbf{x}, t)] + \tau_T \nabla\left[\lambda \frac{\partial \nabla T(\mathbf{x}, t)}{\partial t}\right] \tag{3}$$

where c is the volumetric specific heat.

In biological tissue, internal heat sources should also be considered, which gives the following dual-phase lag equation (DPLE) [11]

$$\begin{aligned} &c\left[\frac{\partial T(\mathbf{x},t)}{\partial t}+\tau_q\frac{\partial^2 T(\mathbf{x},t)}{\partial t^2}\right] \\ &\qquad=\nabla[\lambda\nabla T(\mathbf{x},t)]+\tau_T\nabla\left[\lambda\frac{\partial \nabla T(\mathbf{x},t)}{\partial t}\right]+Q(\mathbf{x},t)+\tau_q\frac{\partial Q(\mathbf{x},t)}{\partial t}+Q_F(\mathbf{x},t) \\ &\qquad+\tau_q\frac{\partial Q_F(\mathbf{x},t)}{\partial t} \end{aligned} \tag{4}$$

Internal heat source $Q(\mathbf{x},t)$, related to blood perfusion and metabolism, is defined as the following sum

$$Q(\mathbf{x},t)=wc_b[T_a-T(\mathbf{x},t)]+Q_{met}(T) \tag{5}$$

where w is the blood perfusion rate, c_b is the specific heat of the blood, T_a is the arterial blood temperature, Q_{met} is the metabolic heat source.

The internal volumetric heat source $Q_F(\mathbf{x},t)$, related to the phase change, is defined as

$$Q_F(\mathbf{x},t)=L\frac{\partial S(\mathbf{x},t)}{\partial t}=L\frac{\mathrm{d}S(T)}{\mathrm{d}T}\frac{\partial T(\mathbf{x},t)}{\partial t} \tag{6}$$

where L is the volumetric latent heat of freezing and $S(\mathbf{x},t)$ is the frozen state fraction.

After determining the derivatives and some transformations, one can obtain

$$\begin{aligned} &Q(\mathbf{x},t)+\tau_q\frac{\partial Q(\mathbf{x},t)}{\partial t}+Q_F(\mathbf{x},t)+\tau_q\frac{\partial Q_F(\mathbf{x},t)}{\partial t} \\ &\qquad=\tau_q\left[\frac{\mathrm{d}w(T)}{\mathrm{d}T}c_b(T_a-T)-w(T)c_b+\frac{\mathrm{d}Q_{met}(T)}{\mathrm{d}T}\right]\frac{\partial T}{\partial t}+w(T)c_b(T_a-T) \\ &\qquad+Q_{met}(T)+L\frac{\mathrm{d}S(T)}{\mathrm{d}T}\frac{\partial T}{\partial t}+\tau_q L\left[\frac{\mathrm{d}^2S(T)}{\mathrm{d}T^2}\left(\frac{\partial T}{\partial t}\right)^2+\frac{\mathrm{d}S(T)}{\mathrm{d}T}\frac{\partial^2 T}{\partial t^2}\right] \end{aligned} \tag{7}$$

and by denoting the derivative of the blood perfusion rate and the derivative of the metabolic heat as

$$v(T)=\frac{\mathrm{d}w(T)}{\mathrm{d}T}, \qquad P_{met}(T)=\frac{\mathrm{d}Q_{met}(T)}{\mathrm{d}T} \tag{8}$$

the dual-phase lag equation (4) can be written as follows

$$\begin{aligned} &\left[C(T)+\tau_q\{w(T)c_b-v(T)c_b(T_a-T)-P_{met}(T)\}\right]\frac{\partial T}{\partial t}+\tau_q C(T)\frac{\partial^2 T}{\partial t^2}+\tau_q\frac{\mathrm{d}C(T)}{\mathrm{d}T}\left(\frac{\partial T}{\partial t}\right)^2 \\ &\qquad=\nabla(\lambda(T)\nabla \mathrm{T})+\tau_T\nabla\left[\lambda(T)\nabla\left(\frac{\partial T}{\partial t}\right)\right]+w(T)c_b(T_a-T)+Q_{met}(T) \end{aligned} \tag{9}$$

where C is a substitute thermal capacity of the intermediate region [11]

$$C(T) = c(T) - L\frac{\mathrm{d}S(T)}{\mathrm{d}T} \tag{10}$$

In this paper, the following functions have been assumed that describe the dependence of the blood perfusion rate and the metabolic heat source on the temperature

$$w(T) = \begin{cases} w_0 & T > T_1 \\ w_0\dfrac{T - T_2}{T_1 - T_2} & T_2 \le T \le T_1 \\ 0 & T < T_2 \end{cases} \tag{11}$$

$$Q_{met}(T) = \begin{cases} Q_{met0} & T > T_1 \\ Q_{met0}\dfrac{T - T_2}{T_1 - T_2} & T_2 \le T \le T_1 \\ 0 & T < T_2 \end{cases} \tag{12}$$

where T_1 is the beginning temperature of freezing and T_2 is the ending temperature of freezing. If part of the tissue is frozen and thawed, then the blood perfusion rate and the metabolic heat source are permanently equal to zero due to the destruction of blood vessels in that region [11]. The derivatives of these two parameters are given as follows

$$v(T) = \begin{cases} 0 & T > T_1 \\ w_0\dfrac{1}{T_1 - T_2} & T_2 \le T \le T_1 \\ 0 & T < T_2 \end{cases} \tag{13}$$

$$P_{met}(T) = \begin{cases} 0 & T > T_1 \\ Q_{met0}\dfrac{1}{T_1 - T_2} & T_2 \le T \le T_1 \\ 0 & T < T_2 \end{cases} \tag{14}$$

where w_0 and Q_{met0} are the values determined for the tissue in its natural state.

In the temperature range $[T_2, T_1]$ a linear function was assumed to describe the frozen state fraction

$$S(T) = \frac{T_1 - T}{T_1 - T_2} \tag{15}$$

Therefore, the substitute thermal capacity can be written as

$$C = \begin{cases} c_N & T > T_1 \\ c_P + \dfrac{L}{T_1 - T_2} & T_2 \le T \le T_1 \\ c_F & T < T_2 \end{cases} \tag{16}$$

where c_N, c_P, and c_F denote the thermal capacity of tissue in its natural, intermediate, and frozen state, respectively. Substitute thermal capacity (eq. (16)) in this article is a step function, thus eq. (9) can be written in the following form

$$\begin{gathered}\left[C(T)+\tau_q\{w(T)c_b-v(T)c_b(T_a-T)-P_{met}(T)\}\right]\frac{\partial T}{\partial t}+\tau_q C(T)\frac{\partial^2 T}{\partial t^2}\\=\nabla(\lambda(T)\nabla \mathrm{T})+\tau_T\nabla\left[\lambda(T)\nabla\left(\frac{\partial T}{\partial t}\right)\right]+w(T)c_b(T_a-T)+Q_{met}(T)\end{gathered}\tag{17}$$

Also, the thermal conductivity is a step function given as

$$\lambda=\begin{cases}\lambda_N & T>T_1\\ \lambda_P & T_2\le T\le T_1\\ \lambda_F & T<T_2\end{cases}\tag{18}$$

where λ_N, λ_P, λ_F denote the thermal conductivity of the tissue, in its natural, intermediate and frozen state, respectively.

On the contact surface between the tip of the cryoprobe and the skin tissue, the Dirichlet condition is assumed

$$T=T_D(t)\tag{19}$$

and on the other surfaces the adiabatic boundary condition is assumed

$$-\lambda\left[\boldsymbol{n}\cdot\nabla T(\mathbf{x},t)+\tau_T\frac{\partial[\boldsymbol{n}\cdot\nabla T(\mathbf{x},t)]}{\partial t}\right]=0\tag{20}$$

where $\boldsymbol{n}\cdot\nabla T(\mathbf{x},t)$ is the normal derivative.

The initial conditions are given as follows

$$t=0:\quad T=T_p,\quad \left.\frac{\partial T}{\partial t}\right|_{t=0}=u\tag{21}$$

where T_p is the initial temperature and u is the initial cooling rate.

As mentioned earlier, the phase lags τ_q and τ_T are small constant values [10]. DPLM can be reduced to thermal wave equations (hyperbolic model) for $\tau_T = 0$ s [10, 12]. The classic Fourier equation (parabolic model) can be obtained by substituting $\tau_q = \tau_T = 0$ s. It should be noted that the dual-phase effect in biological tissue is important even for $\tau_q \approx \tau_T$ [13].

14.3. Numerical model

In a cylindrical coordinate system eq. (17) can be written as

$$\begin{aligned}\left[C(T)+\tau_q\{w(T)c_b-v(T)c_b(T_a-T)-P_{met}(T)\}\right]\frac{\partial T}{\partial t}+\tau_q C(T)\frac{\partial^2 T}{\partial t^2}\\ =\frac{1}{r}\lambda(T)\left[\frac{\partial T}{\partial r}+\tau_T\frac{\partial}{\partial r}\left(\frac{\partial T}{\partial t}\right)\right]+\lambda(T)\left[\frac{\partial^2 T}{\partial r^2}+\tau_T\frac{\partial^2}{\partial r^2}\left(\frac{\partial T}{\partial t}\right)\right]\\ +\lambda(T)\left[\frac{\partial^2 T}{\partial z^2}+\tau_T\frac{\partial^2}{\partial z^2}\left(\frac{\partial T}{\partial t}\right)\right]+w(T)c_b(T_a-T)+Q_{met}(T)\end{aligned} \tag{22}$$

Numerical computations are carried out using the finite difference method (FDM) [14]. The time grid with the constant time step Δt is introduced. The quadratic mesh with the constant grid step h is applied.

Using the explicit FDM scheme eq. (22) can be written as

$$T_{i,j}^{f}=\frac{M_{i,j}^{f-1}}{K_{i,j}^{f-1}}T_{i,j}^{f-1}+\frac{P_{i,j}^{f-1}}{K_{i,j}^{f-1}}T_{i,j+1}^{f-1}+\frac{R_{i,j}^{f-1}}{K_{i,j}^{f-1}}T_{i,j-1}^{f-1}+\frac{S_{i,j}^{f-1}}{K_{i,j}^{f-1}}\left(T_{i-1,j}^{f-1}+T_{i+1,j}^{f-1}\right)+\frac{W_{i,j}^{f-1}}{K_{i,j}^{f-1}} \tag{23}$$

where

$$\begin{aligned}K_{i,j}^{f-1}&=\frac{\left[C_{i,j}^{f-1}+\tau_q\left\{w_{i,j}^{f-1}c_b-v_{i,j}^{f-1}c_b\left(T_a-T_{i,j}^{f-1}\right)-{P_{met}}_{i,j}^{f-1}\right\}\right]\Delta t+\tau_q C_{i,j}^{f-1}}{(\Delta t)^2}\\ M_{i,j}^{f-1}&=\frac{\left[C_{i,j}^{f-1}+\tau_q\left\{w_{i,j}^{f-1}c_b-v_{i,j}^{f-1}c_b\left(T_a-T_{i,j}^{f-1}\right)-{P_{met}}_{i,j}^{f-1}\right\}\right]\Delta t+2\tau_q C_{i,j}^{f-1}}{(\Delta t)^2}\\ &\quad-4\frac{\lambda_{i,j}^{f-1}(\Delta t+\tau_T)}{h^2\Delta t}-w_{i,j}^{f-1}c_b\\ P_{i,j}^{f-1}&=\frac{\lambda_{i,j}^{f-1}(\Delta t+\tau_T)}{h^2\Delta t}+\frac{\lambda_{i,j}^{f-1}(\Delta t+\tau_T)}{2hr_{i,j}\Delta t}\\ R_{i,j}^{f-1}&=\frac{\lambda_{i,j}^{f-1}(\Delta t+\tau_T)}{h^2\Delta t}-\frac{\lambda_{i,j}^{f-1}(\Delta t+\tau_T)}{2hr_{i,j}\Delta t}\\ S_{i,j}^{f-1}&=\frac{\lambda_{i,j}^{f-1}(\Delta t+\tau_T)}{h^2\Delta t}\\ W_{i,j}^{f-1}&=\left[4\frac{\lambda_{i,j}^{f-1}\tau_T}{h^2\Delta t}-\frac{\tau_q C_{i,j}^{f-1}}{(\Delta t)^2}\right]T_{i,j}^{f-2}+\left[-\frac{\lambda_{i,j}^{f-1}\tau_T}{2hr_{i,j}\Delta t}-\frac{\lambda_{i,j}^{f-1}\tau_T}{h^2\Delta t}\right]T_{i,j+1}^{f-2}\\ &\quad+\left[\frac{\lambda_{i,j}^{f-1}\tau_T}{2hr_{i,j}\Delta t}-\frac{\lambda_{i,j}^{f-1}\tau_T}{h^2\Delta t}\right]T_{i,j-1}^{f-2}+\left[-\frac{\lambda_{i,j}^{f-1}\tau_T}{h^2\Delta t}\right]\left(T_{i-1,j}^{f-2}+T_{i+1,j}^{f-2}\right)+w_{i,j}^{f-1}c_bT_a\\ &\quad+{Q_{met}}_{i,j}^{f-1}\end{aligned} \tag{24}$$

and using the FDM implicit scheme eq. (22) can be written as

$$T_{i,j}^{f} = \frac{B_{i,j}^{f-1}}{A_{i,j}^{f-1}} T_{i,j-1}^{f} + \frac{D_{i,j}^{f-1}}{A_{i,j}^{f-1}} T_{i,j+1}^{f} + \frac{F_{i,j}^{f-1}}{A_{i,j}^{f-1}} \left(T_{i-1,j}^{f} + T_{i+1,j}^{f} \right) + \frac{G_{i,j}^{f-1}}{A_{i,j}^{f-1}} \tag{25}$$

where

$$\begin{aligned}
A_{i,j}^{f-1} &= \frac{\left[C_{i,j}^{f-1} + \tau_q \left\{ w_{i,j}^{f-1} c_b - v_{i,j}^{f-1} c_b \left(T_a - T_{i,j}^{f-1} \right) - P_{met}{}_{i,j}^{f-1} \right\} \right] \Delta t + C_{i,j}^{f-1} \tau_q}{(\Delta t)^2} + w_{i,j}^{f-1} c_b \\
&\quad + 4 \frac{\lambda_{i,j}^{f-1} (\Delta t + \tau_T)}{h^2 \Delta t} \\
B_{i,j}^{f-1} &= \frac{\lambda_{i,j}^{f-1} (\Delta t + \tau_T)}{h^2 \Delta t} - \frac{\lambda_{i,j}^{f-1} (\Delta t + \tau_T)}{2 h r_{i,j} \Delta t} \\
D_{i,j}^{f-1} &= \frac{\lambda_{i,j}^{f-1} (\Delta t + \tau_T)}{h^2 \Delta t} + \frac{\lambda_{i,j}^{f-1} (\Delta t + \tau_T)}{2 h r_{i,j} \Delta t} \\
F_{i,j}^{f-1} &= \frac{\lambda_{i,j}^{f-1} (\Delta t + \tau_T)}{h^2 \Delta t} \\
G_{i,j}^{f-1} &= \left[\frac{\left[C_{i,j}^{f-1} + \tau_q \left\{ w_{i,j}^{f-1} c_b - v_{i,j}^{f-1} c_b \left(T_a - T_{i,j}^{f-1} \right) - P_{met}{}_{i,j}^{f-1} \right\} \right] \Delta t + 2 C_{i,j}^{f-1} \tau_q}{(\Delta t)^2} \right] T_{i,j}^{f-1} \\
&\quad - \frac{\lambda_{i,j}^{f-1} \tau_T}{h^2 \Delta t} \left(T_{i,j-1}^{f-1} + T_{i,j+1}^{f-1} + T_{i-1,j}^{f-1} + T_{i+1,j}^{f-1} - 4 T_{i,j}^{f-1} \right) \\
&\quad - \frac{\lambda_{i,j}^{f-1} \tau_T}{2 h r_{i,j} \Delta t} \left(T_{i,j+1}^{f-1} - T_{i,j-1}^{f-1} \right) - \frac{C_{i,j}^{f-1} \tau_q}{(\Delta t)^2} T_{i,j}^{f-2} + w_{i,j}^{f-1} c_b T_a + Q_{met}{}_{i,j}^{f-1}
\end{aligned} \tag{26}$$

The system of linear equations in the implicit scheme is solved iteratively.

Both schemes presented must be supplemented by boundary conditions. In an explicit scheme, stability condition must be fulfilled [15]. The implicit scheme of the finite difference method for the DPLE is always stable [16].

14.4. Results of computations

The cylindrical tissue domain of dimensions R = 0.04 m, Z = 0.04 m at initial temperature T_p = 37°C is considered. The tip of the cryoprobe of diameter D = 0.025 m is subjected to the skin. Thermophysical parameters are as follows: thermal conductivities λ_N = 0.52 W/(m K), λ_P = 1.26 W/(m K), λ_F = 2 W/(m K), volumetric specific heats c_N = 3.6 MJ/(m^3 K), c_P = 2.78 MJ/(m^3 K), c_F = 1.93 MJ/(m^3 K), blood perfusion rate in natural state w_0 = 0.5 kg/(m^3 s), metabolic heat source in natural state Q_{met0} = 245 W/m^3, specific heat of blood c_b = 3770 J/(kg K), arterial blood temperature T_a = 37°C, volumetric latent heat of freezing L = 330 MJ/m^3, intermediate zone [–8°C, –1°C], that is T_2 = –8°C and T_1 = –1°C [11].

Stage III malignant melanoma is considered. The diameter of tumor infiltration is 30 mm (approximately 20 mm of the main tumor accompanied by satellite tumors) and its depth is 5 mm [17, 18]. The initial cryoprobe temperature is 37°C. The final temperature of the cryoprobe is –160°C and four freeze-thaw cycles are considered [19]. The cooling rate is set at 120°C / min and the thawing rate is set at 5°C / min. After the cooling stage, the constant temperature is maintained for 5 min. The dependence of the cryoprobe temperature on time is shown in Fig. 2.

Four cases of phase lags values were considered. Case 1 in which τ_q = 0 s and τ_T = 0 s represents the Pennes model. Case 2 in which τ_q = 0.48 s and τ_T = 0.48 s is based on the results published in [13]. Case 3 in which τ_q = 3 s and τ_T = 0.1 s is discussed in more detail in [11, 12]. Case 4 in which τ_q = 15 s and τ_T = 10 s is taken from [20].

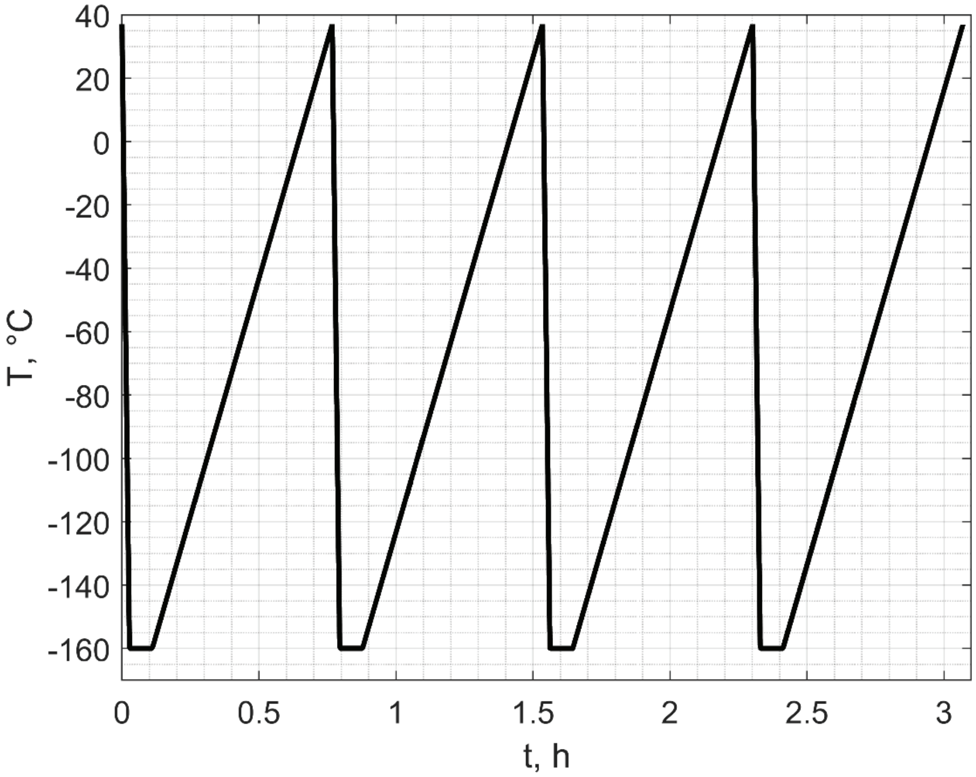


Fig. 2. Temperature of the cryoprobe
Rys. 2. Temperatura kriosondy

The influence of delay times on the obtained results was tested on the first freezing (Fig. 3). Explicit (solid line) and implicit (dashed line) schemes were compared. The results were obtained for: the number of nodes 51 × 51, grid step h = 0.0008 m, time step Δt = 0.05 s. The results are similar for both schemes for all cases. The differences between the Pennes model and the DPLM are most notable near the axis of symmetry and near the cryoprobe.

The next calculations were performed using the implicit scheme of the finite difference method under the assumption that the time step is equal to Δt = 1 s and phase lags are equal to τ_q = 3 s, τ_T = 0.1 s, respectively (case 3). As shown in Fig. 4, this time step gives results similar to those for Δt = 0.1 s and Δt = 0.05 s, but the computation time is much shorter.

In Fig. 5–7 the temperature distributions after first, second, and fourth freezing are shown. The red dashed line marks the area of tumor infiltration.

Fig. 8 presents a temperature history at the selected points. Two points along the axis were chosen, r = 0 m, z = 0.004 m (blue line), r = 0 m, z = 0.008 m (red line), and one point near the necrosis temperature border, r = 0.008 m, z = 0.008 m (orange line). The temperature history is similar for each freezing.

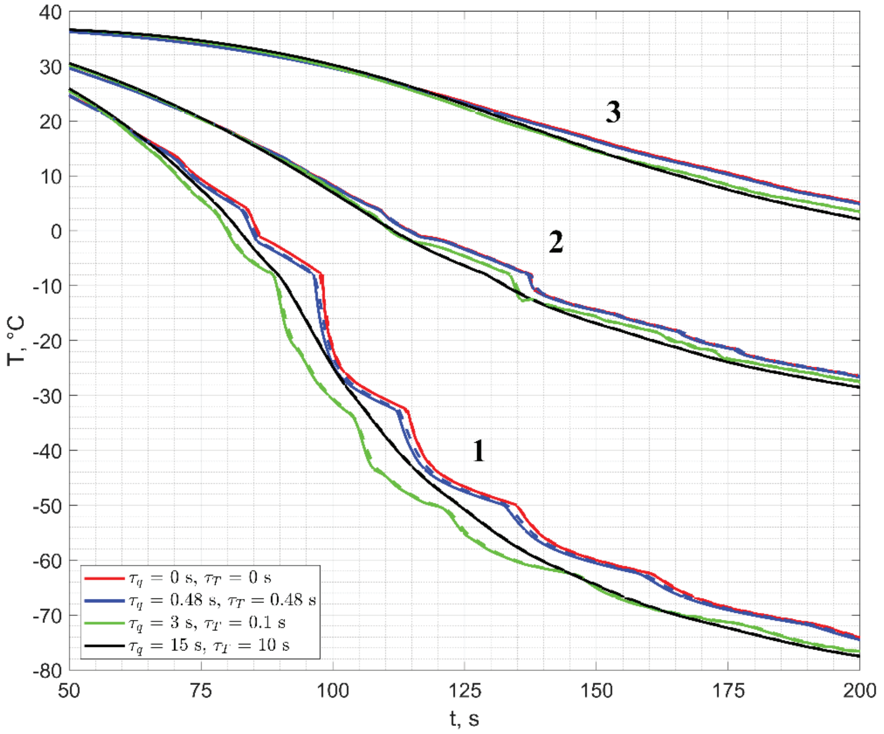


Fig. 3. Temperature history at the points 1 (0, 0.004 m), 2 (0.012 m, 0.004 m), 3 (0.008 m, 0.008 m) from 50 s to 200 s

Rys. 3. Historia temperatury w punktach 1 (0, 0.004 m), 2 (0.012 m, 0.004 m), 3 (0.008 m, 0.008 m) od 50 s do 200 s

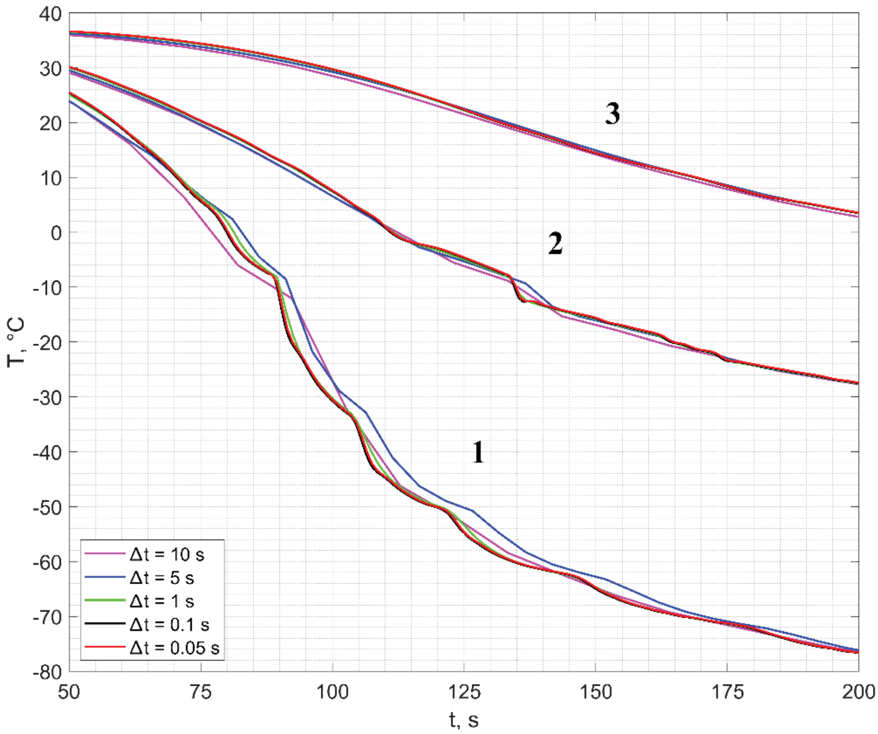


Fig. 4. Temperature history points 1 (0, 0.004 m), 2 (0.012 m, 0.004 m), 3 (0.008 m, 0.008 m) depending on the time step in the implicit scheme for case 3, from 50 s to 200 s

Rys. 4. Krzywe stygnięcia w punktach 1 (0, 0.004 m), 2 (0.012 m, 0.004 m), 3 (0.008 m, 0.008 m) w zależności od kroku czasowego w schemacie niejawnym dla przypadku 3, od 50 s do 200 s

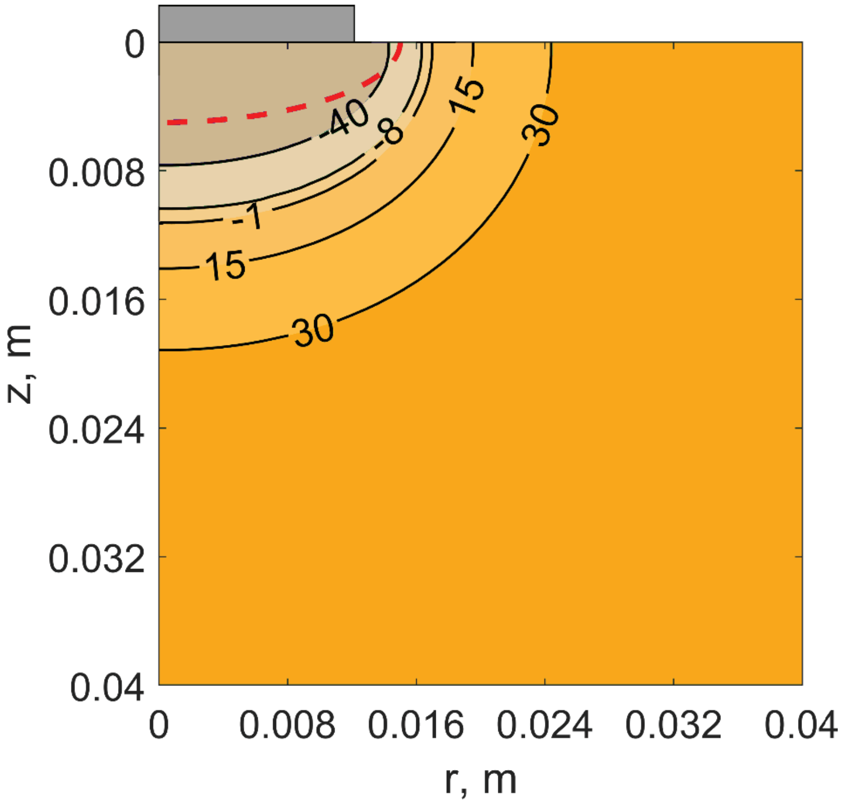


Fig. 5. Temperature distribution after first freezing (t = 6 min 40 s)

Rys. 5. Rozkład temperatur po pierwszym zamrażaniu (t = 6 min 40 s)

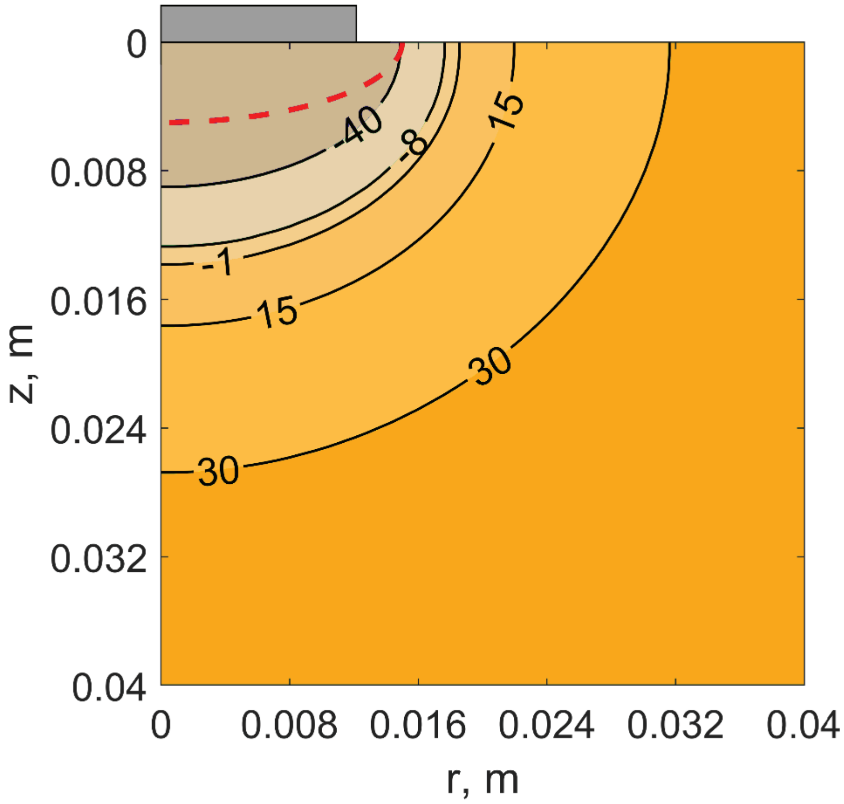


Fig. 6. Temperature distribution after second freezing (t = 52 min 42 s)
Rys. 6. Rozkład temperatur po drugim zamrażaniu (t = 52 min 42 s)

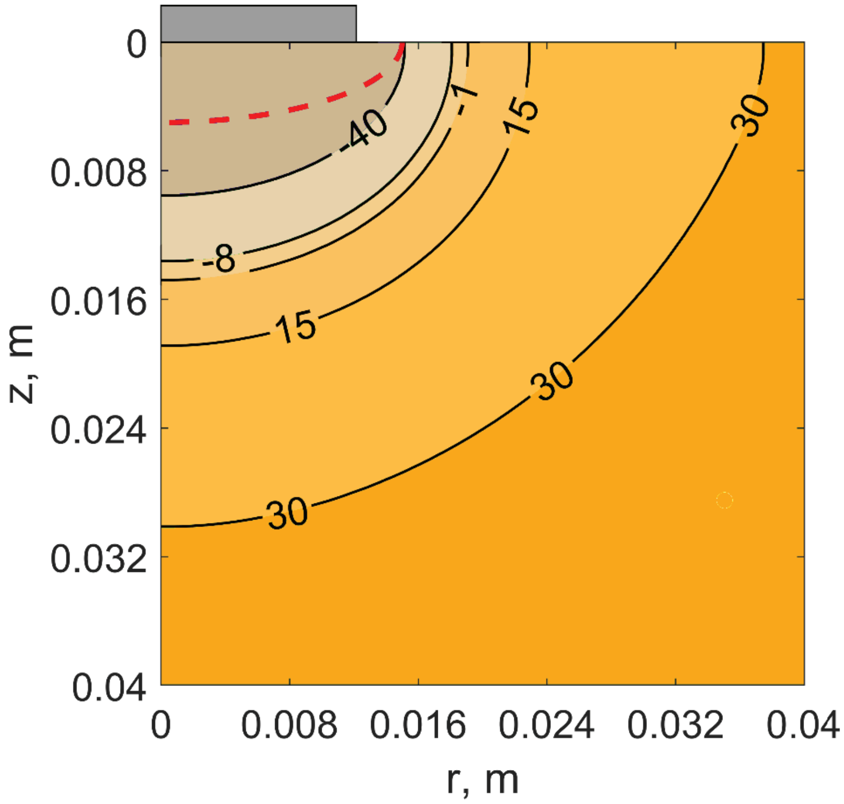


Fig. 7. Temperature distribution after fourth freezing (t = 2 h 24 min 46 s)
Rys. 7. Rozkład temperatur po czwartym zamrażaniu (t = 2 h 24 min 46 s)

After the third and fourth freezing, a slight increase in the frozen subdomain and the necrosis subdomain is observed (Tab. 1).

The numerical analysis carried out shows that for the considered malignant melanoma, a fourth (and possibly a third) freezing during the cryosurgical procedure is not necessary.

Table 1

Frozen and necrosis volume after each freezing

After freezing	Frozen volume (below –8°C) cm^3	Necrosis volume (below –40°C) cm^3	Maximum temperature inside the tumor, °C	Is the entire tumor below the lethal dose?
1	6.17	3.45	–32.1	N
2	8.46	4.29	–42.5	Y
3	9.32	4.59	–45.3	Y
4	9.66	4.71	–46.4	Y

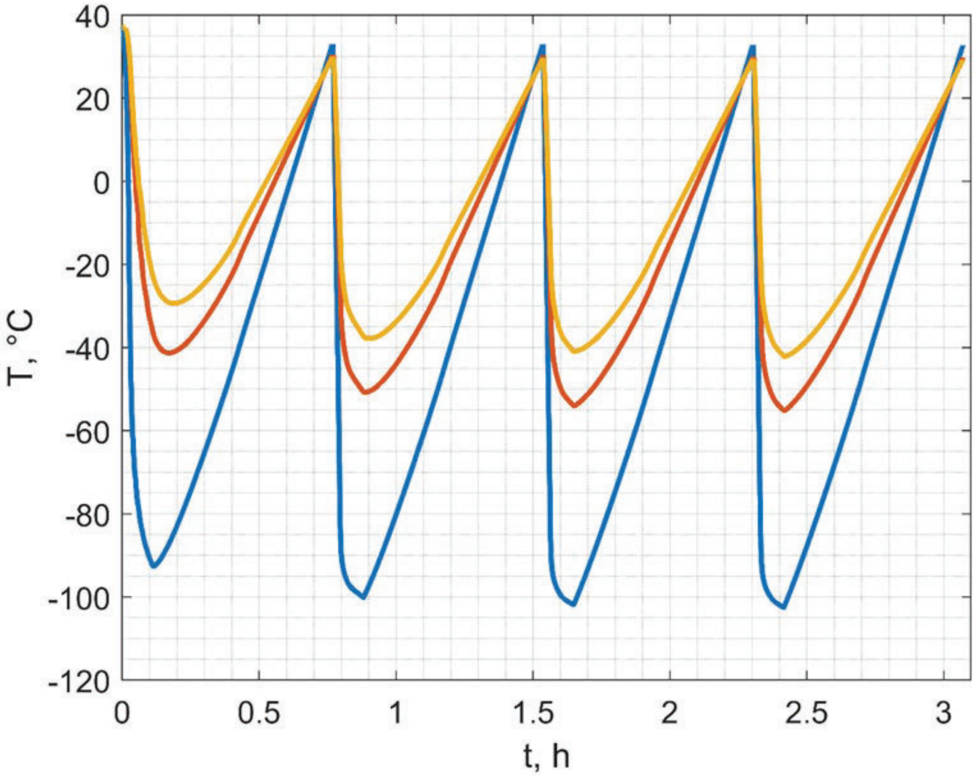


Fig. 8. Temperature history at selected points
Rys. 8. Krzywe stygnięcia/nagrzewania w wybranych punktach

14.5. Conclusions

Numerical modeling of biological tissue freezing can be used as a supporting tool in planning the cryosurgery treatment procedure. Using numerical modeling, one can select the appropriate cryoprobe diameter, tip temperature, freezing and thawing time, cooling rates, number of freezes, predict the size of the necrosis area and determine whether the lethal temperature can reach the entire tumor. The frozen subdomain can also be determined, allowing one to estimate the risk of freezing surrounding healthy tissue.

Bibliography

1. A.A. Gage, J.G. Baust: Mechanisms of tissue injury in cryosurgery, *Cryobiology* (1998) **37(3)**:171–186.
2. B. Rubinsky, C. Lee, J. Bastacky, G. Onik: The process of freezing and the mechanism of damage during hepatic cryosurgery, *Cryobiology* (1990) **27(1)**:85–97.
3. A.M. Granov, D.G. Prokhorov, A.P. Andreev, G.P. Pinaev, G.G. Prokhorov, A.V. Vlasova: Temperature measuring and evaluation of tumor cell viability in different zones of an ice ball. Practical application of in vitro experimental results. In *Basics of Cryosurgery* (Vienna, Austria, 2001) pp. 15–27.
4. J.G. Baust, A.A. Gage, T.E. Bjerklund Johansen, J.M. Baust Mechanisms of cryoablation: clinical consequences on malignant tumors, *Cryobiology* (2014) **68(1)**:1–11.
5. J.G. Baust, A.A. Gage: Progress toward optimization of cryosurgery, *Technology in Cancer Research & Treatment* (2004) **3(2)**:95–101.
6. H.B. Neel, A.S. Ketcham, W.G. Hammond: Requisites for successful cryogenic surgery of cancer, *Archives of Surgery* (1971) **102(1)**:45–48.
7. W.B. Bald, J. Fraser: Cryogenic surgery, *Reports on Progress in Physics* (1982) **45(12)**:1381–1484.
8. A.A. Gage, K. Guest, M. Montes, J.A. Caruana, D.A. Whalen Jr: Effect of varying freezing and thawing rates in experimental cryosurger, *Cryobiology* (1985) **22(2)**: 175–182.

9. E. Mallon, R. Dawber: Cryosurgery in the treatment of basal cell carcinoma: assessment of one and two freeze-thaw cycle schedules, *Dermatologic surgery* (1996) **22(10)**:854–858.
10. D.Y. Tzou: A unified field approach for heat conduction from macro-to micro-scales, *Journal of Heat Transfer* (1995) **117(1)**: 8–16.
11. B. Mochnacki, E. Majchrzak: Numerical model of thermal interactions between cylindrical cryoprobe and biological tissue using the dual-phase lag equation, *International Journal of Heat and Mass Transfer* (2017) **108**:1–10.
12. S. Singh, K. Sushil: Numerical study on triple layer skin tissue freezing using dual phase, *International Journal of Thermal Sciences* (2014) **86**: 12–20.
13. Y. Zhang: Generalized dual-phase lag bioheat equations based on nonequilibrium heat, *International Journal of Heat and Mass Transfer* (2009) **52(21–22)**: 4829–4834.
14. E. Majchrzak, B. Mochnacki: *Metody numeryczne: Podstawy teoretyczne, aspekty praktyczne i algorytmy*, Wydawnictwo Politechniki Śląskiej, Gliwice 1998.
15. E. Majchrzak, B. Mochnacki: Implicit scheme of the finite difference method for 1D dual-phase lag equation, *Journal of Applied Mathematics and Computational Mechanics* (2017) **16(3)**:37–46.
16. E. Majchrzak, B. Mochnacki: Dual-phase lag equation. Stability conditions of a numerical algorithm based on the explicit scheme of the finite difference method, *Journal of Applied Mathematics and Computational Mechanics* (2016) **15(3)**:89–96.
17. A. Breslow: Thickness, cross-sectional areas and depth of invasion in the prognosis of cutaneous melanoma, *Annals of Surgery* (1970) **172(5)**:902–908.
18. H. Drepper, B. Bieß, B. Hofherr, M. Hundeiker, A. Lippold, F. Otto, G. Padberg, A. Peters, H. Wiebelt: The prognosis of patients with stage III melanoma prospective long-term study of 286 patients of the fachklinik hornheide, *Cancer* (1993) **71(4)**:1239–1246.
19. S. Tanaka: Cryosurgery for malignant melanoma. In *Basics of Cryosurgery* (Vienna, Austria, 2001), pp. 289–293.
20. R. Verma, S. Kumar: Computational study on constant and sinusoidal heating of skin tissue using radial basis functions, *Computers in Biology and Medicine* (2020) **121**:103808.

NUMERICAL ANALYSIS OF SKIN TUMOR FREEZING USING DUAL-PHASE LAG MODEL

Abstract

In the paper, the freezing of skin tumors (e.g. malignant melanoma) is considered. The tissue damage mechanism and the freeze-thaw cycle are presented. The thermal interaction between the cryoprobe and tissue is described using a dual-phase lag model (DPLM) in which two phase lags are defined: the relaxation time (associated with the heat flux) and the thermalization time (the temperature gradient lag). Numerical calculations were carried out with the finite difference method (FDM). Implicit and explicit schemes are derived and compared. Finally, numerical modeling of an exemplary cryosurgical procedure is described and analyzed.

Keywords: bioheat transfer, dual phase lag equation, freezing of biological tissue, finite difference method

DOI:10.34918/85115

Michał BERNYŚ[1], Bartłomiej MELKA[1,*], Maria GRACKA[1], Ziemowit OSTROWSKI[1], Marek ROJCZYK[1], Krzysztof PSIUK-MAKSYMOWICZ[2], Damian BORYS[2], Jarosław WASILEWSKI[3], Jan GŁOWACKI[3], Ryszard BIAŁECKI[1]

Chapter 15. NUMERICAL MODELING OF BLOOD FLOW IN BIFURCATION AND TRIFURCATION GEOMETRIES USE OF DIFFERENT BOUNDARY CONDITIONS

15.1. Introduction

The study on 50 million people showed that the world's leading cause of death is ischemic heart disease, which is responsible for 11.3% of deaths, followed by a stroke with an 11% mortality rate [1]. Over the past 19 years, the World Health Organization and CDC reported a growing trend in the increasing number of deaths caused by heart diseases [2, 3].

New methods have been introduced in medicine to a patient-specific. The physicians in fighting the increasing trend of heart disease incidence to protect human life. One of the techniques introduced is patient-specific numerical modelling. Computational Fluid Dynamics (CFD) is the most frequently used simulation method used in the field of blood flow simulations. Nowadays, such methods are used routinely in a variety of applications ranging from modelling vessel prosthetics to the operation of artificial heart valves [4]. CFD is also frequently used to model heart malefactions. Such an approach was used in the case of chronic stenosis of the blood artery [5]. The research objective was to study the ill effects on blood flow. To increase the accuracy of the results, the 3D CT scan was made to reproduce the actual vessel geometry used in the simulation [6, 7].

---

[1] Department of Thermal Technology, Faculty of Environmental and Energy Engineering, Silesian University of Technology, Gliwice, Poland.
* Corresponding author: bartlomiej.melka@polsl.pl.
[2] Department of Systems Biology and Engineering, Faculty of Automatic Control, Electronics and Computer Science, Silesian University of Technology, Gliwice, Poland.
[3] Department of Diagnostic Imaging, Silesian Center for Heart Diseases, Zabrze, Poland.

The building up of plaque deposition may be one of many reasons for coronary artery disease (CAD), which causes abnormalities in blood flow resulting in a reduced supply of oxygen and other vital nutrients to the heart. Studies show that the left coronary artery (LCA) is more prone to plaque formation compared to the right coronary artery (RCA) [8]. It is caused mainly by uniform flow in RCA compared to LCA. The local wall shear stress in the LCA is low and intensively oscillating, especially in areas of bifurcation [9].

The presented study aims to perform a blood flow through two LCA geometries and analyze the differences in the flow behaviour for a typical LCA case with bifurcation and a rarer case where trifurcation is present. As a result, it will be possible to identify locations prone to plaque deposition, the determination of which may be necessary for proper patient diagnosis and, in the future, for an effective treatment or prevention.

15.2. Mathematical model

Every fluid moving within the model boundaries is subjected to general mass and momentum conservation equations. Such an approach is used to ensure the simulation's proper course.

The mass conservation equation:

$$\frac{\partial \rho}{\partial t} + \nabla(p\vec{v}) = S_M \tag{1}$$

The momentum conservation equations:

$$\frac{\partial}{\partial t}(\rho\vec{v}) + \nabla(\rho\vec{v}) = -\nabla \mathrm{p} + \nabla\bar{\tau} + \rho\vec{g} + \vec{F} \tag{2}$$

where: ρ is a fluid density, t is time, $\vec{v}$ represents velocity vector field, $\bar{\tau}$ is stress tensor, p is static pressure, $\rho\vec{g}$ *and* $\vec{F}$ are the gravitational body force and external body forces

When performing simulation in ANSYS Fluent package, choosing the most suitable model is essential to obtain the intended effect and reflect the flow behaviour occurring in blood vessels. To carry out the simulation, it is necessary to correctly identify the principles of flow behavior in biofluids. The viscosity value depends on the share rate and shear stress. In this study, the Carreau model has been selected [12].

The Carreau model's viscosity and parameters η, k, and n are fluid dependents. k is the time constant; n is the power-law exponent (described in non-Newton power-law), and η_0 and η_∞ are the upper and lower viscosity limits, corresponding to low and high shear stresses, respectively

The Carreau viscosity (η_c) can be calculated as:

$$\eta_c = \eta_\infty + (\eta_0 - \eta_\infty)[1 + k^2\gamma^2]^{\frac{n-1}{2}}$$

where: (3)

$$\gamma = \sqrt{\frac{1}{2}\Pi}$$

The two coefficients, k, and n are empirically determined [15] and Π denotes the second invariant strain tensor. The values of the coefficients are taken as: η_0 = 0.056 Pa s, η∞ = 0.00345 Pa s, k = 3.313 s and n = 0.3568

The OSI measures the oscillation of shear stresses at the arterial wall and is computed from Eq. (4).

$$OSI = \frac{1}{2}\left[1 - \frac{\left|\int_0^T \tau_w dt\right|}{\int_0^T |\tau_w| dt}\right] \tag{4}$$

where T is the cardiac period and τ_w is the stress at the artery wall. Wall Shear Stress (WSS) represents the mean shear stresses at the artery wall obtained from Eq. (4).

Wall shear stress was calculated on the base of Eq. (5).

$$\text{WSS} = \frac{1}{T}\left[1 - \int_0^T |\tau_w| dt\right] \tag{5}$$

These indicators are essential since they are believed to give an idea of locations that tend to develop atheroma plaque deposition and, therefore, stenosis growth [17, 18].

The model should accurately perform both near-wall calculations and core fluid motion in the presented research. The shear wall stress (WSS) has a vital role in the flow of blood morphotic elements. Therefore it is unacceptable to omit WSS [10] in the computed simulation so that the standard $k-\varepsilon$ model could not be used, due to imperfect near-wall computations. To meet the need for both requirements, the model of $k{-}\omega$ was proposed however in the current reseach the laminar flow was assumed. The estimated Re number during the cycle did not exeed 3000.

A time-varying velocity profile was used to represent the actual pulsatile nature of blood flow [11]. This implementation provides more detailed results that could reflect the actual heart behaviour. Figure 1 represents the used velocity temporal profile, which was determined by dividing data published in [12] into five parts. The equations of curves describing changes in velocity were adjusted using the sixth-order polynomial. Moreover, the blood flow rate delivered to LCA was estimated as a 2% value of the patient's stroke volume. The user-defined function (UDF) implemented the velocity profile as the inlet boundary condition. UDF will set the proper polynomial for the given time step in running the simulation. Figure 2 shows that the velocity profile has been adjusted so that one cycle lasts about 0.83 s, and it was consistent with a patient pulse of 72 bpm.

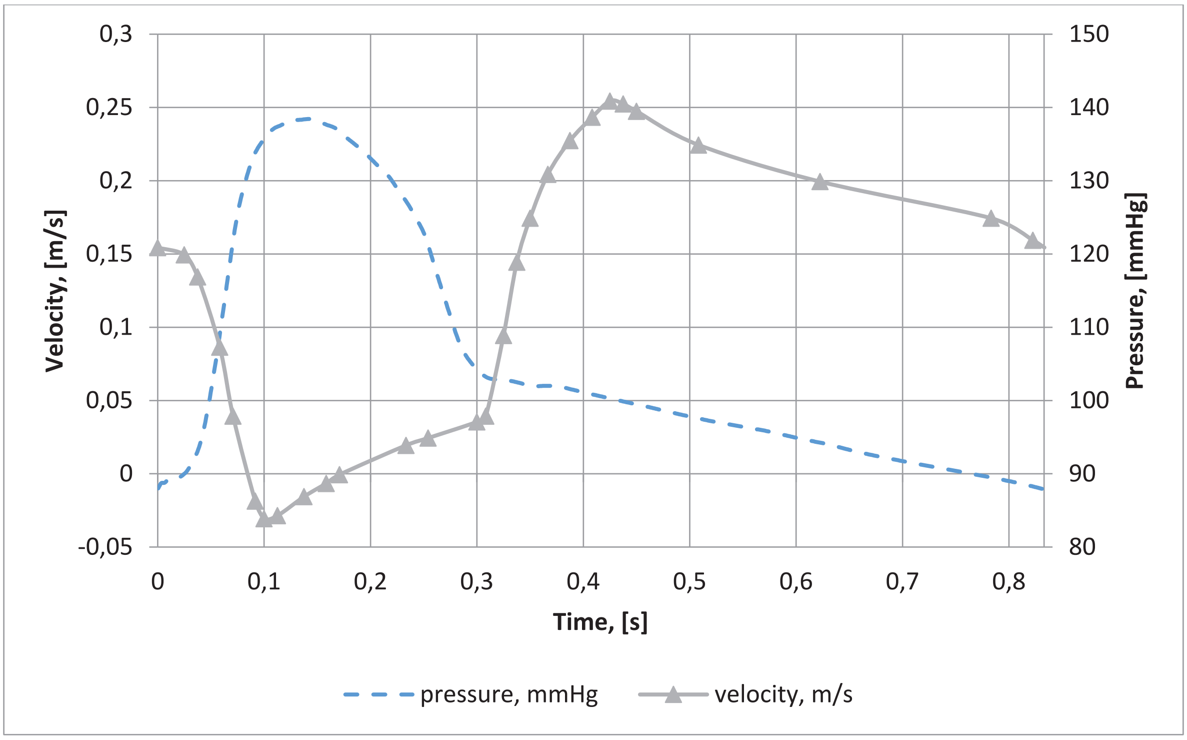


Fig. 1. Inlet velocity and outlet pressures change in a blood vessel
Rys. 1. Zmienny profil prędkości na wlocie oraz zmienny profil ciśnienia na wylocie

A variable velocity profile shown in Fig. 1 was used for the first set of calculations while maintaining constant outlet pressures. The same velocity variable profile at the inlet was used for the second set of calculations. The various pressure was used for the outlets, which was presented in Fig. 2 using a blue dotted curve. The results obtained from such a procedure allow an estimate of the influence of the pulsatile character of the pressures in the domain on OSI and WSS indicators.

15.2.1. Geometry and discretization procedure

In the simulation, two geometries were examined. The first is a bifurcation geometry shown in Fig. 2 with blunt green. While the second geometry includes trifurcation, and in Fig. 2, it is a geometry with blunt and vibrant green colour. The CT images have been segmented [16]. Produce STL files were later used to define the fluid domain.

Properties. Mesh was created in ANSYS Mesher, using tetrahedral cells. The discretization and the comparison of the parameters of mesh quality for both geometries are shown in Table 1.

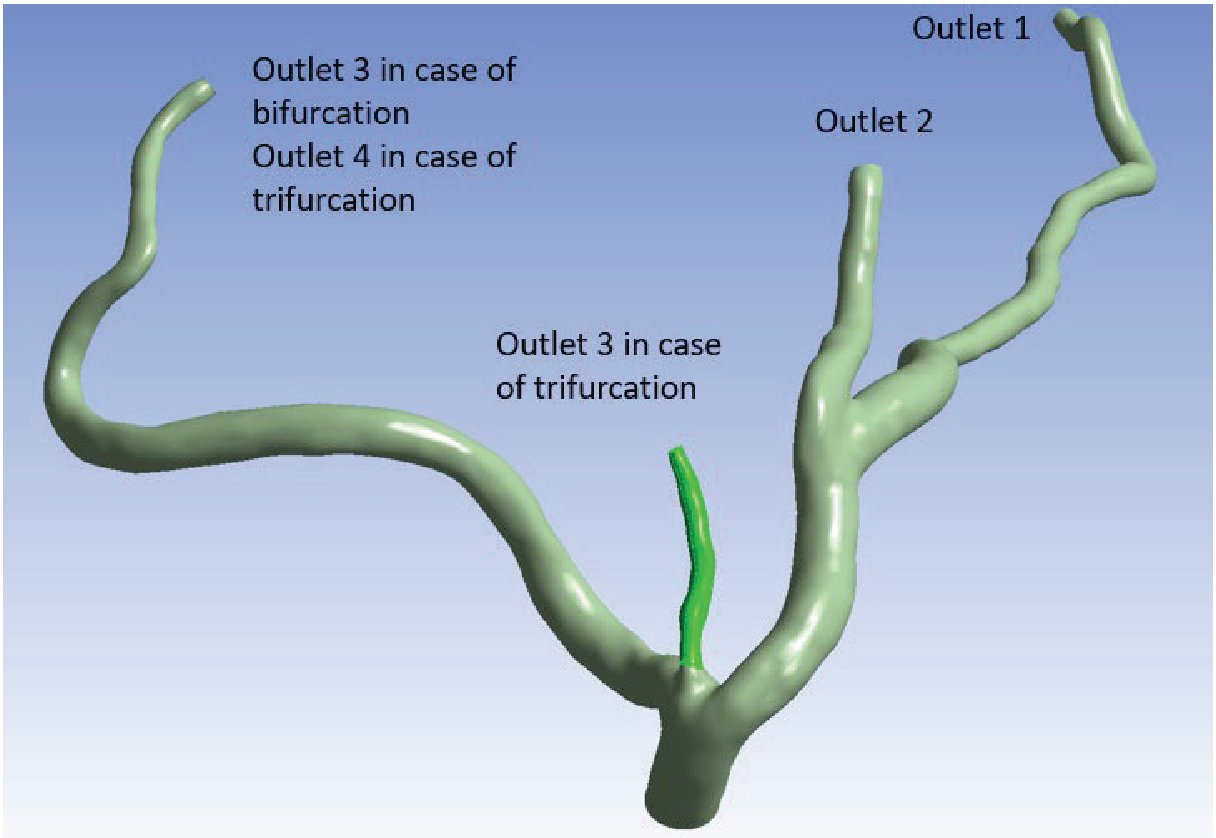


Fig. 2. Geometry comparison in between bifurcation and trifurcation
Rys. 2. Porównanie geometri pomiędzy bifurkacją a trifurkacją

Table 1

Mesh metrics comparison

*	Bifurcation mesh	Trifurcation mesh
Average Element Size	0.25 mm	0.25 mm
Number of Elements	1319788	1338590
Number of Nodes	244986	248878
Average Orthogonal Quality	0.78489	0.78439
Minimal Orthogonal Quality	0.14816	0.31457
Maximal Orthogonal Quality	0.99487	0.99517

15.3. Results and discussion

For both geometries with bifurcation and trifurcation, calculations were performed for two scenarios. The first scenario assumed a constant outlet pressure of 0 Pa, while the second scenario remaining boundary conditions were listed in Section 1.2. As a result of the performed calculations, contours corresponding to WSS and OSI indicators were created.

The resulting OSI and WSS contours show a high probability of atherosclerotic plaque deposition. In the case of OSI, special attention should be paid to locations with a high value of this factor. On the contrary, in the case of WSS, attention should be paid to low values, as the high values will suggest sweeping off the plaque [17, 18].

In Fig. 3 a), a black circle can indicate the site of the plaque diagnosed at the stage of performing geometry scans. This means that the performed calculations indicated the probability of new plaque deposition in the area where calcification was diagnosed

During the calculations, no significant differences were found in the cases for variable and constant outlet pressure. The only difference was the case of bifurcation. Slightly increased OSI values can be observed at 0.32 compared to 0.28 in trifurcation geometries. A similar characteristic can be seen in the WSS values. Lower values at the level of 2.21 Pa occur in the domain with bifurcation and in the domain with trifurcation of values at 2.30 Pa. The flowing blood has more drainage possibilities than trifurcation, so it does not accumulate in the branch with the largest cross-section area.

Fig. 4 and Fig. 5 show the magnitudes of the mass flow rates in case of changing velocity profiles. The flows are distributed accordingly to cross-section areas of a different branch. For example, for the outlet with the largest cross-section area, the blood flow is 7.25 ml/min compared to 1.57 ml/min in the smallest cross-section area. The mass balance remained at the level of 0.01 ml/min, which proves that the blood mass is conserved

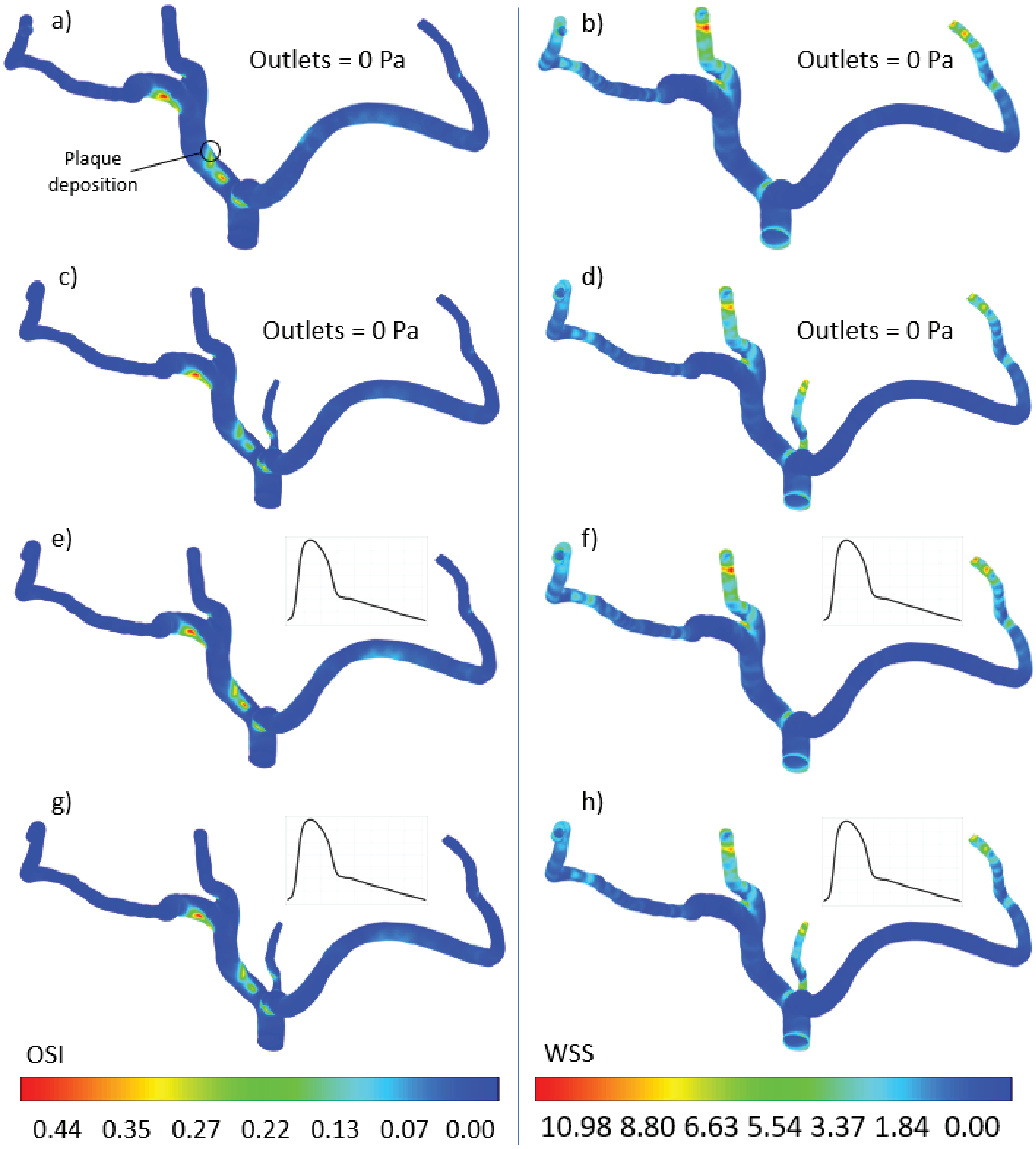


Fig. 3. Wall contours of OSI (left) and WSS (right): a) and b) bifurcation geometry with constant pressure outlets, c) and d) trifurcation geometry with constant pressure outlets, e) and f) bifurcation geometry with pulsatile pressure outlets, g) and h) trifurcation geometry with pulsatile pressure outlets

Rys. 3. Kontury OSI (po lewej) i WSS (po prawej): a) i b) geometria bifurkacji ze stałym profilem ciśnienia wylotowego, c) i d) geometria trifurkacji ze stałym profilem ciśnienia wylotowego, e) i f) geometria rozgałęzienia ze pulsacyjnym profilem ciśnienia na wylocie, g) i h) geometria trifurkacji z pulsacyjnym profilem ciśnienia na wylocie

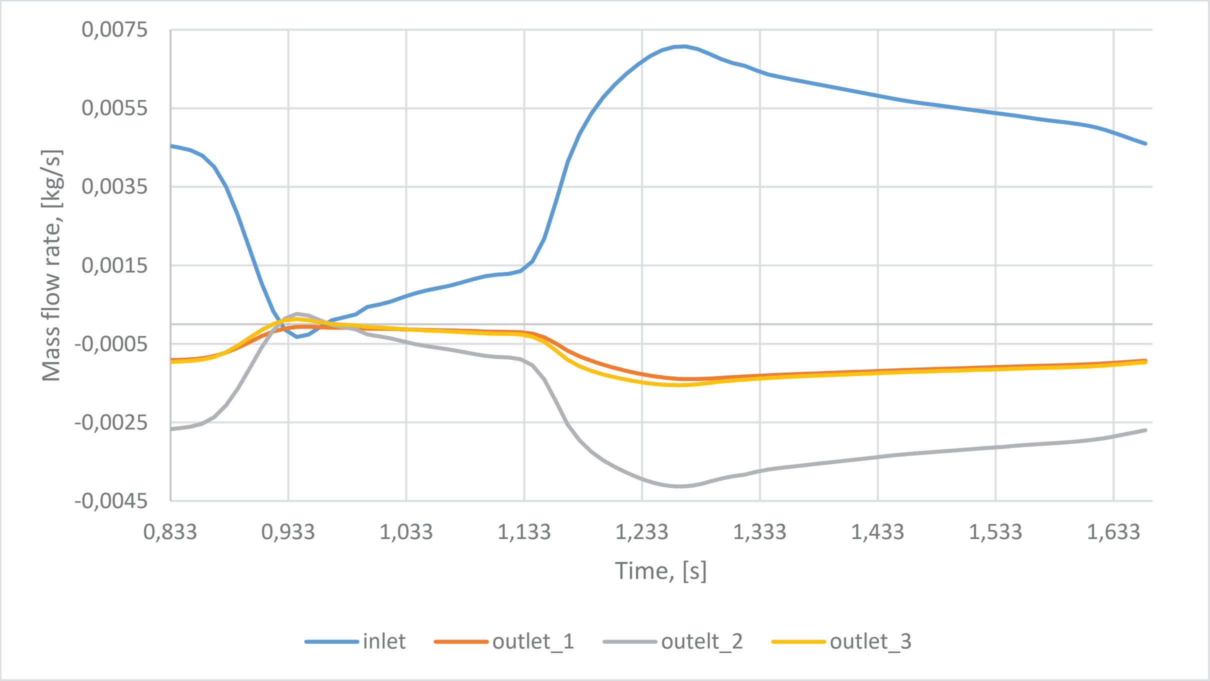


Fig. 4. Mass flow rates in bifurcation with changing velocity profile
Rys. 4. Przepływ masowy przez geometrię bifurkacji ze zmienną prędkością wlotową

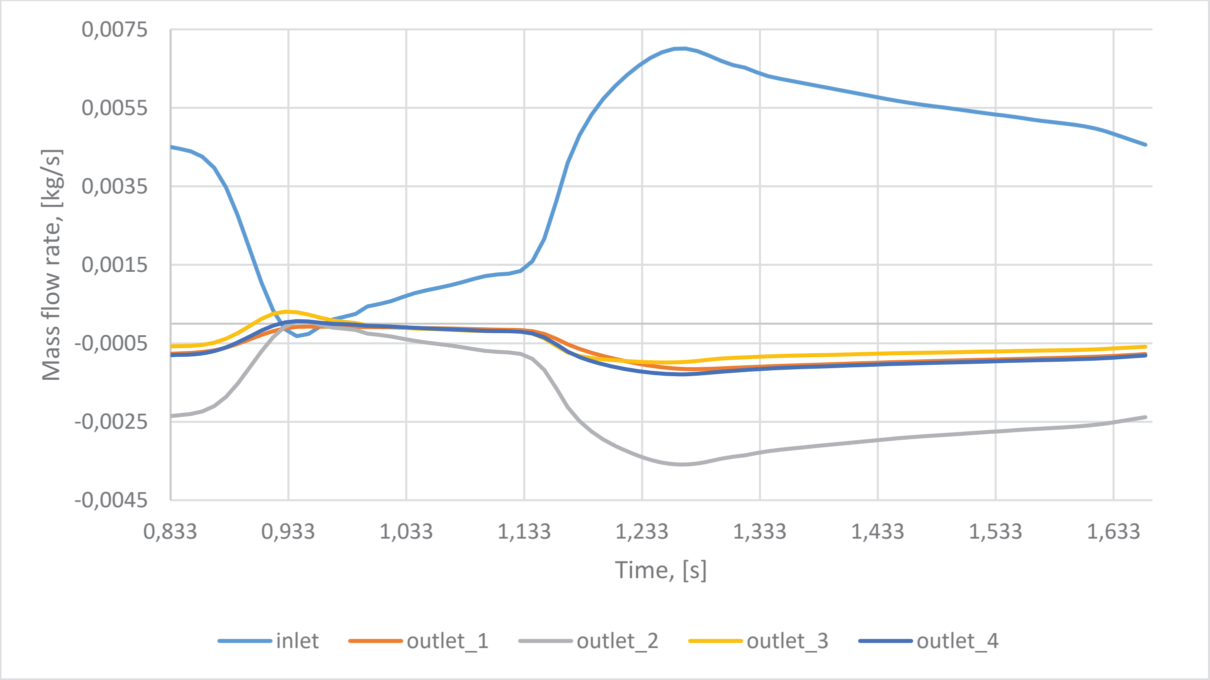


Fig. 5. Mass flow rates in trifurcation with changing velocity profile
Rys. 5. Przepływ masowy przez geometrię trifurkacji ze zmienną prędkością wlotową

15.4. Conclusions

The work was aimed at numerical simulations for two different geometries, which were then used to test the calculations on two different boundary conditions. One of the conditions was to set a variable velocity profile with the simultaneous setting of constant outlet pressures. The second case was the setting of variable inlet and outlet pressures.

The obtained OSI and WSS contours identify areas with a high potential for plaque deposition. The indexes, as mentioned above, indicated, among other things, where the atherosclerotic plaque was deposited in the patient's coronary arteries. The tool could potentially be used to predict the sites of plaque deposition

The obtained results are the basis for the master thesis. It is worth mentioning that the results need to be refined. Particular emphasis should be placed on the quality of the mesh in future calculations and the improvement of the UDF to better reflect the conditions prevailing in the heart.

Acknowledgements

This research is supported by National Science Centre (Poland) project No. 2017/27/B/ST8/01046 and project No. 2019/34/H/ST8/00624. This help is gratefully acknowledged.

Bibliography

1. C.J. McAloon et al.: The changing face of cardiovascular disease 2000–2012: An analysis of the world health organization global health estimates data, *Int. J. Cardiol.*, vol. 224, 2016, pp. 256–264, doi: 10.1016/j.ijcard.2016.09.026.
2. World Health Organization (WHO): The top ten leading causes of death by broad income group, *Fact sheet*, vol. 8, no. 3, 2009, pp. 1–5.
3. “CDC National Health Report Highlights CS251163,” 2012, [Online]. Available: https://www.cdc.gov/healthreport/publications/compendium.pdf.
4. H.A. Simon, L. Ge, F. Sotiropoulos, A.P. Yoganathan: Simulation of the three-dimensional hinge flow fields of a bileaflet mechanical heart valve under aortic conditions, *Ann. Biomed. Eng.*, vol. 38, no. 3, 2010, pp. 841–853, doi: 10.1007/s10439-009-9857-0.

5. T. Kuribara et al.: Preoperatively estimated graft flow rate contributes to the improvement of hemodynamics in revascularization for Moyamoya disease, *J. Stroke Cerebrovasc. Dis.*, vol. 30, no. 1, 2021, p. 105450, doi: 10.1016/j.jstrokecerebrovasdis.2020.105450.
6. W. Wu et al.: 3D reconstruction of coronary artery bifurcations from coronary angiography and optical coherence tomography: feasibility, validation, and reproducibility, *Sci. Rep.*, vol. 10, no. 1, 2020, pp. 1–11, doi: 10.1038/s41598-020-74264-w.
7. D. Lopes, H. Puga, J. Teixeira, R. Lima: Blood flow simulations in patient-specific geometries of the carotid artery: A systematic review, *J. Biomech.*, vol. 111, 2020, p. 110019, doi: 10.1016/j.jbiomech.2020.110019.
8. B.T. Chan, E. Lim, K.H. Chee, N.A. Abu Osman: Review on CFD simulation in heart with dilated cardiomyopathy and myocardial infarction, *Comput. Biol. Med.*, vol. 43, no. 4, 2013, pp. 377–385, doi: 10.1016/j.compbiomed.2013.01.013.
9. J.V. Soulis, T.M. Farmakis, G.D. Giannoglou, G.E. Louridas: Wall shear stress in normal left coronary artery tree, *J. Biomech.*, vol. 39, no. 4, 2006, pp. 742–749, doi: 10.1016/j.jbiomech.2004.12.026.
10. B. Su, J.M. Zhang, H. Zou, D. Ghista, T.T. Le, C. Chin: Generating wall shear stress for coronary artery in real-time using neural networks: Feasibility and initial results based on idealized models, *Comput. Biol. Med.*, vol. 126, no. September, 2020, p. 104038, doi: 10.1016/j.compbiomed.2020.104038.
11. M. Lodi Rizzini et al.: Does the inflow velocity profile influence physiologically relevant flow patterns in computational hemodynamic models of left anterior descending coronary artery?, *Med. Eng. Phys.*, vol. 82, pp. 58–69, 2020, doi: 10.1016/j.medengphy.2020.07.001.
12. M. Abbasian, M. Shams, Z. Valizadeh, A. Moshfegh, A. Javadzadegan, S. Cheng, Effects of different non-Newtonian models on unsteady blood flow hemodynamics in patient-specific arterial models with in-vivo validation, *Comput. Methods Programs Biomed.*, vol. 186, 2020, p. 105185, doi: 10.1016/j.cmpb.2019.105185.
13. J. Wasilewski, K. Mirota, S. Peryt-Stawiarska, A. Nowakowski, L. Poloński, M. Zembala: Wprowadzenie do numerycznej mechaniki płynu na podstawie komputerowej symulacji przepływu w lewej tętnicy wieńcowej, *Kardiochirurgia i Torakochirurgia Pol.*, vol. 9, no. 3, 2012, pp. 366–374, doi: 10.5114/kitp.2012.30851.

14. M. Abbasian, M. Shams, Z. Valizadeh, A. Moshfegh, A. Javadzadegan, S. Cheng: Effects of different non-Newtonian models on unsteady blood flow hemodynamics in patient-specific arterial models with in-vivo validation, *Comput. Methods Programs Biomed.*, vol. 186, 2020, p. 105185, doi: 10.1016/j.cmpb.2019.105185.
15. S. Tabakova, E. Nikolova, S. Radev: Carreau model for oscillatory blood flow in a tube, *AIP Conf. Proc.*, vol. 1629, no. February, 2014, pp. 336–343, doi: 10.1063/1.4902290.
16. B. Berthier, R. Bouzerar, C. Legallais: Blood flow patterns in an anatomically realistic coronary vessel: Influence of three different reconstruction methods, *J. Biomech.*, vol. 35, no. 10, 2002, pp. 1347–1356, doi: 10.1016/S0021-9290(02)00179-3.
17. A.S. Anayiotos, S.A. Jones, D.P. Giddens, S. Glagov, C.K. Zarins: Shear stress at a compliant model of the human carotid bifurcation, *J. Biomech. Eng.*, vol. 116, no. 1, pp. 98–106, 1994, doi: 10.1115/1.2895710.
18. D.N. Ku, D.P. Giddens, C.K. Zarins, S. Glagov: Pulsatile flow and atherosclerosis in the human carotid bifurcation. Positive correlation between plaque location and low and oscillating shear stress, *Arteriosclerosis*, vol. 5, no. 3, 1985, pp. 293–302, doi: 10.1161/01.atv.5.3.293.

NUMERICAL MODELING OF BLOOD FLOW IN BIFURCATION AND TRIFURCATION GEOMETRIES W USE OF DIFFERENT BOUNDARY CONDITIONS

Abstract

The presented work compares the flow through the coronary vessels under different boundary conditions. The geometries were developed as a result of patient scans. It is a modern and non-invasive diagnostic method to achieve highly accurate results. The simulations' results will allow us to identify places susceptible to plaque deposition, which causes collapses or death.

The obtained results for the cases with constant and variable pressure turned out to be similar. The only difference was the case of bifurcation. Slightly increased OSI values result from a lower number of outflows from the domain. The flowing blood has more drainage possibilities than trifurcation, so it does not accumulate in the ebb with the largest cross-section area. Other than the conclusion, there were no significant discrepancies indicative of inaccuracies in any of the methods.

For future simulations, the generated mesh should be refined, and the model's sensitivity should be assessed. The next step is to improve the UDF to give the most accurate results compared to those measured.

Keywords: bifurcation, trifurcation, blood, coronary vessel, CFD

DOI:10.34918/85116

Wesley Chebii RUTTO[1] , Benard KIPSANG[2,*]

Chapter 16. INVESTIGATION OF ADIPOSE MECHANICAL PROPERTIES UNDER COMPRESSION LOADING FOR IMPROVING HUMAN BODY ARMOR MODELING

To improve blast absorption characteristics of personal protection equipment (PPEs), understanding the effects of blasts on the human body itself is a critical step toward improving protective device design. Adipose tissue mechanical properties were investigated under quasi-static and dynamic compression loading at varying strain rates on specimens made from porcine subcutaneous adipose tissue. The specimens were subjected to a uniaxial compression test on an MTS universal testing machine to obtain stress versus strain response data. The tests were carried out at three strain rate regimes: 0.05%/s, 0.5%/s, and 5%/s. For modeling the passive behavior of subcutaneous adipose tissue, three-term QLV and one-term Ogden models were chosen. Python codes were used to numerically simulate experimental data. The hyperelastic material parameters obtained were as follows: $\mu = 16.4$, $\alpha = 8.41$, $C_1 = -136.980$, $C_2 = -225.60$, and $C_3 = 3636.85$.

To simulate adipose tissue, a finite element model was created. The average elastic and shear modulus for strain rates ranging from 0.05 to 5% were determined to be: Initial $E \approx 12.71 \pm 10.93$ kPa, final $E \approx 1300 \pm 66.60$ kPa. Initial modulus of rigidity $G \approx 4.23 \pm 2.30$ kPa, final $G \approx 433.30 \pm 22.20$ kPa. For the Ogden and quasilinear viscoelastic models, the correlation coefficients were $R^2 = 0.9809$ and $R^2 = 0.9986$, respectively. When comparing the model to the experimental data at a 0.5% strain rate, these were the coefficients of determination. A comparison of experimental and simulation results revealed that the QLV model could reproduce the stress-strain curve with acceptable accuracy.

---

[1] College of Mechanical and Vehicle Engineering, Hunan University, Changsha, 410082, China.

[2] Department of Advanced Materials and Technologies, Faculty of Material Science, Politechnika Slaska, 40-019 Katowice Poland.

[*] Corresponding author benard.kipsang@polsl.pl.

Under steady loading, it was discovered that subcutaneous adipose tissue is strain rate dependent and exhibits a non-linear stress-strain response. Because it is soft and compliant under quasi-static loading but stiff and resilient under dynamic loading, it exhibits both viscous and elastic behavior.

16.1. Introduction

Technology advancement has not only helped to improve counterterrorism measures, but it has also enabled insurgent bomb makers to develop new, more sophisticated explosives used on battlefields to attack. Blast injuries are more common than not encountered on battlefields. In such a blast situation, severe injuries are frequently accompanied by traumatic effects throughout the body. Post-traumatic effects, such as ear injuries, lung injuries, abdominal injuries, brain injuries, and soft tissue damage, occur more frequently than death [1].

A blast wave is generated from an explosion by the sudden release of a large amount of energy in a very small volume. A blast wave is commonly comprised of a shock wave and blast wind [2]. When a bomb explodes, the area around the explosion becomes over-pressurized, resulting in highly compressed air particles that travel faster than the speed of sound. This wave will dissipate over time and distance and will exist only for a matter of milliseconds. This initial blast wave inflicts the most damage. When this blast wave reaches a human body, the body will feel the force of the blast, which is the primary and initial impact of the shockwave [3]. This damages the body on impact. The force with which a blast occurs is called its load. Blast loading can inflict enough stress on the body to cause it to be flung onto a structure. After a blast wave strikes the body, high-velocity shockwaves, or stress waves, will continue to pass through it. Then travel through the organs and tissues. Shockwaves carry energy through the medium they pass through; they're supersonic and transport more energy than sound waves. At the explosion site, a vacuum is created by the rapid outward movement of the blast. This vacuum will almost immediately refill itself with the surrounding atmosphere. This creates a very strong pull on any nearby person or structural surface after the initial push effect of the blast has been delivered. As this void is refilled, it creates a high-intensity wind that causes fragmented objects, glass and debris to be drawn back in toward the source of the explosion [3]. Therefore, in a blast scenario, two parts namely positive and negative sections are created. Blast wave leads to injury through spalling, implosion, and inertia [4].

Blast injuries have been categorized into four major types: primary, secondary, tertiary and mutilating [2, 4, and 5]. Primary blast injury is caused by a blast wave striking the body and transmitting energy directly into the body. Currently, there are no effective ways to prevent shockwaves from passing through protective clothing, and in some cases protective measures may even amplify the destructive effects [6]. Secondary blast injury occurs by debris propelled onto or into the body by the blast or its blast wave. The most common cause of death in a blast event is secondary blast injuries. These injuries are caused by flying debris generated by the explosion. Terrorists often add screws, nails, and other sharp objects to bombs to increase injuries. Soft tissue injuries are one of the most common types of secondary blast injuries. Thus, it is of substantial interest in this research.

Following blast dynamics, Personal protection equipment (PPEs) have continually been designed. However, to date, military-grade armor materials designed to mitigate ballistic and shrapnel attacks are less effective in resisting blast impacts. In order to improve blast absorption characteristics of armors, the first key step is thoroughly understanding the effects of blasts on the human body itself. In literature, an extensive study on the effects of blast on some parts of a human body has been conducted using experimental and computational modeling techniques. Abundant computational models to study blast effects on the human head were developed by Ganpule et al. [7, 8]. Arnab Chanda et al. modeled skin and bone sections from six locations of the body i.e. the elbow, figure, wrist, cheek bone, forehead and skin basing on the knowledge of human anatomy and eye estimation of three subjects [1]. Subcutaneous adipose tissue, like other soft tissues in human body, is vulnerable to blast loading and shrapnel penetrations among other impacts. Furthermore, studies have shown that subcutaneous adipose tissue plays an important role as a mechanical load absorbing and distributing member that absorbs shock and protects against local stresses [9, 10]. However, there is a paucity of studies characterizing biomechanical response of adipose to loading conditions. This provides adequate grounds to study the mechanics of adipose tissue loading, which will be a vital step towards improving the design of protective devices.

Most experiments have been conducted on surrogates, animal models, cadavers and few on live human subjects [10, 11]. Adipose tissue is commonly characterized using human heel pad tissue, human breast and porcine subcutaneous [9–11] whereby its biomechanics have been studied experimentally and numerically using finite element methods. In this case, a pig's subcutaneous adipose tissue specimen was used as a model for studying human subcutaneous tissue because comparatively, pigs have genetical, physiological and anatomical traits similar to humans, which make them one of the most useful and versatile animal models. Owing to these similarities, data generated from porcine models are more likely to lead to viable human treatments than those from murine work. In addition, the similarity in size and physiology to humans allows pigs to be used for many experimental approaches not feasible in mice. Research areas that employ pigs range from neonatal development to translational models for cancer therapy [30]. Specimen(s) were subjected to varying strain rates [10–12] to demonstrate different loading conditions. Adipose, as with most biological tissues, exhibits heterogeneous, rate-dependent, viscoelastic behavior and experiences large non-linear deformations [15]. It is suggested that adipose is approximately isotropic in structure and due to the large lipid content, is almost incompressible [11]. Notwithstanding, a recent study by [14] observed some anisotropy in the specimen.

This work aims to future look into adipose tissue mechanics by investigating its mechanical properties under quasi-static and dynamic compression loading at varying strain rates so that important factors that affect its physical properties are understood. Thus, give an insight into predicting the behavior based on the components it is made of.

16.2. Materials and methods

MTS machine was used to perform the uniaxial compression hence obtaining the stress versus strain response data of porcine fat tissue from the material testing mechanical laboratory.

16.2.1. Specimen preparation and Experimentation

A pig's subcutaneous adipose tissue was obtained from the nearby market, fresh and succulent with the skin on. The skin was dissected off the adipose tissue using a scalpel blade so that preloading conditions are not altered much by the cutting force. More so, correct measurements and the right shape are maintained on the surfaces of the specimen. This was to ensure the consistency of all the samples. The resulted block of adipose was cut accurately into cubes of 10mm in dimensions. The measurement of dimensions in the specimens' unloaded state was done using a conventional ruler. To enable the generation of data consistent with the specified specimen shape, the dimensions of every specimen were confirmed prior to loading.

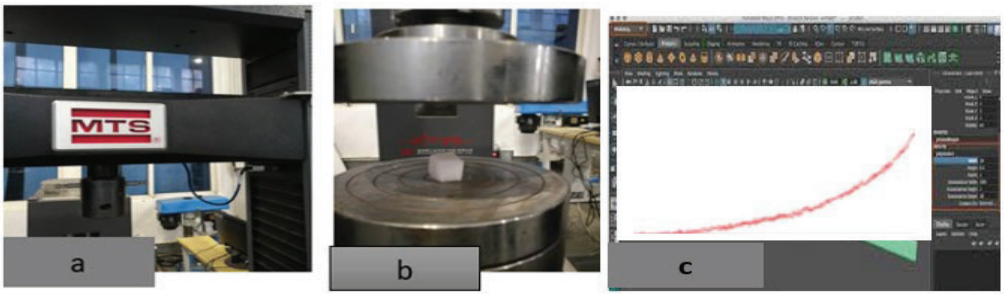


Fig. 1. (a) MTS universal testing machine (b) Specimen placement (c) Computer data generation and plotting
Rys. 1. (a) Uniwersalna maszyna wytrzymałościowa MTS (b) umieszczenie próbki (c) generowanie i komputerowe kreślenie danych

The specimen was placed on the center of the lower platen and oriented so its remaining four surfaces were perpendicular to both the platens. This allows the compression to be uniaxial without limiting shear loading because the test was unconfined. The upper platen was lowered down to partially touch the top surface of the specimen. This point was assumed to be the datum. The two platens were made sure that they were frictionless to avoid lateral stresses. A compressive loading regime was programmed into the software i.e. zero load, zero position and find contact. The test was executed using the software program where the sample was loaded under compression. These tests were performed over 3 strain rate regimes which is 0.05%, 0.5% and 5%. The same procedure was followed for all the specimens. The ambient temperature was 10°C.

It was assumed that the experiment occurred under symmetric conditions, with a homogeneous, isotropic and incompressible material, and that imposed deformations were small compares to the original size of the cubic specimen.

16.2.2. Experimental data analysis

The experimental engineering stress was found by dividing the reaction force at every sample point by the un-deformed contact area. The engineering strain corresponded to the displacement of the platen divided by the un-deformed (original) height of the specimen. Nominal stress and strain were also calculated.

Unlike traditional solid materials that are typically described with a linear elastic or hyperelastic behavior, most soft tissues are inhomogeneous, anisotropic, and frequently have integral nonlinear viscoelastic mechanical behavior because they consist of significant amounts of interstitial fluids. Therefore, this mechanical behavior involves a response that changes with time as a product of tissue relaxation. All the above-mentioned characteristics together with a nonlinear stress-strain relationship make the characterization of biological tissues complicated [23]. Three--term QLV and one-tern Ogden models were selected for modeling the passive behavior of subcutaneous adipose tissue. The equation below expresses the Ogden hyperelastic material law.

$$W = \sum_{i=1}^{n} \frac{\mu_i}{\alpha_i} \left(\lambda_1^{\alpha_i} + \lambda_2^{\alpha_i} + \lambda_3^{\alpha_i} - 3 \right) + K(J - 1 - InJ) \tag{1}$$

Where W is the strain energy of the model, **μ** is the initial shear modulus, **α** is the strain hardening exponent, **λ_i** is the stretch ratio in three directions: **x**, **y** and **z** and **J** represents the bulk modulus. Subcutaneous AT was considered incompressible, in this way, **J** equals to 1. Consequently, one-order Ogden law simplifies a can be expressed as follows,

$$W = \frac{2\mu}{\alpha^2} (\lambda_1^{\alpha} + \lambda_2^{\alpha} + \lambda_3^{\alpha} - 3) \tag{2}$$

Quasilinear viscoelastic (QLV) constitutive model is perhaps the most common viscoelastic constitutive model used to characterize biological soft tissues. It has the capability of modeling materials with time-dependent viscoelastic behavior that undergo large deformation. The constitutive equation, which is defined as the convolution integral of the time-dependent reduced relaxation function with the time derivative of the elastic response function, can be used to fit the experimental data and to estimate the QLV parameters [25, 26].

QLV theory models the viscoelastic response of a material based on a stress relaxation function and the instantaneous stress resulting from a ramp strain as:

$$\sigma(t) = G(t) * \sigma^{e}(\varepsilon) \tag{3}$$

Where **σ (t)** is the stress at any time **t**, **$\sigma^{e}(\varepsilon)$** is the stress corresponding to an instantaneous strain, **G(t)** is the reduced relaxation function representing the stress of the material divided by the stress after the initial ramp strain noted that * represents the convulsion of **G** and **$\sigma^e(\varepsilon)$,** which is **G(t)** multiplied by **$\sigma^e(\varepsilon)$** and then integrated over time. **G(t)** is defined as:

$$\mu\alpha G(t) = \frac{\sigma(t)}{\sigma(0^+)} \; and \; G(0^+) = 1 \tag{4}$$

The complete stress history at any time is then the convolution integral:

$$\sigma(t) = \int_{-\infty}^{t} G(t-\tau)\,\frac{\delta\sigma^e(t)}{\delta\varepsilon}\frac{\delta\varepsilon}{\delta\tau} \tag{5}$$

Where G is the reduced relaxation function, $\frac{\delta\boldsymbol{\sigma}^e(\mathbf{t})}{\delta\varepsilon}$ represents the instantaneous elastic response, and $\frac{\delta\varepsilon}{\delta\boldsymbol{\tau}}$ is the strain history. It can be assumed that t starts at zero instead of negative infinity for the exponential situation. The reduced relaxation function is:

$$G(t) = ae^{-bt} + ce^{-dt} + ge^{-ht}$$

Or

$$G(t) = \sum_{i=1}^{n} G_i e^{\beta_i t} \tag{6}$$

Where **a**, **b**, **c**, **d**, **g**, **h** are all constants to be determined experimentally. **β_i** is the decay constant.

The instantaneous stress response is assumed to be represented through the nonlinear elastic relationship:

$$\sigma^e(\varepsilon) = A(e^{B\varepsilon} - 1) \tag{7}$$

In the nonlinear elastic function, **A** and **B** are constants that must be determined by experiment. They are also constants to be fit with experimental data. Three terms of coefficients are adopted in this study.

16.2.3. Analytical solution

To determine the strain rate dependent material properties, an inverse FE algorithm was used. Selected models (Ogden and QLV) were simulated numerically in python using the available package weave namely: matplotlib, numpy and scipy. Parameter optimization was carried out using a pycham optimization package. Thus, material parameters were calibrated by minimizing the square of the absolute error between the stress-strain data obtained numerically with the experimental measurements. This produced optimal material parameters. The use of absolute errors instead of relative errors is justified as it provides a better fit for large deformations [27]. Since the main output of numerical and experimental tests was the compressive force measured over time on the adipose tissue specimen by the load cell, the objective function was formed to calculate and minimize the difference between the experimental and numerical force measurements.

$$f(x) = \sqrt{\frac{\sum_{i=1}^{n}(y_{iExp} - y_{iNum})^2}{n}} \tag{8}$$

Where **f(x)** is the objective function, $\mathbf{y_{iExp}}$ is the value of experimental force, $\mathbf{y_{iNum}}$ is the responsive numerical force and **n** is the number of data points in experimental tests.

Hyperelastic material parameters; **μ**, **α**, **C1**, **C2,** and **C3** were obtained from the numerical inverse process.

16.2.4. Modelling and Simulation

Commercial solver LS-DYNA_SMP_R11.0.0×64 was used in conjunction with Ls-prepost(R) V4.5.24 for modeling and simulations of the adipose tissue. Ls-prepost can be used to model simple components that do not have intricacies in its parts. This version of the software have Implicit and Explicit finite element codes.

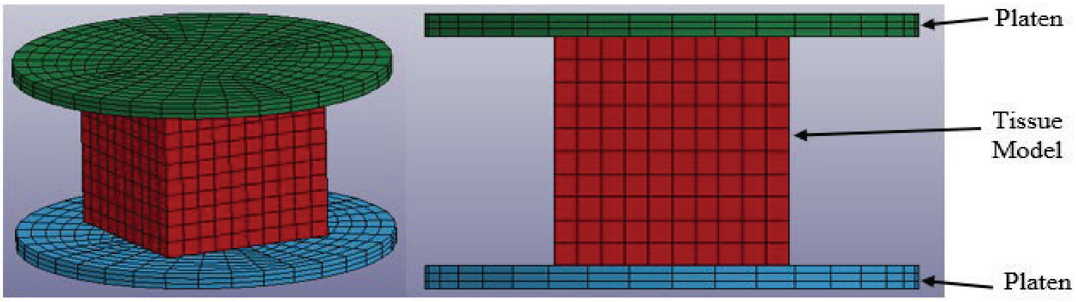


Fig. 2. Adipose tissue model between two steel platens
Rys. 2. Model tkanki tłuszczowej pomiędzy dwiema stalowymi płytami

The model was meshed to obtain accurate results. Automatic surface to surface contact type was chosen with segment sets between the slave and/or master and the specimen. Static coefficient of friction (FS) was set to 0.01 for both cases.

16.2.5. Boundary conditions

The model setup is that all translational and rotational degrees of freedom (DOF) about the bottom platen are constrained while on the top platen, the translational DOF in compression direction was unconstrained. This allows the top platen to move downwards during compression. The specimen model remains unconfined in the middle of the top and bottom platens thus, undergoing uniaxial compression. To execute the compression work, top platen (slave) was given a prescribed motion of 0.5%/s and the load curve scale factor set to negative 1.

16.2.6. Hyperelastic Material models

Ls-dyna present material models that can be used to analyze biological materials. However, few can reasonably be used to analyze the constitutive properties of adipose tissue. The models that can be used for this particular one are: MAT_077-O OGDEN_RUBBER and MAT_176 QUASILINEAR_VISCOELASTIC.

The former is a hyperelastic nearly incompressible model in principal directions and the latter is a quasi-linear, isotropic, viscoelastic material which represents biological soft tissues, such as the brain. Therefore suitable for simulating adipose tissue because it undergoes finite deformation when stressed.

16.2.7. Inverse finite element method/analysis

To implement finite element analysis in ls-dyna, material parameters found experimentally were input into the respective material models. Ogden material model (Fig. 4) required the input of parameters: density of material, Poisson's ratio, mu (μ), alpha (α) and optional relaxation shear modulus (G) and relaxation constant(s) (β). QLV material model (Fig. 5) required: density of material, Poisson's ratio, bulk modulus, relaxation shear modulus (G) and relaxation constant(s) (β) and elastic constants C_1, C_2 & C_3. The experimentally acquired parameters can be substituted with load curve ID in the material model keyword cards.

Table 1

Parameters input to material model cards

Density (ρ)	**920 kg/cm^3**
Poisson's ratio (υ)	0.499983
Bulk modulus (k)	5e5 Pa
Relaxation shear modulus (G)	$G_1 = 3000$, $G_2 = 2770$
Relaxation constants (β)	$\beta_1 = 310$, $\beta_2 = 100$
Ogden parameters	$\mu = 16.40$, $\alpha = 8.410$
QLV parameters	$C_1 = -136.98$, $C_2 = -225.60$, $C_3 = 3636.85$

The parameter values of ρ, υ, G_1 and β_1 used above were obtained from [46] simulation of the calibration experiment with low strain rate (0.2%/s).

During the simulation, the specimen was compressed at a strain rate of 0.5%/s until 50% of its initial length as demonstrated in Fig. 3 below. Ogden model simulation was carried by invoking implicit solver while QLV simulation used explicit solver because it cannot be simulated in implicit solver. Both simulations converged successfully.

Stress-strain data from the elastic response of the material during simulation was obtained by plotting designated node(s) and/or element(s) stress-strain responses that were enabled by invoking in DATABASE_HISTORY_NODE/SOLID_ID. The data was saved to excel files for later analysis.

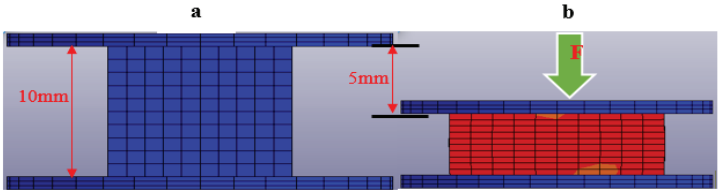


Fig. 3. (a) Tissue sample before compression (b) After compression
Rys. 3. (a) Próbka tkanki przed kompresją (b) po kompresji

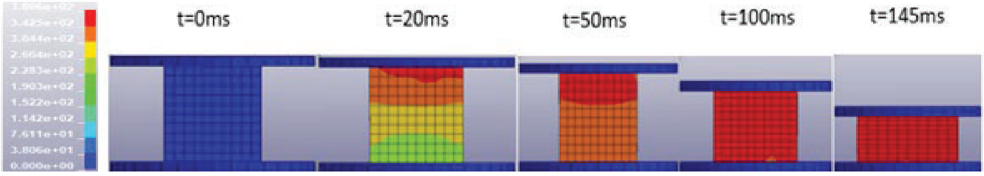


Fig. 4. Demonstrates simulation using Ogden material model
Rys. 4. Demonstracja symulacji z wykorzystaniem modelu Ogdena

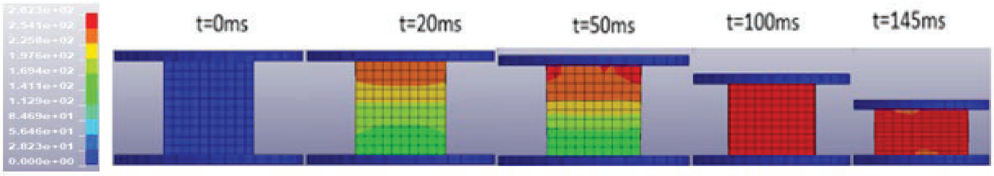


Fig. 5. Demonstrates simulation using quasilinear viscoelastic (QLV) material model
Rys. 5. Demonstracja symulacji z wykorzystaniem kwaziliniowego modelu lepkosprężystego

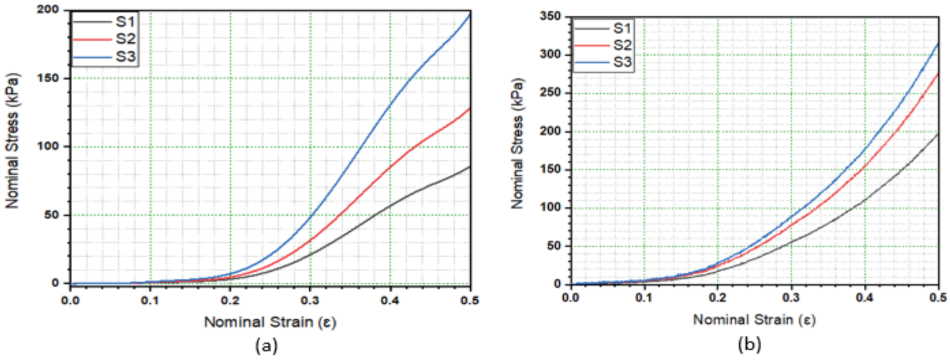
From the above demonstrations, it’s observable that the material models respond to the applied load differently. This is depicted by change of stress colorations and shape with respect to time. Ogden model gives almost immediate response to the load compared with QLV model. On examining further, the specimen during the compression test, lateral elongation was observed proving the presents of Poisson’s ratio in the material. Poisson’s ratio contributes to the form of stress-strain diagram. The initial and final cross-sectional area of the specimen in compression changed. In this way, nominal and true strain diagrams cannot be identical. Until 50% compression, the volume of the material was conserved.

The experimental curve at 0.5%/s was chosen for comparison with the developed finite element model. The reason for choosing only one curve is that there is not always a continuous increasing stiffness from lower to higher strain rates within the levels, lower, intermediate and high of strain rate, hence one curve is sufficient. More curves would not increase the accuracy [27].

16.3. Results

16.3.1. Experimental results

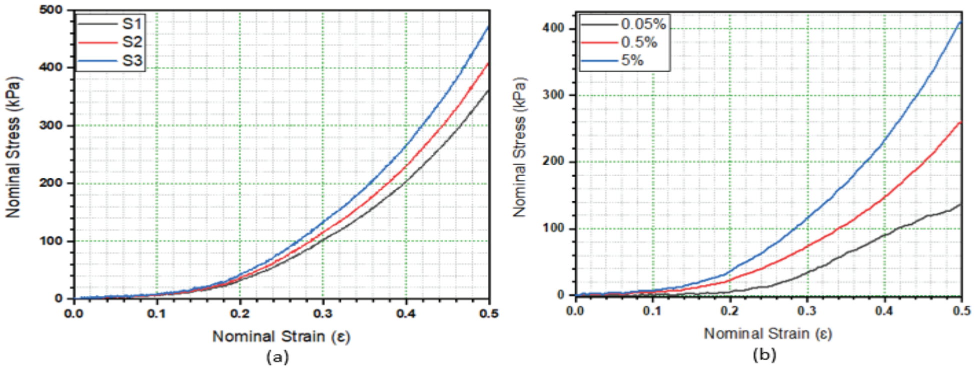
Experimental data obtained from compression of adipose tissue (see Table 1 above) in the mechanical laboratory were represented graphically using OriginPro 2019b. It is a graphing and data analysis software. Curves of the three samples are plotted for each strain rate regimes i.e. 0.05/s, 0.5/s and 5/s as shown in Fig. 6 (a), 6 (b) and 7 (a) below.



(a) (b)

Fig. 6. (a) Three specimen samples at 0.05% strain rate. (b) Three specimen samples at 0.5% strain rate

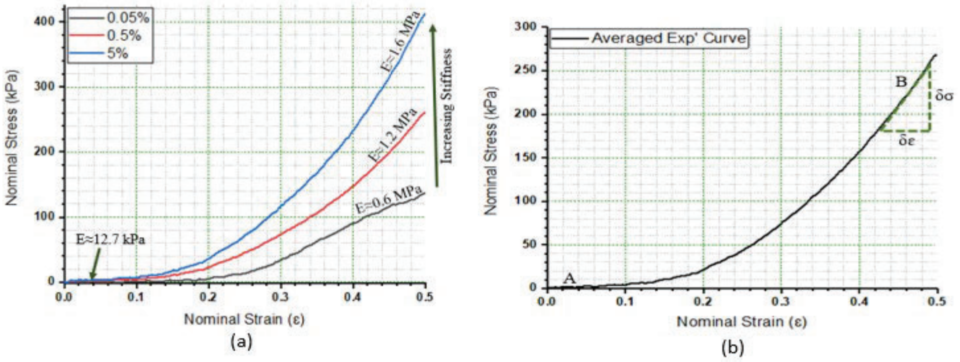
Rys. 6. (a) Trzy próbki przy odksztalceniu 0.05%. (b) Trzy próbki przy odkształceniu 0.5%



(a) (b)

Fig. 7. (a) Three specimen samples at 5% strain rate. (b) Average of 0.05%, 0.5% and 5% strain rates from figure 6 (a), 6 (b) and 7 (a) respectively

Rys. 7. (a) Trzy próbki przy odkształceniu 5%. (b) Średnie dla odkształceń 0.05%, 0.5%, 5% z rysunków 6 (a), 6 (b) i 7 (a)



(a) (b)

Fig. 8. (a) Increasing stiffness as strain rate increases. (b) Average from figure 3.3

Rys. 8. (a) Wzrastająca sztywność przy wzrastającym odkształceniu. (b) Średnia z rysunku 3.3

The values of elastic modulus and modulus of rigidity shown in the table below was obtained from calculating the slope of the curve in Fig. 8 (b) above.

Table 2

Average approximate elastic and shear modulus with standard deviations for strain rates between 0.05% and 5%

	Initial (A) (kPa)	**STD (kPa)**	**Final (B) (kPa)**	**STD (kPa)**
Elastic modulus (E)	12.71	10.93	1300	66.60
Shear modulus (G)	4.23	2.30	433.30	22.20

The graph describes the stress-strain relationship of the adipose tissue and more so clearly depicts the strain rate dependence of the tissue. The elastic modulus is within the range reported in [9, 11, 20, 21] and the shear modulus calculated from the curve slope is in agreement with the values reported in [11].

16.3.2. Analytical results

Using python algorithm parameters i.e. $\boldsymbol{\mu}$, $\boldsymbol{\alpha}$ and $\mathbf{C_1}$, $\mathbf{C_2}$, & $\mathbf{C_3}$ from Ogden law and Quasilinear-Viscoelastic law respectively, are obtained as tabulated in Table 3 and 4 below.

Table 3

Ogden parameters for the listed strain rates

Strain Rates	μ	α	R^2	RMSE
0.05%	7.820	8.763	0.9973	2.090
0.5%	16.062	8.330	0.9870	8.804
5%	25.380	8.331	0.990	13.91

Table 4

QLV parameters for the listed strain rates

Strain Rates	C_1	C_2	C_3	R^2	RMSE
0.05%	-124.36	840.40	-15.280	0.990	4.251
0.5%	-32	593.20	1041.14	0.9993	2.032
5%	-51.11	937.26	1645.00	0.9993	3.210

Table 5

Averaged Ogden parameters

μ	α	R^2	RMSE
16.40	8.410	0.9809	11.195

Table 6

Averaged QLV parameters

C_1(-)	C_2(-)	C_3	R^2	RMSE
136.980	225.60	3636.85	0.9986	3.90

These averaged parameters are used in the respective material models for simulations.

16.3.3. Inverse FEM results

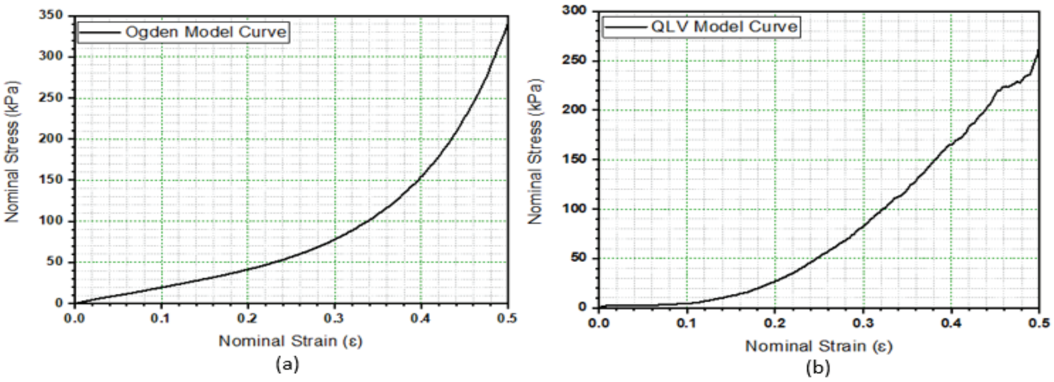


Fig. 9. (a) Stress-Strain response of Ogden material model. (b) Stress-Strain response of quasilinear viscoelastic material model

Rys. 9. (a) Krzywa naprężenie-odkształcenie dla modelu Ogdena. (b) Krzywa naprężenie-odkształcenie dla kwaziliniowego modelu lepkosprężystego

Ogden model depicts some sensitivity to applied load even from the beginning and increasingly non-linear from 0 to 0.5 Strain. According to ls-dyna material selector, Ogden material model is in the family of rubber (RB). This can explain the low viscoelasticity and stress sensitivity seen at initial stages of stress application as shown on Fig. 9 (a) above. Apart from the graphical representation, its stress-strain responses can be observed in Fig. 4.

QLV finite model curve shows pliability of the specimen during compression as seen in the graph above, from 0 to 0.1 strain. QLV material model is in the family of biological materials (BIO). Stress response of this model can also be described by Fig. 5.

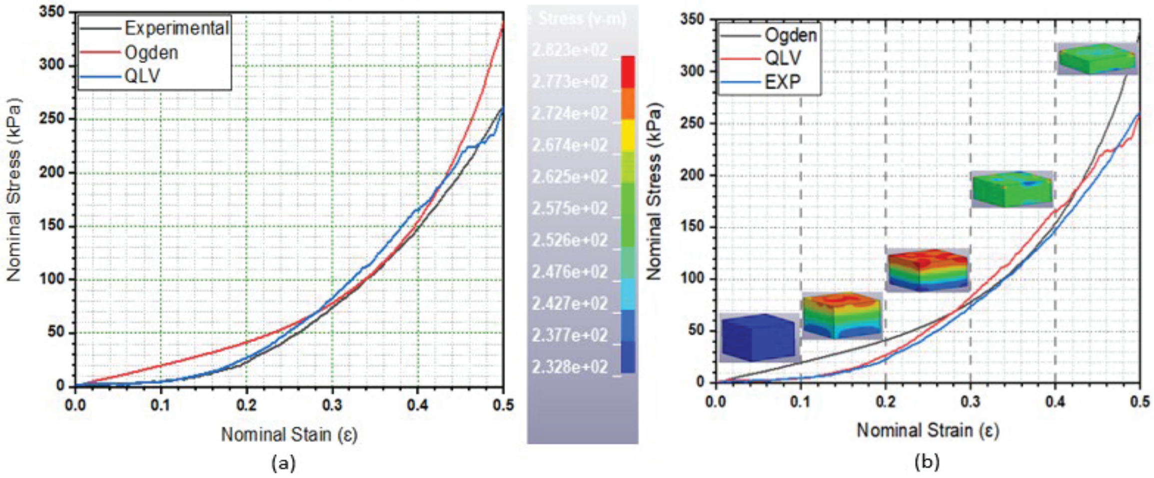


Fig. 10. (a) Experimental and finite element models data comparison. (b) Models deformation and curves under compression loading

Rys. 10. (a) Porównanie danych eksperymentalnych I modelu elementow skończonych. (b) Modele deformacji I krzywych pod obciążeniem ściskającym

A clear understanding of the relationship among model compression, load-deformation curves and effective stresses can be obtained from the figure below.

16.4. Discussion

Uniaxial compression test was conducted in the laboratory to obtain the stress-strain data when the specimen was subjected to three different loading rates as mentioned earlier on in the previous chapter(s). The compiled data were represented graphically to bring into perspective the stress-strain response when subcutaneous adipose is loaded at varying rates. Three specimen samples were tested at every strain rate. That is: 0.05%, 0.5% and 5%. At 0.05% strain rate, it is observed that the curves are less inclining and the stress for deforming the specimen to 50% of its initial shape is also low as shown by Fig. 6 (a). Curves at 0.5% strain rate are more inclined and its deformation stress is higher than the one for 0.05%. The slope of curves at 5%/s increased further than for the first two strain rates. The three samples in each strain rate are averaged into one as illustrated in Fig. 7 (b). These clearly show the stress-strain response with changing loading rate. The elastic modulus and stress strain curve were found to be dependent on the strain rate. As the strain rate increases, the elastic modulus increases and the stress strain curve becomes stiffer.

The general behavior of the curve shows that from 0 to 0.1 strain, the material deforms with low sensitivity to the applied load - it remains viscous. From 0.1 to 0.2, sensitivity to loading rises. From 0.2 to 0.3, a sharp stiffness is observed and between 0.3 and 0.5, the material's stiffness increases significantly. This behavior gives a non-linear stress strain curve. The viscous behavior of subcutaneous adipose can be attributed to the 3D foam cell structure of interlobular septa and large amount of interstitial fluids that fill the spaces between cells. Whereas the elasticity is attributed to the existence of adipose ECM in varying degrees of stiffness and elasticity which is a property primarily dependent on collagen and elastin concentration(s). The combination of these properties yields a viscoelastic material that is manifested in adipose tissues.

Regression analysis was conducted on stress-strain data from experiment to obtain the constitutive hyperelastic parameters of the tissue. Ogden and QLV models were employed to characterize the hyperelastic properties and are presented in Table 5 and 6 respectively. Strain hardening exponents remained relatively same for all the strain rates. The value of shear response (μ) see table 3.2, depicts that the material's deformation behavior under steady loading is strain rate dependent because it increases with strain rate. The negative values of μ is attributed to artifacts of the fitting. It is magnified when some of the powers in the Ogden model are negative. In such case, a small but positive shear modulus can artificially appear as negative. This could be experienced more often when the material undergoes finite deformation. In this study, the values of μ were taken as positive. QLV elastic constants C_1, C_2 and C_3 were used as they are in material model keyword cards.

Stress-strain data from finite element model analysis were plotted. Ogden model and QLV model showed different responses during compression test, see Fig. 9 (a) and 9 (b) The major difference is that curve from Ogden data shows instant sensitivity to loading whereas QLV curve shows low sensitivity to loading initially as observed on experimental curves.

To validate the results of the experiment, the stress strain curve at a strain rate of 0.5%/s was compared to the stress strain curves from finite element simulation. As observed in Fig. 10 (a), there is a pronounced difference between Ogden curve and experimental curve but slight difference between QLV and experimental curve. The correlation coefficients for Ogden and quasilinear viscoelastic models were $R^2 = 0.9809$ and $R^2 = 0.9986$ respectively.

A Comparison between the experimental and simulation results shows the capability of QLV model to reproduce the stress strain curve with acceptable accuracy. Therefore, the suggested material parameters are capable of fairly predicting the load and deformation of subcutaneous fat tissue.

16.4.1. Conclusions and recommendation

Subcutaneous adipose tissue is strain rate dependent and shows a non-linear stress-strain response under steady loading. It manifests both viscous and elastic behavior by the virtue that, it is soft and compliant under quasi-static loading but stiff and resilient under dynamic loading.

Quasilinear viscoelastic model gives a better prediction of the load-deformation response of subcutaneous fat tissue than Ogden. Model validation was performed employing the coefficients of determination in which the one for QLV nears 1 and its RMSE is small compared to Ogden.

A fresh sample of porcine can possibly give better results especially when all conditions are recorded prior to testing. Therefore, it is recommended that fresh samples be used in future to investigate the mechanical properties of subcutaneous adipose tissue so that comparative results can be obtained.

Credit author statement

Wesley C. Rutto: Conceptualization, Methodology, Investigation, Data collection, Data organization, Software, Visualization, and writing – Original draft preparation.

Benard Kipsang: Writing – Reviewing and Editing.

Conflict of interest

The author declares that the research was conducted in the absence of any relationships that can be interpreted as a potential conflict of interest.

Acknowledgements

We thank the Chinese government ministry of commerce (MOFCOM) for funding this research.

Bibliography

1. Effect of blasts on subject-specific computational models of skin and bone sections at various locations on the human body. Arnab Chanda 1, Rebecca Graeter 2, and Vinu Unnikrishnan 1, Published date 12 November 2015.
2. Ballistic helmets – Their design, materials, and performance against traumatic brain injury. S.G Kulkarni, X.-L. Gao, S.E. Horner, J.Q. Zheng, N.V. David (2013). US Army Research. 201.
3. TOM SCHEVE; How Bomb Blasts Cause Damage: howstuffworks 2008.
4. Zara R. Mathews, MD*, Alex Koyfman, MD†. The Journal of Emergency Medicine, Vol. 49, No. 4, pp. 573–587, 2015; BLAST INJURIES
5. Yan Zhao, Yuan-Guo Zhou*. The past and present of blast injury research in China; Available online 12 November 2015.
6. Christopher George Thom: Soft Materials under Air Blast Loading and Their Effect on Primary Blast Injury; Waterloo, Ontario, Canada, 2009.
7. Ganpule S.G., Chandra N., Salzar R., (2013): Mechanics Of Blast Loading On Post-Mortem Human Heads in The Study Of Traumatic Brain Injury (TBI) Using Experimental And Computational Approaches [PhD Dissertation]: University of Nebraska-Lincoln, 289.
8. Ganpule S., Chandra N., (2013): Mechanics of Interaction of Blast Waves on Surrogate Head: Effect of Head Orientation. ASME 2013 Summer Bioengineering Conference: American Society of Mechanical Engineers, V01BT55A028-029.
9. Alkhouli N., Mansfield J., Green E., Bell J., Knight B., Liversedge N., Tham J.C., Welbourn R., Shore A.C., Kos K., Winlove C.P., (2013): The mechanical properties of human adipose tissues and their relationships to the structure and composition of the extracellular matrix.
10. Thomas Payne, Sean Mitchell, Richard Bibb, Mark Waters; 30 September 2014; the evaluation of new multi-material human soft tissue simulants for sports impact surrogates.
11. Kerstyn Comley, Norman Fleck, (2012): The compressive response of porcine adipose tissue from low to high strain rate.

12. Grigoris Grigoriadis, Nicolas Newell, Diagarajen Carpanen, Alexandros Christou, Anthony M.J. Bull, Spyros D. Masouros; 8 September 2016; Material properties of the heel fat pad across strain rates.
13. Bruno Bordoni, DPT, DO, PhD,[1,2,3] Fabiola Marelli, DPT, DO,[2,3] Bruno Morabito, DPT, DO, PhD,[2,3] and Roberto Castagna, DPT, DO[2]; A New Concept of Biotensegrity Incorporating Liquid Tissues: Blood and Lymph
14. Sommer G., Eder M., Kovacs L., Pathak H., Bonitz L., Mueller C., Regitnig P., Holzapfel G.A., (2013): Multiaxial mechanical properties and constitutive modeling of human adipose tissue: a basis for preoperative simulations in plastic and reconstructive surgery. Acta Biomater. 9 (11), 9036–9048.
15. Sapozhnikov S., Ignatova A., (2013): Experimental and theoretical investigation of deformation and fracture of subcutaneous fat under compression. Mech. Compos. Mater., 48 (6), 649–654.
16. Peckham M.: Extracellular Matrix – 'Ground substance'. Faculty of Biological Sciences, University of Leeds.
17. Assist. Prof. Dr. Pınar Tulay, Ph.D. (Molecular biology of THE CELL). Extracellular Matrix (ECM) and Mechanisms of Cell Communication. Faculty of Medicine Near East University.
18. Mithieux S.M., Weiss A.S., (2005): Elastin. Advances in protein chemistry, 70: 437–61.
19. G.A. Holzapfel (Paper No. 7 November 2000). Biomechanics of Soft Tissue
20. Azar F.S., Metaxas D.N., Schnall M.D., (2002): Methods for modeling and predicting mechanical deformations of the breast under external perturbations. Med. Image Anal. 6 (1), 1–27.
21. Samani A., Zubovits J., Plewes D., (2007): Elastic moduli of normal and pathological human breast tissues: an inversion- technique-based investigation of 169 samples. Phys. Med. Biol., 52 (6), 1565–1576.
22. Deepak-Kumar, (2015): Mathematical Modelling of soft Biological Tissues. Department of Mechanical Engineering Indian Institute of Technology.
23. Carlos Bustamante-Orellana, Robinson Guachi, Lorena Guachi-Guachi, Simone Novelli, Francesca Campana, Fabiano Bini, Franco Marinozzi, (2019): Biomechanics of Soft Tissues: The Role of the Mathematical Model on Material Behavior.
24. Cees W.J. Oomens, Gerrit W.M. Peters, (2017): Skin Mechanics. Hyperelastic models – Biomechanics of Living Organs.

25. Christian Quaia, Howard S. Ying, Lance M. Optican, (2010): The Viscoelastic Properties of Passive Eye Muscle in Primates. III: Force Elicited by Natural Elongations.
26. Elizabeth Mesa-Múnera, Juan F. Ramírez-Salazar, Pierre Boulanger, John W. Branch, (2012): Inverse-FEM Characterization of a Brain Tissue Phantom to Simulate Compression and Indentation.
27. Kristofer Engelbrektsson, (2011): Evaluation of material models in LS-DYNA for impact simulation of white adipose tissue. Master's Thesis in Solid and Fluid Mechanics. Department of Applied Mechanics. Chalmers University of Technology SE-412 96, Goteborg Sweden.
28. Cato T. Laurencin, Yusuf Khan, (2013): Regenerative Engineering. Text book.
29. Mei Song, Yi Liu, Ling Hui, (2017): Preparation and characterization of acellular adipose tissue matrix using a combination of physical and chemical treatments. 138–146. https://doi.org/10.3892/mmr.2017.7857
30. Schook L.B., Collares T.V., Darfour-Oduro K.A., De A.K., Rund L.A., Schachtschneider K.M., Seixas F.K.: Unraveling the swine genome: implications for human health. Vol. 3:219–244 (Volume publication date February 2015).

INVESTIGATION OF ADIPOSE MECHANICAL PROPERTIES UNDER COMPRESSION LOADING FOR IMPROVING HUMAN BODY ARMOR MODELING

Abstract

Adipose tissue mechanical properties were investigated under quasi-static and dynamic compression loading at varying strain rates on specimens made from porcine subcutaneous adipose tissue. The specimens were subjected to a uniaxial compression test to obtain stress versus strain response data. The tests were carried out at three strain rates: 0.05, 0.5, and 5%/s. For modeling three-term, QLV and one-term Ogden models were chosen. Python was used to numerically simulate experimental data. The hyperelastic material parameters obtained were as follows: $\mu = 16.4$, $\alpha = 8.4$, $C_1 = -137$, $C_2 = -225.6$, and $C_3 = 3636.85$ To simulate adipose tissue, a finite element model was created. The average elastic and shear modulus for strain rates ranging from 0.05 to 5% were determined to be: For the Ogden and quasilinear viscoelastic models, the correlation coefficients were $R^2 = 0.9809$ and $R^2 = 0.9986$, respectively. When comparing the model to the experimental data at 0.5% strain rate, these were the coefficients of determination. A comparison of empirical and simulation results revealed that the QLV model was capable of reproducing the stress-strain curve with acceptable accuracy. Results showed that subcutaneous adipose tissue is strain rate dependent and exhibits a non-linear stress-strain response. It was found to be soft and compliant under quasi-static loading but stiff and resilient under dynamic loading, it exhibits both viscous and elastic behavior.

Keywords: subcutaneous adipose tissue, strain rates, finite deformation, viscoelastic, uniaxial compression, ls-dyna

DOI:10.34918/85117

Joanna BARTNICKA[1,*], Piotr BOJAR[2], Patrycja KABIESZ[1],
Agnieszka ZIĘTKIEWICZ[1], Katarzyna MLECZKO[1], Cristina HERRERA[3],
Constanza SAN MARTIN VALENZUELA[4], Harm PETERS[5]

Chapter 17. FUNCTIONAL EVALUATION IN PHYSIOTHERAPY EDUCATION. INTERNATIONAL RESEARCH WITH USE E-LEARNING PLATFORM

17.1. Introduction

One of the important health challenges of the 21st century is facing with development of civilization diseases and decreasing quality of life associated with it. In this regard, the gradual ageing of the population will be one of the crucial social change agents. According to Eurostat [1] the impact of demographic ageing within the European Union (EU) is likely to be of major significance in the coming decades. Additionally, low birth rates and higher life expectancy are changing the shape of the EU-28's age pyramid. It is estimated that the share of people aged 65 years or over will account for 28.7% of the EU-28's population by 2080 and people aged 80 years or above in the EU-28's population is projected to more than double between 2015 and 2080, from 5.3% to 12.3%.

While most people are aware that successive generations are living longer, less is known about the health of the EU's ageing population. In 2014, the number of healthy life years at birth was estimated at 61.4 years for men and 61.8 years for women in the EU-28; this represented approximately 79% and 74% of total life expectancy for men and women [2]. According to Eurostat [3] more than one-quarter of the EU-28 population which was self-reported regarding health experienced long-standing

---

[1] Silesian University of Technology, Gliwice, Poland.
* Corresponding author: Joannna.Bartnicka@polsl.pl.
[2] Cuiavian University in Włocławek, Poland.
[3] Instituto de Biomecánica, Spain.
[4] Universitat de València, Spain.
[5] Association of Medical Schools in Europe, German.

limitations due to health problems. Moreover, 26.9% of the population aged 16 and over reported some (moderate or severe) long-standing limitations in their usual activities due to health problems in 2013; 18.3% reported moderate long-standing limitations and 8.7% reported severe long-standing limitations. Taking into account severe long-standing limitations regarding the individual EU Members States, the highest shares of people reporting them were registered in Slovakia, Austria, Italy, Latvia, the United Kingdom, Germany and Greece, all 10% or higher. It can be concluded that the problem of heavy functional and activity limitations of (even young people) is present in countries regardless of level development.

Another research namely the European health interview survey (EHIS) conducted between 2006 and 2010 surveyed persons aged 15 and over and included questions about any physical or sensory limitations indicated that in all of the participating EU Member States, among people aged 65–74, around two thirds or more reported physical and sensory limitations, a share that peaked at close to 90% in Latvia and Romania. Moving to the oldest age group for which data are available, persons aged 85 and over, the share of people reporting physical and sensory limitations exceeded 95% in 11 of the 12 Member States for which data are available, the exception being Hungary (91%).

The presented data above and results of research on healthy life years are strictly linked to the quality of life, and represent those years that may be enjoyed by individuals free from the limitations of illness or disability. Chronic disease, frailty, mental disorders and physical disability tend to become more prevalent in older age, and may result in a lower quality of life for those who suffer from such conditions. The consequences are not only addressed for people suffered but also for healthcare system and economy (increased social expenditure related to population ageing, in the form of pensions, healthcare and institutional or private (health) care). In fact, the healthy life years monitor health as a productive or economic factor. Hence an increase in healthy life years is one of the main goals of EU health policy, given that this would not only improve the situation of individuals (as good health and long life are fundamental objectives of human activity) but would also lead to lower public healthcare expenditure and would likely increase the possibility that people continue to work later into life. If healthy life years increase more rapidly than life expectancy, then not only are people living longer, but they are also living a greater proportion of their lives free from health problems [2].

Taking into account this problematic background the important question is how to improve the extra years of life gained through increased longevity to spend them in good health and how to predict and prevent functional impairments of people.

The crucial areas, through which this improvement is feasible, are: learning new or supplemented competencies of physicians and creating new or supplemented healthcare procedures, the aim of which is to predict and prevent functional and activity limitations of people of different ages. Nowadays most medical students are trained to deal with structure and physiology and not function and disability. Actually, according to the research, it can be stated that function, impairment or disability are words in which many physicians have little interest or are ignored [4, 5].

Another problem of today's approach to diagnosing and treatment of physical impairments is focusing particularly on the musculoskeletal system (physical health) and lack of including in the evaluation model an integration approach taking into account also neurological, cognitive system (mental health) and social environment (social integration). According to research, such an approach is confirmed as equally important [6].

EU countries should be obligated to health promotion policy, however, it is needed the creation of new knowledge on how age, disability onset/duration, and type of limitation influence health and quality of life across adulthood. According to recent research this could identify when and for whom health promotion services are most critical in greater to optimize health and quality of life and reducig the risk of disability [7, 8].

As confirmed in research, the knowledge and technological innovation reached in the field of health sciences have grown considerably in recent years. Particularly, the scientific and technical advances achieved in the field of Functional Evaluation (FE) have become a revolution in the way practitioners identify, treat and assess many of the syndromes and pathologies with the highest impact on the quality of life of people. FE scope goes beyond the physical dimension, and it takes into account perceptions, needs and preferences. Particularly, FE collects the required technologies and methods to assess how people perform daily life activities (DLA), focusing on (1) the musculoskeletal system (physical health), (2) the neurological and cognitive system (mental health), (3) social environment (social integration).

It is believed that FE is an appropriate way to address current and future socio-sanitary challenges in Europe because FE supports the diagnosis. Sometimes, conventional diagnostic techniques do not allow clinicians to make objective decisions because diseases have an indeterminate origin or an organic lesion cannot be found to justify their occurrence. In other cases, results from FE methodologies complement conventional diagnostic techniques and help clinicians gain a broader and more realistic picture of patients' conditions. FE improves monitoring of patient development and treatment effectiveness. Because FE collects quantitative methods, by comparing their results, it is possible to quantify the development of patients or the impact that each type of treatment has on them. As a result, patients will receive better health care based on more accurate diagnoses and improved treatments. Healthcare assistance will be more focused on what patients are doing and how they are doing it, thus improving their quality of life. As FE enables more accurate diagnosis and helps improve treatments, national health systems will increase their efficiency and sustainability. FE supports the prediction of possible diseases and problems related to functional limitations and activity restrictions, making it possible to prevent them at an early stage of diagnosis.

Taking into account the proposition of including FE in the educational path, the original e-learning course was developed as a part of the international project "Development of innovative training solutions in the field of functional evaluation aimed at updating of the curricula of health sciences schools" and disseminate within students and lectures of health sciences schools as well as physiotherapists. Based on this, the paper aims to present research outcomes on the interest level of people who participated in the course, especially from point of view of the FE approach to diagnosis and treatment.

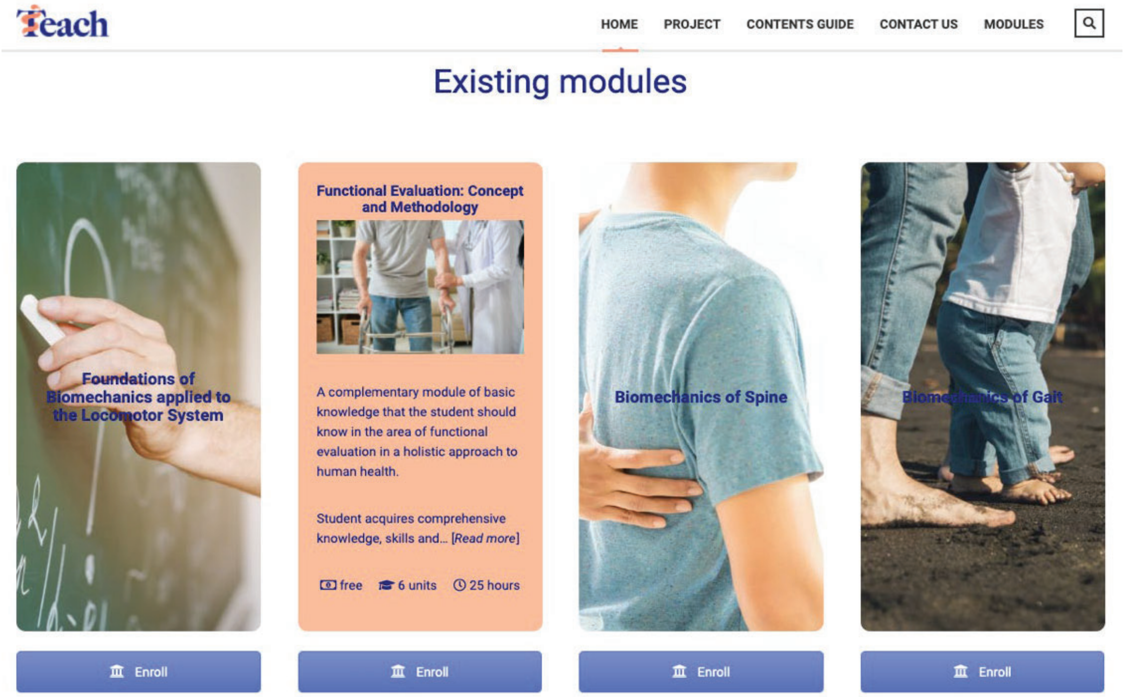
17.2. E-learning course evaluation methodology

The course was created taking into account the progressing phenomenon of population aging and the related problem of long-term functional limitations (physical, mental and social) affecting an increasing number of people, especially the elderly.

The main objective of the course was to familiarize participants with FE issues and, in particular, with elements of biomechanics that are one of the areas of interest in FE.

An online course integrated into the web has consisted of four basic modules divided into chapters: (i) Functional Evaluation: Concept and Methodology; (ii) Foundations of Biomechanics applied to the Locomotor System; (iii) Biomechanics of Spine; (iv) Biomechanics of Gait.

The view of the e-platform is shown in Fig. 1.



A complementary module of basic knowledge that the student should know in the area of functional evaluation in a holistic approach to human health.

Student acquires comprehensive knowledge, skills and... [*Read more*]

Fig. 1. View of the e-learning platform
Rys. 1. Widok platformy e-learnig
https://teach.ibv.org/

Every module includes from 4 to 6 topics with comprehensive and general information presenting the most important aspects of biomechanics, anthropometry and physiology. A module contains theoretical content with elements of practice, that formats are video and pdf as a script ready for download.

The expected study time for a module is 25 hours, which corresponds to 1 ECTS credit.

The course was dedicated to students, lecturers and physiotherapists. The course was very popular among students all over the world, even from India, Bolivia, Mexico and Georgia, however, the largest number of participants came from Europe. Classifying participants in terms of universities, the largest number were students of the University of Bydgoszcz – 535 participants, David Tvildiani Medical University – 36 participants, University of Valencia – 45 participants, Medical University in Lublin – 4 participants, Bydgoszcz University Enterprise of Knowledge – 2 participants. The geographical location map of participants and the percentage from universities are shown in Fig. 2 and Fig. 3 respectively.

Fig. 2. Territorial coverage of the TEACH course countries
Rys. 2. Zasięg terytorialny kursu TEACH w państwach Europy
Own elaboration

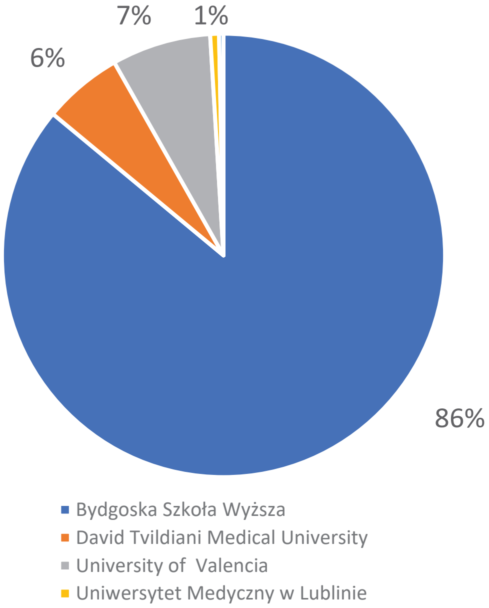
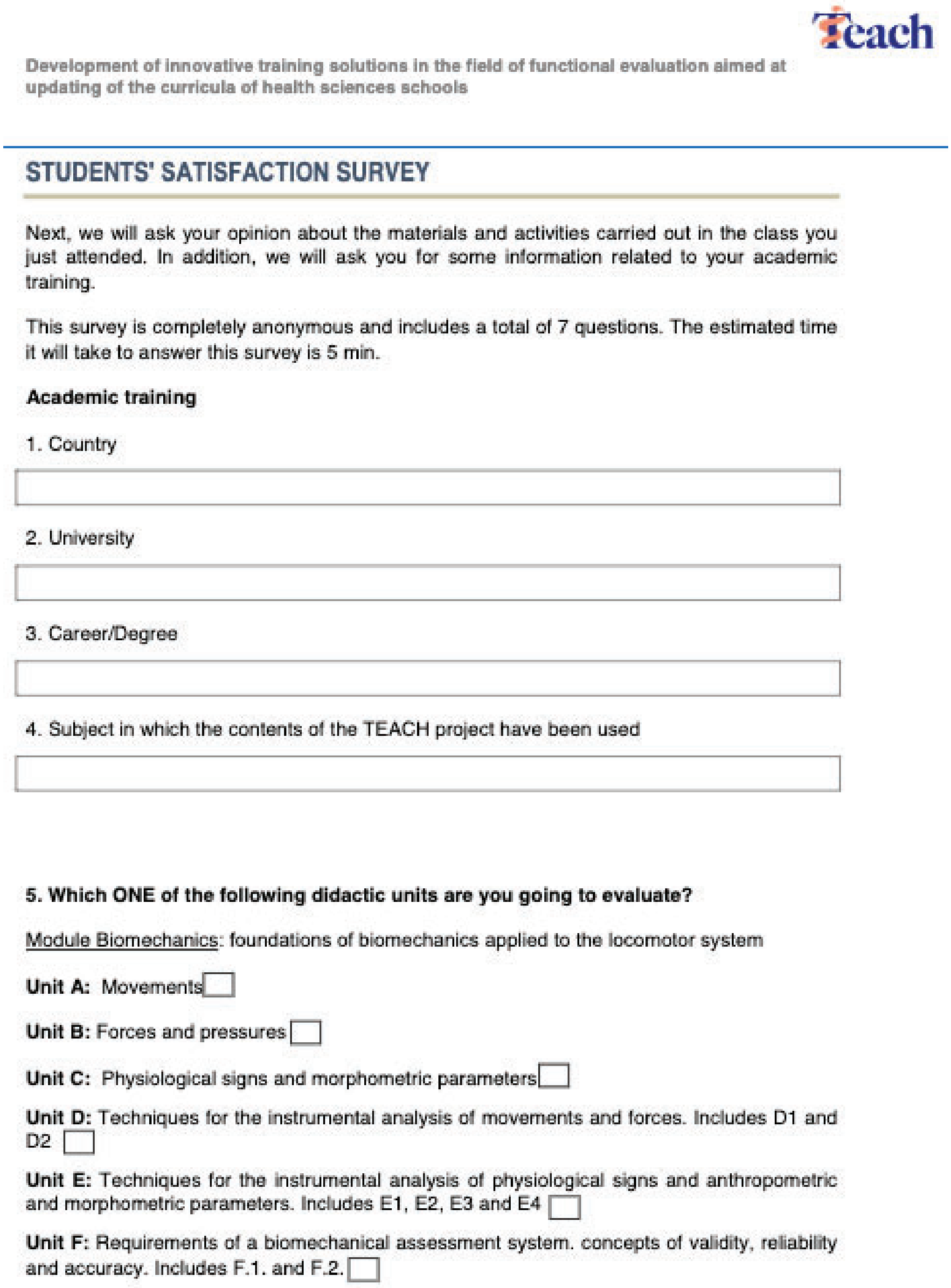


Fig. 3. The percentage of participants from all universities
Rys. 3. Procentowy udział uczestników z badanych uniwersytetów
Own elaboration

For the verification study, a questionnaire form was developed to measure course participant satisfaction, particularly interest in the topics and the relevance of each module, that overall look is presented in the Fig. 4.



Teach

Development of innovative training solutions in the field of functional evaluation aimed at updating of the curricula of health sciences schools

STUDENTS' SATISFACTION SURVEY

Next, we will ask your opinion about the materials and activities carried out in the class you just attended. In addition, we will ask you for some information related to your academic training.

This survey is completely anonymous and includes a total of 7 questions. The estimated time it will take to answer this survey is 5 min.

Academic training

1. Country

2. University

3. Career/Degree

4. Subject in which the contents of the TEACH project have been used

5. Which ONE of the following didactic units are you going to evaluate?

Module Biomechanics: foundations of biomechanics applied to the locomotor system

Unit A: Movements ☐

Unit B: Forces and pressures ☐

Unit C: Physiological signs and morphometric parameters ☐

Unit D: Techniques for the instrumental analysis of movements and forces. Includes D1 and D2 ☐

Unit E: Techniques for the instrumental analysis of physiological signs and anthropometric and morphometric parameters. Includes E1, E2, E3 and E4 ☐

Unit F: Requirements of a biomechanical assessment system. concepts of validity, reliability and accuracy. Includes F.1. and F.2. ☐

Fig. 4. An excerpt from the prepared questionnaire
Rys. 4. Fragment ankiety weryfikacyjnej
https://teach.ibv.org/

17.3. Results and discussion

The results of the survey allowed for quantitative analysis of interest in the course. Moreover, the obtained data allowed for the assessment of the attractiveness of individual modules and the topics included in them.

As illustrated by the data presented in Fig. 5, the module on concepts and methodology in the functional evaluation was the most popular course (Module 1). This course was evaluated by 255 students of health sciences schools. Important information from the point of view of quantitative analysis of the received data would be the information about the course and the semester studied. Then the number of questionnaires of respondents per 100 students of a given course would allow us to obtain absolute values indicating the actual interest in the chosen course (for future use).

The second group that participated in the survey was the academic staff, who, like the students, were asked to complete a survey on how they rated their satisfaction with the course. The first question on the survey was which module you rated. The number of responses in terms of course evaluations mapped the degree of interest in a particular course. Figure 5 presents quantitative data in terms of course modules evaluated. As can be seen from the data presented, the module on the foundations of the biomechanics of the locomotor system "Module Biomechanics: foundations of biomechanics applied to the locomotor system" was the most popular among the group of academic teachers.

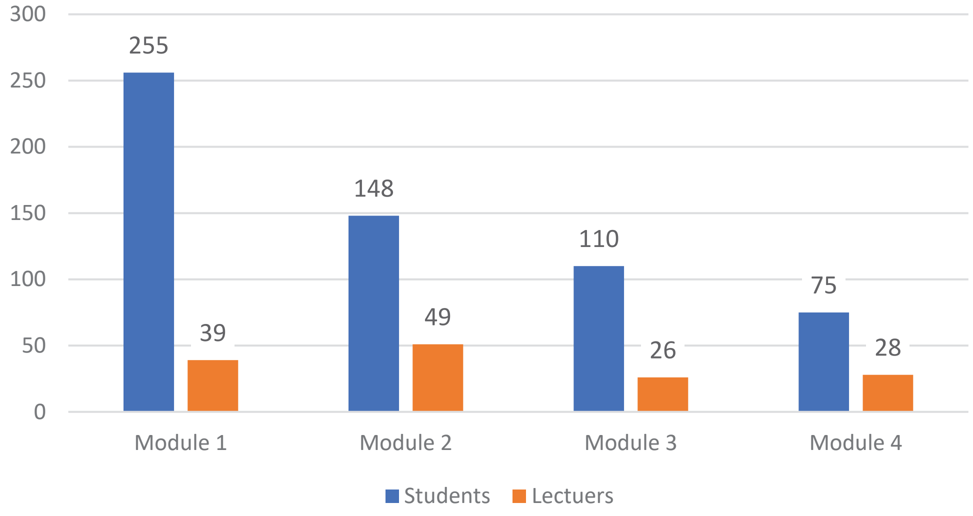


Fig. 5. Number of participants by choice of course modules
Rys. 5. Liczba uczestników w podziale na wybór modułów kursu
Legend: Modul 1: Functional Evaluation: Concept and Methodology; Module 2 Foundations of Biomechanics applied to the Locomotor System; Module 3: Biomechanics of Spine; Module 4: Biomechanics of Gait.

As can be seen from the data presented in Table 1, the greatest interest among students was: Classification of activities and functions according to the International Classification of Functioning, Disability, and Health (ICF) – 28%, The importance of cognitive abilities in performing motor tasks and why it is important to include biomechanical analysis in cognitive impairments – 19%; Importance of functional assessment and its application – 18%; Classification of functional impairments and disabilities – 16%; Socio-health impact of disability. Disability in the workplace – 11%; Functional assessment analysis: classical assessments versus instrumental analysis – 8%.

Table 1

Results of interest for individual chapters of Module 1 "Functional Evaluation: Concept and Methodology"

Name of units	Country				
	Colombia	Georgia	Poland	Spain	Total
Unit A: Classification of activities and functions according to the ICF		2	68	3	73
Unit B: Importance of functional assessment and its applications		1	42	2	45
Unit C: Classification of functional impairments and disability		2	36	2	40
Unit D: Socio-health impact of disability. Disability in working places		1	26	1	28
Unit E: Importance of cognitive abilities in the performance of motor tasks and why it is important to include biomechanical analysis in cognitive impairments	1	1	23	24	49
Unit F: Functional evaluation assessment: classical assessments vs instrumented analysis			19	1	20
Total	1	7	214	33	255

As in the case of lecturers, the Module Functional Evaluation: Concept and Methodology were chosen as the second most interesting. The first one was Module Foundations of Biomechanics applied to the Locomotor System.

As can be seen from the data presented in Table 2, the most interesting chapter in module one was the chapter on the mechanics of body motion (dynamics). Among all the chapters in this module, the first chapter was selected most often – 21 times, which is 43% of the submitted evaluations.

Table 2

Results of lecturers' interests in particular chapters of Module 2 "Biomechanics: foundations of biomechanics applied to the locomotor system"

Name of units	Country		
	Poland	Spain	Total
Unit A: Movements	11	10	21
Unit B: Forces and pressures	4	4	8
Unit C: Physiological signs and morphometric parameters	3	2	5
Unit D: Techniques for the instrumental analysis of movements and forces. Includes D1 and D2	2	2	4
Unit E: Techniques for the instrumental analysis of physiological signs and anthropometric and morphometric parameters. Includes E1, E2, E3 and E4	4	1	5
Unit F: Requirements of a biomechanical assessment system. Concepts of validity, reliability and accuracy. Includes F1 and F2	6		6
Total	30	19	49

Among students, the FE module was the most popular, which may be due to the growing interest in patient diagnosis, which is based on the overall functioning of the body, rather than on a selected disease entity.

It is quite surprising that teachers' interest is directed towards aspects of the biomechanics of movement rather than FE approaches. This may be related to a certain pattern of education developed over the years in which FE topics were not popular.

17.4. Conclusion

The utilitarian goal of the e-learning course of FE is to update and reinforce the skills of the next generation of graduates in health sciences schools. The specific innovative training program was developed taking into account the wide spectra of needs and requirements of health sciences teachers from all over Europe and always regarding the future scenario of European health.

As indicated by the results of the international survey with the participation of students and teachers, the most interesting topic among students was Modul containing knowledge about Functional Evaluation: Concept and Methodology. It means that the FE area is a desirable component for future medics. For this reason, FE should be implemented into the curricula of health sciences schools.

Acknowledgements

This work has been done under the project "Development of innovative training solutions in the field of functional evaluation aimed at updating of the curricula of health sciences schools", contract number 2018-1-PL01-KA203-051084. This project has been funded with support from the European Commission. This communication reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Bibliography

1. Eurostat: *Population structure and ageing*, 2016: http://ec.europa.eu. [Article available: 18.05.2022].
2. Eurostat: *Healthy life years statistics*, 2016: http://ec.europa.eu. [Article available: 18.05.2022].
3. Eurostat: *Functional and activity limitations statistics*, 2015: http://ec.europa.eu. [Article available: 18.05.2022].
4. J.J. Chen, Functional Capacity Evaluation & Disability, *The Iowa Orthopaedic Journal* (2007) **27**:121-127.
5. M.E. Geisser, M.E. Robinson, Q. Miller, M. Quaintance, S.M. Bade, (Psychosocial Factors and Functional Capacity Evaluation Among Persons with Chronic Pain, *Journal of Occupational Rehabilitation* (2003) **13**:259–276.
6. J.J. Chen, Functional Capacity Evaluation & Disability, *The Iowa Orthopaedic Journal* (2007) **27**:121–127.
7. L. Chia-Ming, C. Ching-I, Y. Wen-Ruey, Y. Winnie, H. Chih-Cheng, C. Ching-Yu, Enhancing elderly health examination effectiveness by adding physical function evaluations and interventions, *Archives of Gerontology and Geriatrics*, (2017) **70**:38–43

8. A. Dixon-Ibarra, G. Krahn, H. Fredine, A. Cahill, S. Jenkins, Adults aging ‘with’ and ‘into’ paralysis: Epidemiological analyses of demography and health, *Disability and Health Journal* (2016) **9**:575–583.

FUNCTIONAL EVALUATION IN PSYCHOTHERAPY EDUCATION. INTERNATIONAL RESEARCH WITH USE E-LEARNING PLATFORM

Abstract

This paper presents findings on the importance of functional evaluation (FE) as one of the teaching topics for therapists against the background of an international European project. The problem of the aging population becomes the basis for activities in predicting and preventing functional limitations and activity limitations of people of different ages. From the point of view of health sciences, it seems reasonable to implement FE as an element of the educational path integrating a comprehensive approach to diagnosis and treatment of physical disability. Presented results indicate interest in the mentioned topics among students, lecturers and physiotherapists from many European countries and others.

Keywords: functional evaluation, physiotherapy, e-learning, education

POSTĘPY W ONKOLOGII OBLICZENIOWEJ I SPERSONALIZOWANEJ MEDYCYNIE

TOM 2
WYZWANIA PRZYSZŁOŚCI

Streszczenie

Druga książka z serii wydawniczej *Postępy w Onkologii Obliczeniowej i Spersonalizowanej Medycynie* kontynuuje raportowanie doniesień o najnowszych badaniach i rozwiązaniach stosowanych w obszarze onkologii obliczeniowej i medycyny spersonalizowanej. Zaprezentowane rozdziały są wynikiem badań prowadzonych zarówno przez pracowników Politechniki Śląskiej jak i partnerów zarówno ze świata akademickiego i otoczenia społeczno-gospodarczego. Poruszana tematyka obejmuje badania związane z bioinformatyką (m.in. przetwarzanie obrazów medycznych, klasyfikacja, badania związane z nowotworami), modelowaniem numerycznym w zastosowaniach biomedycznych, e-learningiem (bio)inżynierią materiałową.

Nowe idee i aplikacje prezentowane w ramach serii ukierunkowane są na wdrożenia nowych narzędzi diagnostycznych, leków oraz terapii, które pomogą klinicystom w ich praktyce i ostatecznie doprowadzą do wzrostu jakości życia.

Niniejsza wieloautorska monografia podąża za najnowszymi trendami inżynierii biomedycznej, w szczególności związanymi z onkologią obliczeniową i medycyną spersonalizowaną. Praca zawiera 17 rozdziałów, w których Autorzy dzielą się z Czytelnikiem swoim doświadczeniem i wiedzą w odniesieniu do następujących zagadnień:

a) *bioinformatyki*: przetwarzanie obrazów medycznych (m.in. mikrotomografia w badaniach histopatologicznych, analiza obrazów hodowli komórkowych, przygotowanie obrazów medycznych pacjentów COVID-19 do klasyfikacji AI); systemy klasyfikacji (np. wyodrębnianie cech, wizualizacja wielkich zbiorów danych); onkologiczne (np. mutacje komórek rakowych, algorytmy ekspresji genów); analiza rekurencyjna,

b) *modelowania numerycznego* (m.in. mrożenie komórek rakowych, mechanika płynów układu krążenia),
c) *biomateriałów* (m.in. biofunkcjonalizacja powierzchni) oraz określania *właściwości tkanek*,
d) *e-learningu* (w zastosowaniu do ewaluacji kursów e-learningowych w warunkach pandemicznych).

Publikacja monografii jest jednym z działań realizowanych przez Politechnikę Śląską. Wydawana jest m.in. w ramach programu Inicjatywa Doskonałości – Uczelnia Badawcza, a jej tematyka doskonale wpisuje się w temat Priorytetowego Obszaru Badawczego 1 (POB1; Onkologia Obliczeniowa i Spersonalizowana Medycyna) Politechniki Śląskiej.

WYDAWNICTWO POLITECHNIKI ŚLĄSKIEJ
ul. Akademicka 5, 44-100 Gliwice
tel. (32) 237-13-81, faks (32) 237-15-02
www.wydawnictwopolitechniki.pl

UIW 48600

Sprzedaż i Marketing
tel. (32) 237-18-48
wydawnictwo_mark@polsl.pl

Sprawy wydawnicze
tel. (32) 237-13-81
wydawnictwo@polsl.pl

Nakł. 100 + 44 Ark. wyd. 21 Ark. druk. 16,625 Papier 80 g

Zam. 10/23
Monografia 969

The second book in the series continues to report recent research and applied solutions in the area of computational oncology and personalized medicine, being the topic of the Priority Research Area #1 (POB1) of The Silesian University of Technology. The chapters are the result of the research conducted by the staff of SUT and by the partners from both the academic world and socio-economic environment. The topics covered include research related to bioinformatics (including medical image processing, classification, and cancer-related research), (bio)materials engineering, numerical modelling in biomedical applications and e-learning.

The editorial team believe that the new ideas and applications presented within the series would lead to the implementation of new diagnostic tools, drugs or therapies to assist clinicians in their practice and eventually would lead to an increase in life quality.

ISBN 978-83-7880-876-3
Wydawnictwo Politechniki Śląskiej

Series: Excellence Initiative - Research University

Silesian
University
of Technology

RESEARCH
UNIVERSITY
EXCELLENCE INITIATIVE