I T H E A

International Journal

INFORMATION TECHNOLOGIES & KNOWLEDGE

2016 Volume 10 Number 1

International Journal
INFORMATION TECHNOLOGIES & KNOWLEDGE

Volume 10 / 2016, Number 1

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" (IJ ITK)
is official publisher of the scientific papers of the members of
the ITHEA International Scientific Society

IJ ITK rules for preparing the manuscripts are compulsory.
The **rules for the papers** for IJ ITK are given on www.ithea.org

Responsibility for papers published in IJ ITK belongs to authors.

International Journal "INFORMATION TECHNOLOGIES & KNOWLEDGE" Volume 10, Number 1, 2016
Edited by the **Institute of Information Theories and Applications FOI ITHEA**, Bulgaria, in collaboration with:
Institute of Mathematics and Informatics, BAS, Bulgaria; V.M.Glushkov Institute of Cybernetics of NAS, Ukraine;
Universidad Politechnika de Madrid, Spain; Hasselt University, Belgium;
St. Petersburg Institute of Informatics, RAS, Russia; Institute for Informatics and Automation Problems, NAS of the Republic of Armenia.
Printed in Bulgaria
Publisher ITHEA®
Sofia, 1000, P.O.B. 775, Bulgaria. www.ithea.org, e-mail: office@ithea.org
Technical editor: Ina Markova

Издател: ИТЕА®, София 1000, ПК 775, България, www.ithea.org, e-mail: office@ithea.org

ISSN 1313-0455 (printed) **ISSN 1313-048X (online)**

TERM MATCHING APPROACHES IN E-COMMERCE

Mariam Gawich, Marco Alfonse, Mostafa Aref, Abdel-Badeeh M. Salem

***Abstract**: Information of a particular domain is presented in different terms with different structures. Terms can be found in unstructured text, semi structured forms (e.g. HTML and XML) and structured forms (e.g. databases and ontology). Computer can understand the semantic of a term from its relations with other terms. Term matching approaches are applied to detect the similar terms which yields to discover and present new integrated information. This paper investigates the term matching approaches applied in e-commerce.*

***Keywords**: Term Matching, String Similarity, Ontology Mapping, E-Commerce.*

***ACM Classification Keywords**: F.2.2 Nonnumerical Algorithms and Problems - Pattern matching*

Introduction

Term matching (TM) [Strzalkowski,1999] (also called term conflation) is considered as a field of text processing which belongs to information retrieval system. The objective of term matching is to cluster textual fragments in a document to similar terms or concepts. In addition, term matching is considered as a component of text analysis [Staab & Studer,2009] which is a stage of the process of information extraction. Term matching types are classified into syntactic matching and semantic matching. Syntactic matching [Giunchiglia & Shvaiko,2003] focuses on the string similarity between terms based on their form and it involves the syntax driven approaches. Semantic matching concerns about the interpretation of term meaning. The meaning of a term can be detected by its association with external knowledge sources such as semantic network, lexicon and ontology.

E-commerce is generally defined as the process of buying and selling goods and services on the internet. If there are two organizations want to make a deal on the internet, they will face a difficulty to understand each other if they don't have a homogeneous terminology. The matching between terms can solve this problem. The objective of this paper is the investigation of term matching approaches applied in e-commerce. This paper is organized as follows; section 1 introduces the term matching approaches in e-commerce, section 2 presents a framework proposal for comparative study and the last section contains the conclusion.

1. E-Commerce Term Matching Approaches

There are two main approaches for term matching in e-commerce which are the automated approach to product taxonomy mapping and the LexOnt matching approach. This section discusses in details the two approaches.

1.1 Automated Approach to Product Taxonomy Mapping in E-Commerce.

Lennart Nederstigt and colleagues proposed this approach [Nederstigt et al ,2014] to match two heterogeneous taxonomies provided by two ontologies. The algorithm receives two inputs; the first input is category taxonomy and its paths of the source taxonomy (the path is a list of nodes that start from root and end to the current node), while the second input involves the categories of target taxonomy. The approach was executed following these stages:

- **Preprocessing of category name**

The algorithm decomposes the name of a category of the source taxonomy into ampersands, comma. The result of this preprocessing is a set of multiple terms called split term set.

- **Word sense disambiguation**

This approach is derived from Park and Kim approach [Park & Kim,2007]. Its objective is to identify the correct meaning of a term presented by a leaf node (node which doesn't have a child) in the source taxonomy. The algorithm uses the Wordnet [Miller,1995] to search about the term provided by the split term set. The approach executes a comparison between the hyponyms of the term provided by the Wordnet and the same term provided by the split term set provided by the source taxonomy. The result of this process is an extended term set that is composed of original term and its synonyms. In order to detect the correct closely meaning of each term, the algorithm compares the sense hierarchy (hypernym relations) of the term provided by Wordnet with all ancestor nodes (upper category nodes which have children nodes) of the term provided by the current node located in the taxonomy source. The result of this comparison is a set of matched lemmas. To measure the convenience between each upper category with all sense hierarchy provided by the set of matched lemma, a similarity function [Aanen et al,2015] is applied as shown in equation 1.

$$\text{HyperProximity}(t,S)= \begin{cases} \dfrac{1}{\min_{x \in C}(\text{dist}(x,l))} & if\ C \neq \emptyset \\ 0 & if\ C=\emptyset \end{cases} \quad (1)$$

Where dist refers to the number of edges (connection between node and another), t denotes the upper category (ancestor), S denotes the sense hierarchy, l denotes the leaf located in the sense hierarchy S, C is a set of matching lemmas and x is a matching lemma. The output of this function is the matched lemma which its hypernym in Wordnet has the shortest distance to the leaf of the sense hierarchy. The average of hyper proximity is calculated to determine the overall similarity between the source category path and sense hierarchy.

- **Candidate path identification**

In this stage, the extended term set is used to determine the candidate path of the target taxonomy to be mapped with the current source category. The algorithm matches the terms provided by the extended split term set with paths provided by the target taxonomy. If one term of the extended term set is found to be a substring of the examined category of target taxonomy, the category is considered as candidate path. The candidate path identification is derived from Park and Kim algorithm [Park & Kim,2007] which compares the root node of the target path with the extended term set. The difference between the Park and Kim algorithm and proposed algorithm is that the later if it detects a term that is a substring of the actual tested category, category will be considered as candidate path and the algorithm will check the children of this category. Moreover, the proposed algorithm splits the original category name to multiple sets if it is a composite category. The algorithm compares between the multiple extended term set and the actual tested category name. This comparison requires the matching between every extended term set with its extended split term set. The result of the comparison is a boolean value true, if a term is a substring of the actual tested category or false, if there is no term can be a substring of the actual tested category. The path of the target category will be considered as a candidate path if half of boolean values are true.

- **Aggregated path similarity score**

To determine the best candidate paths of target taxonomy that can be adapted to match with source paths, an aggregated similarity score is calculated for each candidate path. Its objective is to measure the adaptation between target candidate path and the source path. The aggregated function is based on the use of Park and Kim algorithm [Park & Kim,2007] and the parent mapping similarity presented by the proposed algorithm. The aggregated similarity score consists of the cooccurence and order consistency measure. The cooccurence similarity [Aanen et al,2015] is based on lexical matching to detect the overlap between the category of target candidate path and the category of the source path regardless of their nodes order. Cooccurence similarity consists of Levenshtein similarity [Levenshtein,1966] and Jaccard similarity [Jaccard, 1912] .Order consistency is calculated to detect the common nodes that share the same order in the taxonomy hierarchy. It consists of the common PrecRel and consistent functions. The common function adds the node that match the category name

of the source taxonomy or its synonyms with another node of the target taxonomy or its synonyms. The synonyms are provided by Wordnet. The precRel function takes each common node to generate binary associations that indicates the order of relation between nodes (precedence relation). The consistent function determines if both of the two categories which are located in the candidate target path and two categories which are located in the source path have the same precedence relation. If they have the same precedence relation, the result of this function will be 1 and if they don't have the same precedence relation, the result of this function will be 0.

Concerning the evaluation, the algorithm is compared to Anchor Prompt algorithm [Noy& Musen,2004] and Park and Kim algorithm [Park & Kim,2007] .They are tested by three real life datasets; the first is provided by Open directory Project (ODP)- dmoz.org that consists of 44.000 categories, the second dataset is provided by Amzaon.com which more than 9.500 different categories are chosen for the evaluation and the third dataset is provided by Overstack.com that consists of 1.000 categories.

Table 1 shows that the automated approach achieves the highest value of recall (83%) and F1 (66%) measure compared to anchor prompt and Park and Kim. For precision, automated approach has a precision value less than Park and Kim. The high value of recall indicates that the automated approach can work on composite categories.

Table 1.Average results per algorithm [Nederstigt et al ,2014]

Algorithm	Precision	Recall	F1-measure	Computation time(s)
Anchor-PROMPT	28.93%	16.69%	20.75%	0.47
Park and Kim	47.77%	25.18%	32.52%	4.99
Lennart Nederstigt	38.28%	83.66%	52.31%	20.71

1.2 LexOnt Matching Approach

The objective of LexOnt approach [Arabshian et al,2012] is the production of frequent and significant terms provided by the corpus of Programmable Web (PW) directory [Programmable Web,2015]. Frequent and significant terms reflect the general properties of service classes provided by the PW to be automatically classified in ontology. The corpus of PW directory contains API description. The

corpus is encoded as HTML format. LexOnt algorithm relies on the information provided by the HTML text which describes the APIs service and information provided by Wikipedia that describe the domain of the service. Moreover LexOnt uses the Wordnet [Miller,1995] to detect the synonyms of terms to produce top N list words and phrases which can be used to determine distinct features of the service. LexOnt provides a semi-automatic ontology construction.

LexOnt Approach is executed by several algorithms that are outlined in the following stages:

- **TF-IDF (Term Frequency- Inverse Document Frequency)**

TF-IDF is calculated to demonstrate the importance of term appeared in the corpus. TF [Salton,1983] is defined as the frequency of a term t appeared in the corpus while IDF is the inverse document frequency which can be defined using this expression: [log(N/(nj+1))+1] where N is the total number of document and nj is the document frequency of term (t).

- **Significant phrases**

A significant phrase is composed of two or more words that can be a clue that indicates the high level property of a service class. For example, in the service 'Advertising' significant terms are 'Mobile Advertising', 'Facebook Advertising', etc. Significant phrase is detected through two steps; the first step is the determination of collocation, terms that occur together and the second step is the selection of unique collocations. The Chi Square is computed in this phase on collocated words to show the comparison between the numbers of times that words in a phrase are appeared together and the number of times that words appear alone. LexOnt uses the Wikipedia, Wordnet and constructed ontology to cover the main concepts and properties for an API service. For example, for the 'Advertisng' category, LexOnt algorithm will generate 20 top words provided by Wikipedia page which are (advertising, marketing, brand, television, semiotics, advertisement, billboard, radio, product, bowl, sponsor, consumer, advertise, placement, super, logo, commercial, infomercial, message, promotion). In addition, LexOnt applies the use of Wordnet to find the synonyms and related terms for each term listed in top N words. Also, LexOnt applies the matching between terms which are located in the constructed ontology and the generated terms. If there exist matched terms, LexOnt algorithm will rank and label them to mark that they are already existed in the ontology.

Tables 2, 3 and 4 demonstrate the evaluation of LexOnt; table 2 demonstrates the calculation of precision and recall of TF-IDF terms and the generation of significant phrases, table 3 demonstrates a second evaluation that involves the calculation of percentage of terms provided by external knowledge base (Wikipedia, Wordnet and Ontology) and table 4 shows a third evaluation which finds the percentage calculation of matched terms. For the three evaluations, the (Advertising, Real State, Social, Travel and Utility) categories are chosen for LexOnt testing. The (Advertising, Real State)

categories are selected upon the number of API (average of 40 API), specifity (has Wikipedia page) and prior knowledge of the domain (the co-author of the ontology should have a background about the service domain) . The (Social, Travel) categories are selected according to the familiarity of ontology creator with the service domain. The (Utility) category is chosen according to its number of APIs (65 APIs) and has no matched Wikipedia page. The equal terms, similar terms and different terms are illustrated in table 4.

Table 2. Precision/Recall Stats [Arabshian et al,2012]

Category	Sig.Phrase	TF-IDF	Recall
Advertising	3.98%	2.77%	43.88%
Real Estate	1.02%	.92%	9.57%
Social	3.21%	2.8%	20.19%
Travel	1.96%	2.4%	30.91%
Utility	9.58%	3.83%	34.91%

Table 3.Percentage of terms used from KB [Arabshian et al,2012]

Category	Sig.Phrase	TF-IDF
Advertising	41.38%	52.73%
Real Estate	100%	100%
Social	31.90%	11.38%
Travel	82.26%	72.73%
Utility	0%	0%

Table 4. Term Usage [Arabshian et al,2012]

Category	Equal Terms	Similar Terms	Different terms
Advertising	85.71%	100%	65.71%
Real Estate	16.67%	91.67%	66.67%
Social	1.73%	86.9%	79.1%
Travel	6.25%	100%	2.3%
Utility	5%	60%	50%

2. Framework Proposal for Comparative Study

The points of comparison that are used to point out differences between term matching approaches are; input, matching approach type, evaluation and output. Table 5 demonstrates the comparison between approaches.

Table 5. A Comparison betweeen Term Matching Approaches

Approach Name	Input	Matching Approach Type	Evaluation	Output
Automated approach to product taxonomy Mapping	Two ontologies	Semantic technique, statistical measure and substring technique	Yes	Mapping terms and candidate paths
LexOnt	Corpus, Wikipedia page and Ontology	String based technique and semantic matching	Yes	Significant terms and candidate terms for the ontology enrichment

First for the automated approach to product taxonomy mapping in e-commerce, its inputs are two ontologies in the domain of interest. The implementation of matching relies on several techniques; the semantic technique applied by the use of Wordnet and syntactic techniques that involve substring approach that detects the candidate path between target taxonomy and current source. The approach evaluation is based on its comparison with anchor prompt and Park and Kim algorithm. Second for LexOnt, its inputs are a corpus provided by the PW directory, Wikipedia page and constructed ontology in the domain of interest. The string matching approach is executed between word provided by API and top Wikipedia word. The semantic matching is applied through the Wordnet to detect the synonym of term.

Conclusion

Both of automated approach to product taxonomy mapping in e-commerce and LexOnt approach rely on the syntatic matching that focus on the structure of the word. Moreover, both of them execute the semantic matching by the use of Wordnet. Only LexOnt executes the mapping that focuses on the significant terms and suggests these to enrich the ontology which is considered as a technique for knowledge representation that can be used by other tools.

Bibliography

[Aanen et al,2015] SS.Aanen, D.Vandic and F. Frasincar. Automated product taxonomy mapping in an e-commerce environment. Expert Systems with Applications 42.3 (2015): 1298-1313, 2015.

[Arabshian et al,2012] K.Arabshian, P.Danielsenand and S.Afroz. LexOnt: A Semi-Automatic Ontology Creation Tool for Programmable Web. In: AAAI Spring Symposium: Intelligent Web Services Meet Social Computing. 2012.

[Giunchiglia & Shvaiko,2003] F.Giunchiglia and P. Shvaiko. Semantic matching. April 2003 Technical Report # DIT-03-0. University of Trento.Italy,2003.

[Jaccard, 1912] P.Jaccard. The distribution of the flora in the alpine zone." New phytologist 11.2, 1912.

[Levenshtein,1966] V. I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals." *Soviet Physics Doklady*. Vol. 10, 1966.

[Miller,1995] G.A.Miller. WordNet: A Lexical Database for English. Communications of the ACM 38(11), 39–41 .1995.

[Nederstigt et al ,2014] L.Nederstigt,D.Vandic and F.Frasincar. Automated Product Taxonomy Mapping In An E-Commerce. Erasmus University Rotterdam, 2014.

[Noy& Musen,2004] Natalya F. Noy and Mark A. Musen Using prompt ontology-comparison tools in the EON ontology alignment contest." Proceedings of the Third International Workshop Evaluation of Ontology-based Tools (EON). 2004

[Park & Kim,2007] S.Park and W.Kim. Ontology Mapping between Heterogeneous Product Taxonomies in an Electronic Commerce Environment. International Journal of Electronic Commerce 12(2), 69–87, 2007.

[Programmable Web,2015] Programmable Web, http://www.programmableweb.com/. (accessed: 4.05.2015).

[Salton,1983] G.Salton. Introduction to Modern Information Retrieval.McGraw-Hill, 1983.

[Staab & Studer,2009] S.Staab and R.Studer. Handbook On Ontologies.Part V- Ontology-Based Infrastructure and Methods.Page 663-666.Berlin: Springer, 2009. Print.

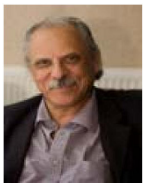
[Strzalkowski,1999] T. Strzalkowski. Natural Language Information Retrieval. NLP For Term Variant Extraction: Snergy Between Morphology, Lexicon And Syntax,1999.

Authors' Information

Mariam Gawich *is a PhD student* at the Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt.

Dr.Marco Alfonse Tawfik *is a Lecturer at the Faculty of Computer and Information Science, Ain Shams University, Cairo, Egypt. He got Ph.D. of Computer Science since August 2014, University of Ain Shams. His research interests: Semantic Web, Ontological Engineering, Medical Informatics, and Artificial Intelligence. He has published around 20 publications in refereed international journals and conferences.*

Prof. Mostafa M. Aref *is Professor of Computer Science, Ain Shams University, Cairo Egypt. He got Ph.D. of Engineering Science, June 1988, University of Toledo, Toledo, Ohio, USA. He has more than 50 journal and conference publications. His research areas are Natural Language Processing, Knowledge Representation, Object-oriented Programming, Ontology and Real-Time Strategy Games.*



Prof. Dr. Abdel-Badeeh M Salem *He is a Professor of Computer Science since 1989 at Ain Shams University, Egypt. His research includes intelligent computing, knowledge-based systems, biomedical informatics, and intelligent e-leaming. He has published around 250 papers in refereed journals and conferences. He has been involved in more than 400 Confs and workshops as a Keynote Speaker, Scientific Program Committee, Organizer and Session Chair. He is a member of many national and international informatics associations.*

PROPOSED MOBILE SYSTEM FOR BONE VISUALIZATION USING GPU RAYCASTING VOLUME RENDERING TECHNIQUE

Yassmin Abdallah, Abdel-Badeeh M.Salem

Abstract: *Despite the rapid development of mobile phones, the existence of medical visualization applications on stores is rare. This paper represents iOS mobile application that helps medical students and doctors to show bones processed over in an orthopedic surgery. The mobile application implements GPU ray casting algorithm. Experimental results obtained visually by comparing the result to the result obtained from ImageVis3D which use slice based volume rendering technique. ImageVis3D's development was initiated in 2007 by the NIH/NCRR Center for Integrative Biomedical Computing and additionally supported by the DOE Visualization and Analytics Center for Enabling Technologies at the SCI Institute. The result shows that proposed application based on GPU ray-casting technique present better visualized result than ImageVis3D. The dataset used in the experiment is CT images obtained from Osirix datasets for surgical repair of facial deformity.*

Keywords: *mHealth, volume visualization, mobile-based biomedical computing , CT images , GPU ray casting*

ACM Classification Keywords: *D.2: SOFTWARE ENGINEERING , I.3: COMPUTER GRAPHICS , J.3: LIFE AND MEDICAL SCIENCES*

Introduction

During common orthopedic surgery training, students must learn how to perform numerous surgical procedures like fixing fractures which requires training on artificial bones with the usage of surgical tools and implants. These artificial bones have a high cost that depends on the bone's type and quality. Thus the idea of using a computer based simulators for orthopedic surgery training appeared. Simulators will decrease the cost and help students to practice various procedures on a large number of available simulated surgeries in a safe and controlled environment. Also as they are designed dedicatedly for training with specific training goals and high quality visualization of bones, they will have much more to offer than artificial patterns. Visual representation of the bone is the key element in orthopedic surgery simulation. Nowadays and with the rapid development of mobile phones that can be similar to laptops, mobile devices became the trend toward information systems and ubiquitous graphical devices and native volume rendering due to their rapid development in the graphics hardware which can be similar

to the PCs. GFXBench [GFXBench, 2016] is a high-end graphics benchmark that measures mobile and desktop performance with next-gen graphics features across all platforms. GFXBench used to compare the current apple laptops, iPhones and iPad to show the rapid development of mobile and tablets devices that in some cases can beat laptops performance. Figure 1 and 2 shows a comparison based on CPU speed between MacBook pros, MacBook, iPhone and iPad. As shown the speed of CPU in mobile phones and tablets begin to be as the same as laptop.

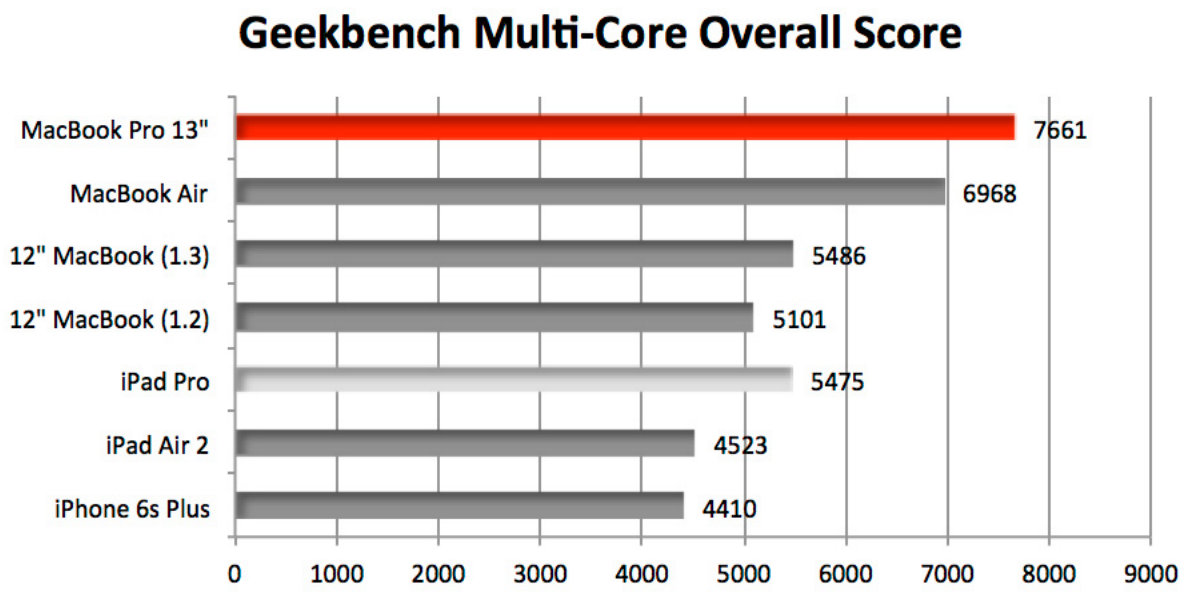


Fig 1 Geekbench single core overall score

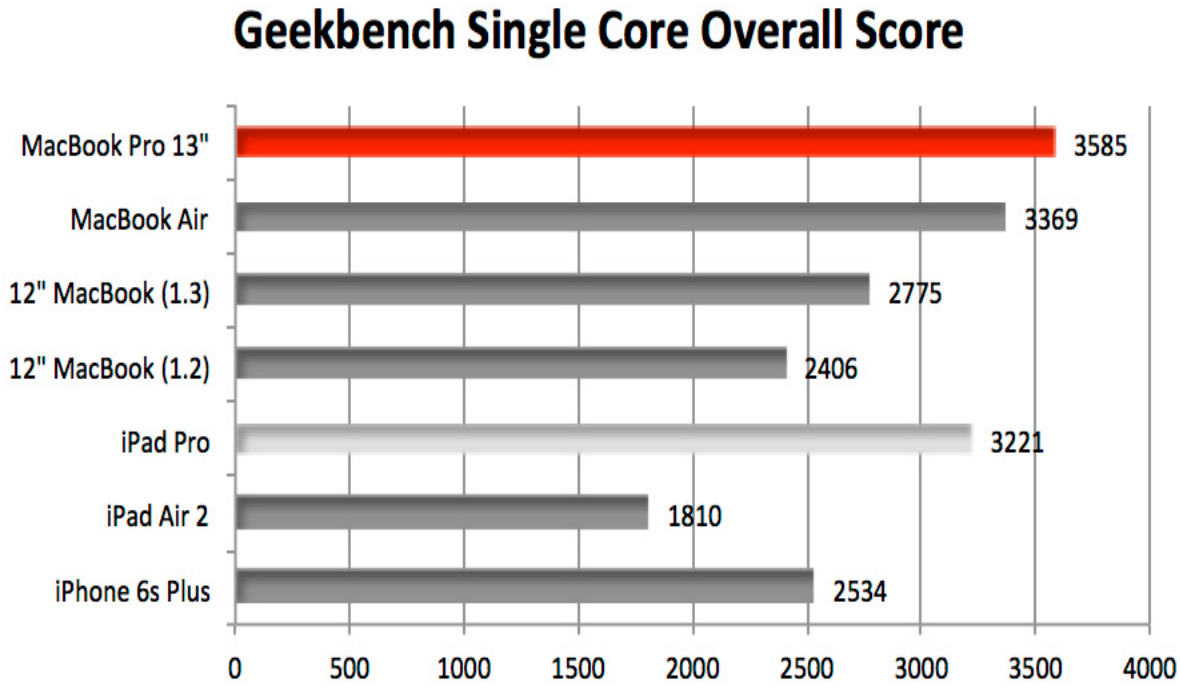


Fig 2 Geekbench Multi-core overall Score

Despite that rapid development, mobile stores has a rare free open source applications that implements the core of medical visualization. This paper presents an iOS mobile application for bone visualization that help medical students to visualize bones.

Paper organized as follows, firstly related work for volume visualization field listed. Secondly methodology used in this paper shown. Thirdly, experimental results shown. Finally conclusion and future work illustrated.

RELATEDWORK

There is two approach for visualization on mobile: first one is server-based rendering where datasets being render in server machine and the result image streamed to the mobile client, the second one is on-device rendering where datasets rendering in mobile directly.

The server-based approach appear due to the limitation on the mobile device and memory issues. [Lamberti and Sanna, 2007] present a powerful render server generates visualization images that are subsequently streamed to the mobile device for display. The advantage of this approach is that the rendering performance can be scaled by scaling the server capabilities and that very large datasets can be visualized.

Unfortunately, it suffers from a potentially substantial degradation of the interactive user experience, depending on the net-work quality [N. Tolia _et_al, 2006]; in particular, network latency on interactive responses can cause serious problems.

Therefore, on-device rendering on the mobile devices is preferable if fast interactive feedback is required and the net-work connection is not optimal. Fortunately, hardware development, along with basic texture mapping capabilities, is becoming increasingly available on mobile devices. [M.Moser and D.Weiskopf, 2006] propose a screen-adaptive hybrid low/high resolution rendering technique that achieves a good compromise between image quality and interactivity. [J.M.Noguera _et_al, 2012] develop a novel volume rendering algorithm perfectly suited to modern GPU-enabled mobile device. It addressed the limitations of mobile device, mainly the lack of 3D texture support. [Mobeen _et_al, 2012] represent two volume rendering algorithms using the WebGL platform for implementing medical image visualization on the mobile devices.

Mobile stores have a few open source application that present volume rendering techniques. Such as, [ImageVis3D, 2016], which will be concentrate on in this paper, is a lightweight open source, volume rendering application for interactive visualization on the iPhone and iPod Touch. It can display the built in visible human datasets as well as any dataset shared by the ImageVis3D application. ImageVis3D's development was initiated in 2007 by the NIH/NCRR Center for Integrative Biomedical Computing and

additionally supported by the DOE Visualization and Analytics Center for Enabling Technologies at the SCI Institute. After an initial development phase of about one year the project was released under the MIT license and since then multiple institutions from America, Europe and Asia are contributing to the software. KiwiViewer [Kiwiviewer, 2016] application is a free, open-source visualization app for exploring scientific and medical datasets that runs on Android and iOS mobile devices with multi-touch interaction.

ToolKits available also in the mobile platform unfortunately, there is only one open source toolkit available called VES. VES [VES, 2016] is the VTK OpenGL ES Rendering Toolkit. It is a C ++ rendering library for mobile devices using OpenGL ES 2.0. VES integrates with the Visualization Toolkit (VTK) to deliver scientific and medical visualization capabilities to mobile application developers. Recently in 2015 VES become deprecated and combined with VTK.

PROPOSED METHOD

Ray casting, a standout amongst the most utilized volume rendering algorithm was initially presented by Levoy. In comparison to the other techniques, raycasting is widely accepted as the best quality volume rendering technique [Christian John Noon, 2012]. Additionally, it supports optimizations such as early ray termination and space leaping. It gives consequences of high quality, normally considered to give the best image quality. In this technique, a ray is produced for each desired picture pixel. Ray casting is a normal image order technique. Since there is no surfaces in direct volume rendering step through the volume must done carefully. The principle is as follows: a ray is shot out from a specific point in the screen, and the ray passes through the volume data, then equidistant sampling is obtained along the ray. The value at every sampling point is repeated accumulated according to its color value and opacity. This composited voxel can be presented the every sampling point in this ray. Eventually, the synthesis value of sampling in every ray show the map image. Figure 3 illustrates the pseudo code used for implement ray cast. Ray casting can produce high-quality 2D images from 3D volume data but the method is computationally demanding, especially when multiple volumes are involved, so a parallel GPU version has been implemented. Which is implemented in this paper.

The implementation is done using visualization toolKit. VTK has long evolved beyond just visualization. VTK consists of a C++ class library and several interpreted interface for visualization technique. Figure 4 shows the processes used to implement ray casting technique on iOS platform using VTK. The main class used from VTK to implement ray casting is vtkGPUVolumeRayCastMapper, which is a volume mapper that performs ray casting on the GPU. Ray casting is inherently an image-order rendering technique, with one or more rays cast through the volume per image pixel. VTK is inherently an object-

order rendering system, where all graphical primitive (points, lines, triangles, etc.) represented by the vtkProps in the scene are rendered by the GPU in one or more passes.

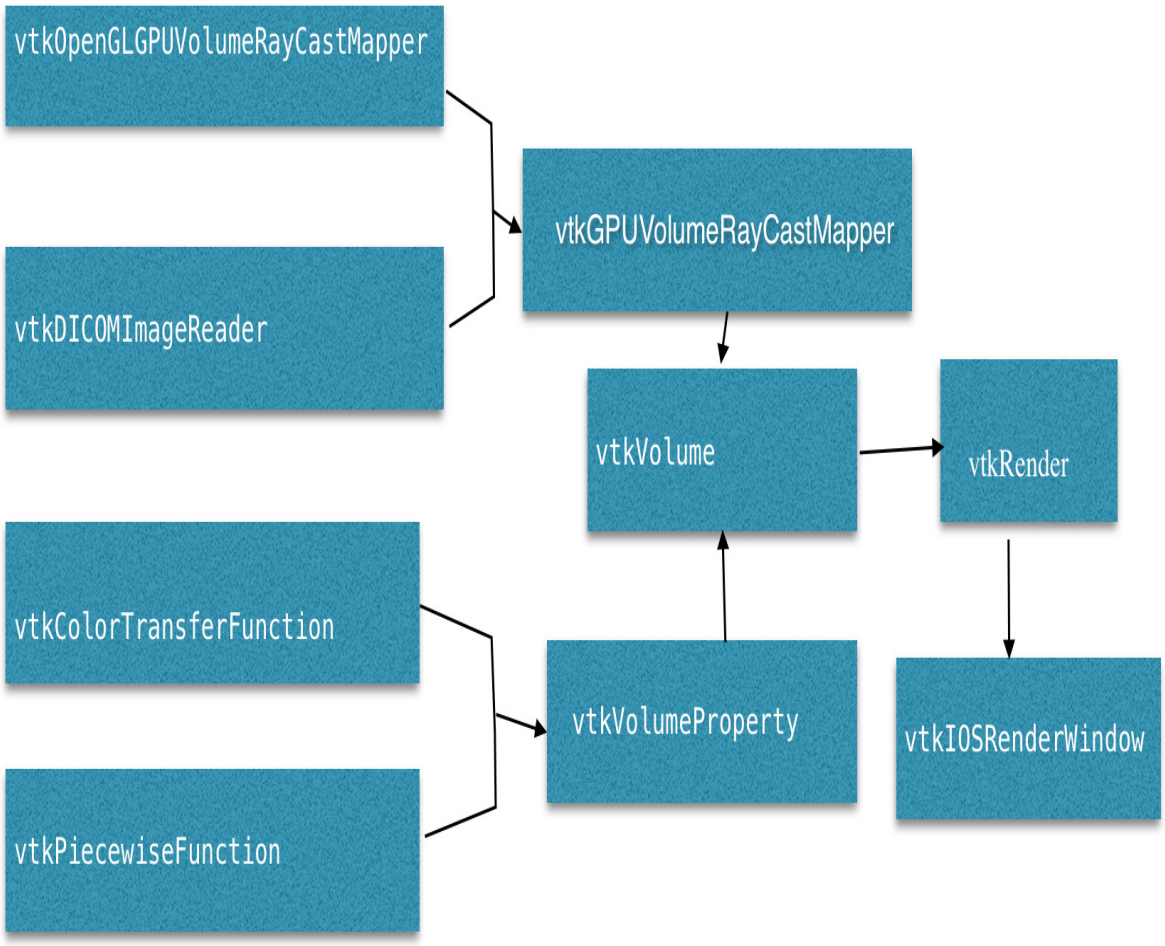


Figure 3: Volume rendering process using VTK

The proposed mobile based application architecture described in Figure 4. Here's the basic idea the user can load the DICOM image of the bone on the mobile, the application accepts images of CT data type. The application reads DICOM images and then apply RayCast volume visualization technique. The 3D object result of RayCast technique shows on mobile screen. The application is designed to be on-device rendering, as its read DICOM image, and implement the volume visualization technique on mobile without need any server side.

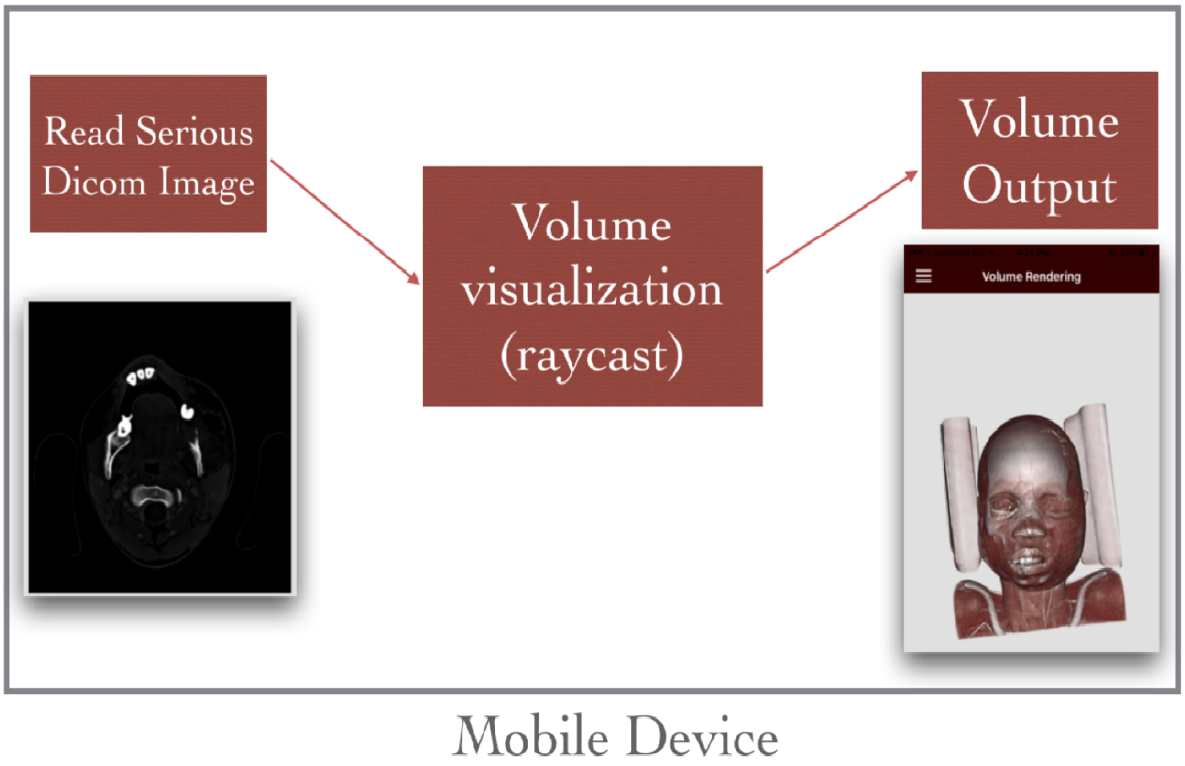


Figure 4: Proposed Mobile Based Application Architecture

EXPERIMENTAL RESULTS

A. Experimental Environment

In our experiment we select iPhone 6s as a test platform that has PowerVR GT7600 GPU which render into 750 x 1334 frame buffers for display at so-called retina resolutions and Apple A9 Dual-core 1.84 GHz Processor. The device runs on iOS 9. The software was developed as a native iOS application.

B. DataSet

The data set used in this experimental is raw images obtained from Osirix [11] datasets . The raw images are CT images. File size is 120 MB compressed. Raw images are from surgical repair of facial deformity.

C. Results

Figure 5show the result obtained from the both application (a) represent the 3D object obtained from ImageVis3D and (b) represent proposed application. It is shown obviously the high visualized result gained by implementing GPU ray cast technique over sliced based volume rendering technique used in ImageVis3D. The surgical repair of facial deformity shown clearly in the implemented system over ImageVis3D.

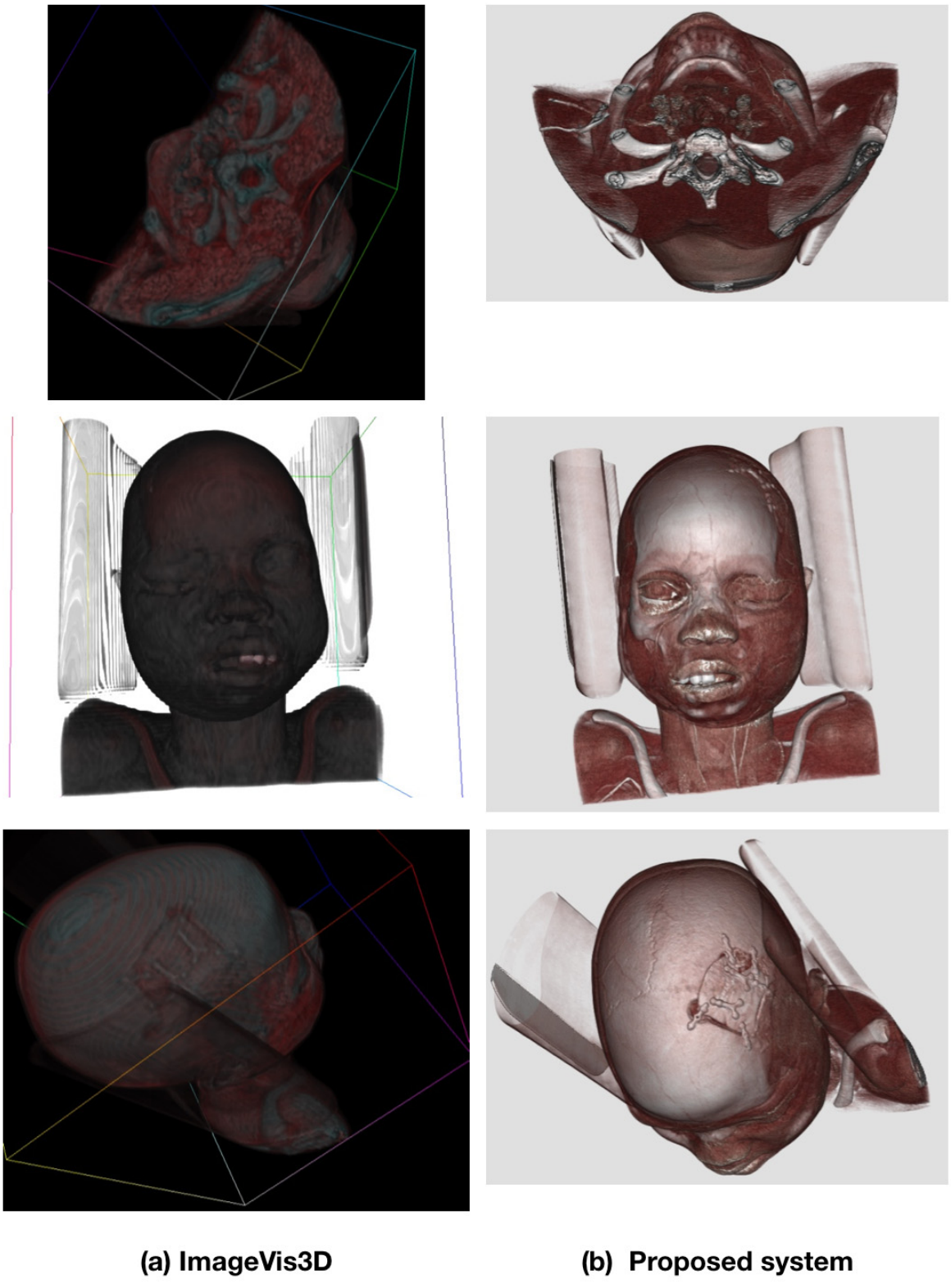


Figure 5: The results from ;(a) ImageVis3D application and (b) proposed application

D. Limitation

This system has been developed based on GPU ray casting technique. The system reads CT images from iPhone and apply the technique then the 3D object result rendering on the screen.

This system has the following recognition limitations :

- Runs only on iPhone and iPad devices
- Accepts only CT images
- Tests on small datasets “120Mb”
- Doesn't support large datasets Gigs

CONCLUSION AND FUTURE WORK

Now it's time for mobile, mobile phones gained high performance due to the rapid development in their hardware which live up to the laptop's hardware with the advantage of their mobility. This development made mobile phones be available for the heavy implementation of medical visualization techniques. Despite these developments, the stores of mobile suffer from the rare existence of medical volume visualization application. In this paper mobile application is represented to help in the orthopedic field. The implemented application shown high visualization result when compared to ImageVis3D.

For the future development there are several areas of focus. Mainly will focus on overcome the limitation of the mobile application for visualization medical image such as support large datasets ,GIGS, and support different datasets techniques like MRI.

The other focus will be virtual reality for 3d medical objects on mobile phones. As currently and with the development of virtual reality glasses for mobile phones and their cheap price, it become possible to develop mobile application that support virtual reality.

Acknowledgement

"The paper is published with partial support by the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua)".

Bibliography

[Christian John Noon, 2012] Christian John Noon. A volume rendering engine for desktops, laptops, mobile devices and immersive virtual reality systems using gpu-based volume raycasting. In: A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of doctor of Philosophy, 2012.

[Dataset, 2016] VES. http://www.osirix-viewer.com/datasets/.

[GFXBench, 2016] GFXBench. https://gfxbench.com/

[ImageVis3D, 2016] ImageVis3D. https://en.wikipedia.org/wiki/Image-Vis3D_Mobile.

[J.M.Noguera _et_al, 2012] J.M.Noguera, J.R.Jiménez, C.J.Ogáyar and R.J.Segura. Volume rendering strategies on mobile devices. In: International Conference on Computer Graphics Theory and Applications, 2012.

[Kiwiviewer, 2016] Kiwiviewer. http://www.kiwiviewer.org/.

[Lamberti and Sanna, 2007] F. Lamberti and A. Sanna. A streaming-based solution for remote visualization of 3D graphics on mobile devices. In: IEEE Transactions on Visualization and Computer Graphics, 1949.

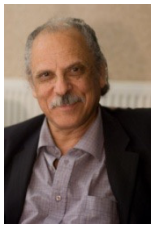
[M.Moser and D.Weiskopf, 2006] M.Moser and D.Weiskopf. Interactive Volume Rendering on Mobile Devices. In: VMV, 2006.

[Mobeen _et_al, 2012] Mobeen, M.M., Lin Feng. Ubiquitous Medical Volume Rendering on Mobile Devices (i-Society). In: International Conference, 2012.

[N. Tolia _et_al, 2006] N. Tolia, D. Andersen, and M. Satyanarayanan. Quantifying inter- active user experience on thin clients. In: Computer, 2006.

[VES, 2016] VES. http://www.vtk.org/Wiki/VES.

Authors' Information



Abdel-Badeeh M.Salem -*Professor of Computer Science and Head of Artificial Intelligence and Knowledge Engineering Research Labs at the department of Computer science , Faculty of Computer and Information sciences , Ain Shams University , Cairo, Egypt; email: absalem@cis.asu.edu.eg*

Major Fields of Scientific Research: *Knowledge engineering, Artificial intelligence, Data mining, Biomedical informatics ,Intelligent e-Learning*

Yassmin Abdallah *-Msc Student at at the department of Computer science , Faculty of Computer and Information sciences , Ain Shams University , Cairo, Egypt; e-mail: yassminabdallah@gmail.com*

Major Fields of Scientific Research: *Medical image processing , medical volume visualization techniques*

COMPUTER DEVICES AND INFORMATION TECHNOLOGIES FOR MEDICINE

Volodymyr Romanov, Ozar Mintser, Igor Galelyuka, Viktor Degtjaruk, Viktor Chernetsky, Illya Chaikovsky

Abstract*: Information technologies and computer devices are modern and very useful tools, which are widely and successfully applied in various areas of human activity and social life, including medicine and health care. Adoption of new information technologies and modern devices in medicine causes increasing of quality and efficiency of medical assistance and in such way lets to decrease the number of complications and unfavourable outcomes, to promote socioeconomic progress and to increase the overall quality of life. New information technologies (IT communicator, remote healthcare monitor) and computer devices (ECG device with vessels system evaluation, and digital phonendoscope) for medical application, developed in Glushkov Institute of Cybernetics in cooperation with Shupyk National Medical Academy of Postgraduate Education, are considered. Proposed devices refer to diagnostic and measuring tools. In spite of fast application of new diagnostic technologies and devices, in most cases it is very important and reasonable to have direct contact between doctor and patient or injured person. Proposed information technology, IT communicator, helps to simplify communication between doctors and patients and increase treatment efficiency beginning from the first contact. The main areas of communicator usage are family and emergency medicine, where it is necessary to support the first contact of doctors with patients (with voice limitation) and get information about patient state, for example, during the preliminary examination of patients and diagnosing. Proposed remote healthcare monitors help doctors to monitor state of their patients in remote mode and prevent possible ill health. Application of proposed computer devices and information technologies in medicine give possibility to improve the quality of life of risk group people and giving timely medical aid to health impairment patients.*

Keywords*: medical computer device, remote healthcare monitor.*

ACM Classification Keywords: *J.3 Life and Medical Sciences - Biology and Genetics*

Introduction

Information technologies and computer devices are modern and very useful tools, which are widely and successfully applied in various areas of human activity and social life, including medicine and health care. Adoption of new information technologies and modern devices in medicine causes increasing of quality and efficiency of medical assistance and in such way lets to decrease the number of complications and unfavourable outcomes, to promote socioeconomic progress and to increase the overall quality of life.

Medical computer devices and information technologies are getting wide application in the world. In the article new medical computer devices and information technologies, developed in the Glushkov Institute of Cybernetics of NAS of Ukraine with cooperation with Shupyk National Medical Academy of Postgraduate Education, are presented. The first one is medical IT communicator [Sergienko, 2013]. It is intended for supporting first contact of doctors with patients (with voice limitation), and getting information about patient state, and also for providing emergency medical aid for such patients. The second is ECG device with evaluation of vessels system. The main idea of this computer device is combining advanced software for analysis of electrocardiogram and pulse wave. The device includes updated ECG-photometric unit, advanced software for deep processing and up-to-date interpretation of ECG and pulse wave results, manuals for patients and doctors. The third computer device is used for measuring performances of cardiovascular system and is used as digital phonendoscope. The fourth one is remote healthcare monitor is based on wireless technologies and intended to remote monitoring the health state of people of risk groups, such as elderly people, chronic patients and risky job people.

Work objectives

Work objectives are analyses of possibilities of application of medical computer devices and information technologies for first aid, treatment, and providing emergency relief for risk group people.

IT communicator

Relationships and communication between doctors and patients is a very important problem in medicine. Ability of correct communication during the disease or separate stage of disease helps to diagnose, prognose a clinical behaviour, and accelerate the recovery. Communication problems between a doctor and a patient influence on correctness of the diagnoses and quality of a doctor aid, and complicate the medical treatment. Relations between doctors and patients are not limited with communication. Making a diagnosis, patient state examination, interpretation of the data received from a patient, treatment planning, and treatment evaluation are very important too. Methods of evaluation and correction of diagnosis influence on treatment planning and treatment effectiveness. It is proposed

to apply special hardware-software means, particularly IT communicators to simplify the communication of doctors with patients (particularly with voice limitation) and increase treatment efficiency beginning from the first contact.

First version of the IT communicator, presented in [Mintser, 2014], simplified the communication between doctors and patients with voice limitation and included only the general information about traumas and diseases. The last version of proposed IT communicator additionally has databases with detailed formalized information about a large number of traumas or diseases and proper methods of medical treatment or care. Such information helps both doctors and patients to act correctly during diagnosing, patient examination and medical treatment.

Developed IT communicator is a mobile tablet computer with special application-dependent software and intended for the following purposes:

1. In medicine, including emergency medical aid: for supporting first contact of doctors with patients (with voice limitation), and getting information about patient state. In this case the IT communicators help such patients to communicate with doctors. Particularly it is important for family doctors who often are the firsts to examine patients.

2. In family medicine: to support doctors, especially family doctors, during patient examination and diagnosing. In such case, if the family doctor hasn't enough experience and knowledge to diagnose in correct way, he can use digital databases with detailed formalized information about a large number of traumas or diseases and proper methods of medical treatment or care.

3. In education: during continuous professional development of doctors and pharmacist.

Special application-dependent software consists of two subsystems. The first subsystem is used for alternative communication of doctors with patient, lost temporarily or permanently possibility of speaking, during first contact or examination. The second subsystem gives detailed information about trauma or disease and proper methods of medical treatment or care to doctors or, in some cases, patients. It is needed to note, that now there are many home medical devices in the world market and the second subsystem of IT communicator consists of manuals of many of them to make easy to use the devices.

One of the aims of IT communicator is to support alternative communication with voice and motion limited patients. The idea of alternative communication is rather simple. Voice limited patient needs communication with surroundings in any way. The process of establishing relations with additional means helps patient with voice limitation to express his needs and wishes. Communicators for alternative communication are used to help voice limitation patients to understand surroundings clearly.

The IT communicator makes it possible communication between doctors and patients with move and voice limitation. Fig. 1 shows the main menu of the smart communicator and the window, where patient can locate the place of pain to help doctor. There are following Menu items: “Needs”, “Requests”, “Pain”, “Emergency aid”, and “Keyboard”. By using this menu and submenus the patient with voice limitation can ask peoples around or doctor for help, service or medicaments, or explain own feelings, troubles or pain. It is possible to convert written by patient needs to voice. It is necessary to add that IT communicator stores the patient medical history.

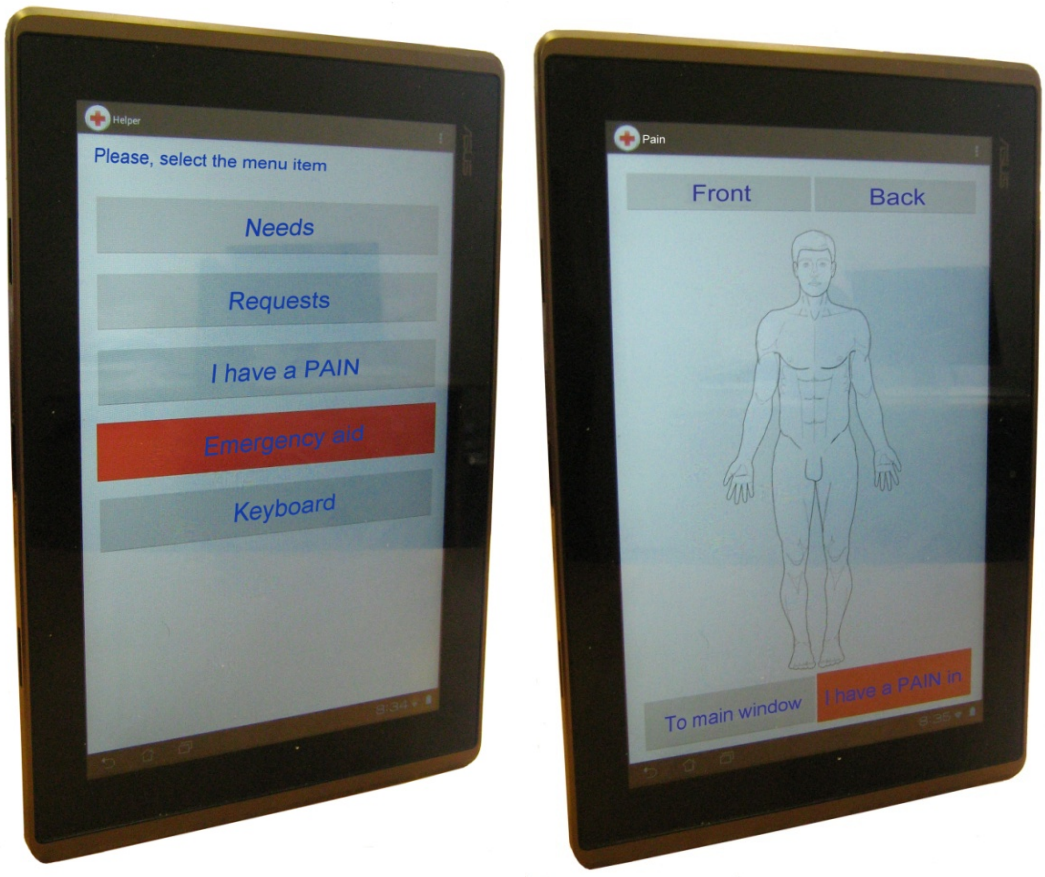


Fig. 1. Main menu and pain location windows

The second subsystem of smart communicator is intended to support family doctors and simplify the communication between doctor and patient. To remove some problem in communication between doctor and patient it is reasonable to computerize the formalized and standard models and situations in communication, what usually lets to avoid the effect of incorrect understanding of certain "message" both data sender and data recipient.

IT communicators are reasonable to be used by family doctors to get detailed information about trauma or disease and proper methods of medical treatment or care during first examination of sick or injured person. It is very urgent, because family doctors meet a broad list of traumas or diseases immediately after their appearance. In such cases family doctors sometimes haven't enough experience and knowledge to diagnose in correct way. Information exchange between doctor and patient already includes some "distortion" of information because of incorrect or insufficient understanding of certain knowledge domain. Proposed communicator removes such problems by standardization of possible situations. For family medicine it is developed more than 500 standardized situations, which include as cases of emergency medical aid, so planed activity of medical care. The window with details of illness or injury, for example the asthma, and window with visual explanation are shown on Fig. 2. In addition, proposed communicator subsystem contains databases with reference information, what helps to avoid a lot of traumas or diseases.

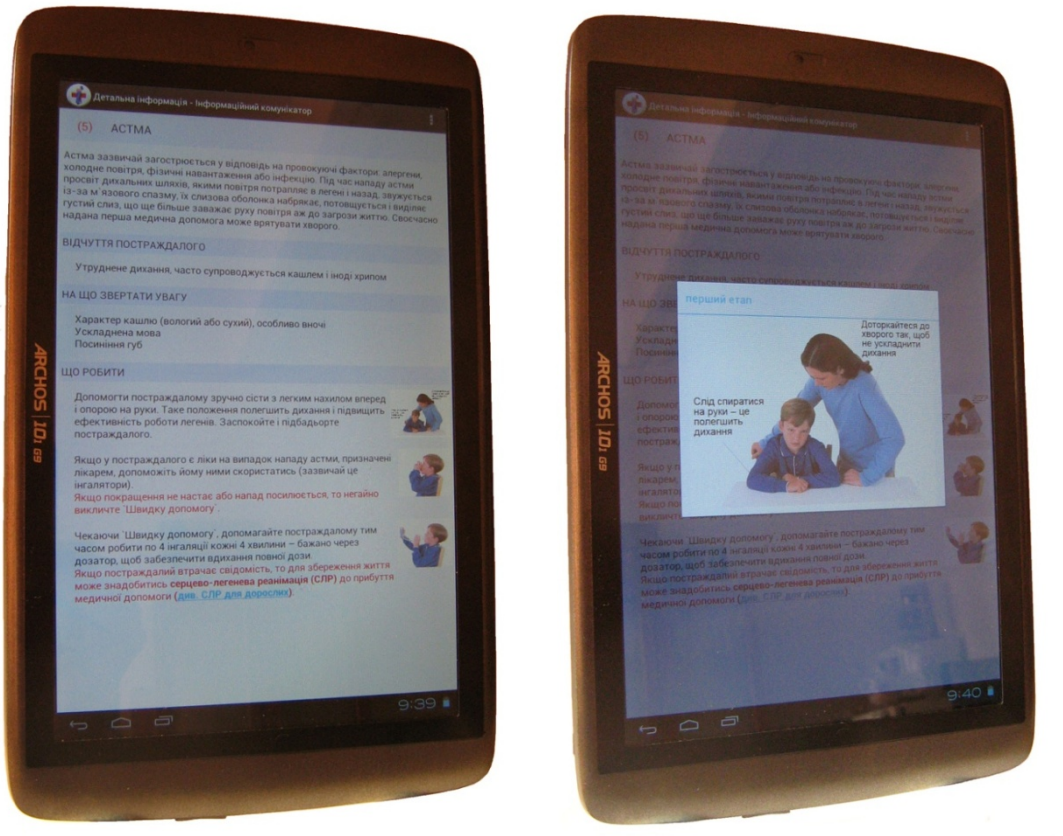


Fig. 2. Detailed information about trauma or disease (with visual explanation)

Very often in distant mode the doctor can't explain the patient how to act with some disease or trauma, because of absence visual materials or knowledge restrictions. To avoid such problems it is reasonable

to give patients access to data bases with detailed information on majority of diseases and traumas, its symptoms, diagnostic methods and first aid actions. In such case the doctor orally explain the patient the actions and direction of operations. The patient by oneself takes actions with using of detailed text explanations, high-resolution images and video materials from databases. In such cases the IT communicator is very useful tool.

It is also proposed to use smart communicators in continuous professional development of doctors and pharmacist.

ECG device with evaluation of vessels system

The main idea of this computer device is combining advanced software for analysis of electrocardiogram and pulse wave. The device includes updated ECG-photometric unit, advanced software for deep processing and up-to-date interpretation of ECG and pulse wave results, manuals for patients and doctors. This device may significantly increase the specificity and sensitivity of resting electrocardiography by using the most modern algorithms for ECG interpretation. Combination of advanced ECG with evaluation of vessels system, including its micro vessel component and endothelial dysfunction using pulse wave analysis, opens new excellent possibilities for early diagnosis of most dangerous heart diseases and for control the efficiency of treatment in clinical condition. Moreover, this devise may be used for home medicine to increase the quality of life of patients with chronic heart diseases and for healthy persons by monitoring their heart pre-conditions. During last three years the pilot series of family of the miniature devices were produced in Ukraine. The device clinical testing was done in several lead hospitals in Ukraine.

Digital phonendoscope

Digital phonendoscope is used for measuring performances of cardiovascular system. Computer device is consists of acoustic pressure sensor, data converter and microcontroller. The acoustic sensor is situated upper the vessel for cyclorama of pressure measuring. The sensor is sensitive to real time changing of pressure in vessel during the cardiac cycle. The main advantage of the device is high sensitivity due to measuring of phase shift between two sinusoidal signals (instead of voltage) caused with changing the blood pressure in vascular walls of the patient. The level of phase shift is equivalent to changing the pressure in vessel during the cardiac cycle. The different phase shifts are converted with ADC to digital codes and then digital data are displayed as time plot of changing the blood pressure in vascular walls of the patient. The device is used as digital phonendoscope.

Fig. 3. General view of portable ECG-photometric diagnostic complex

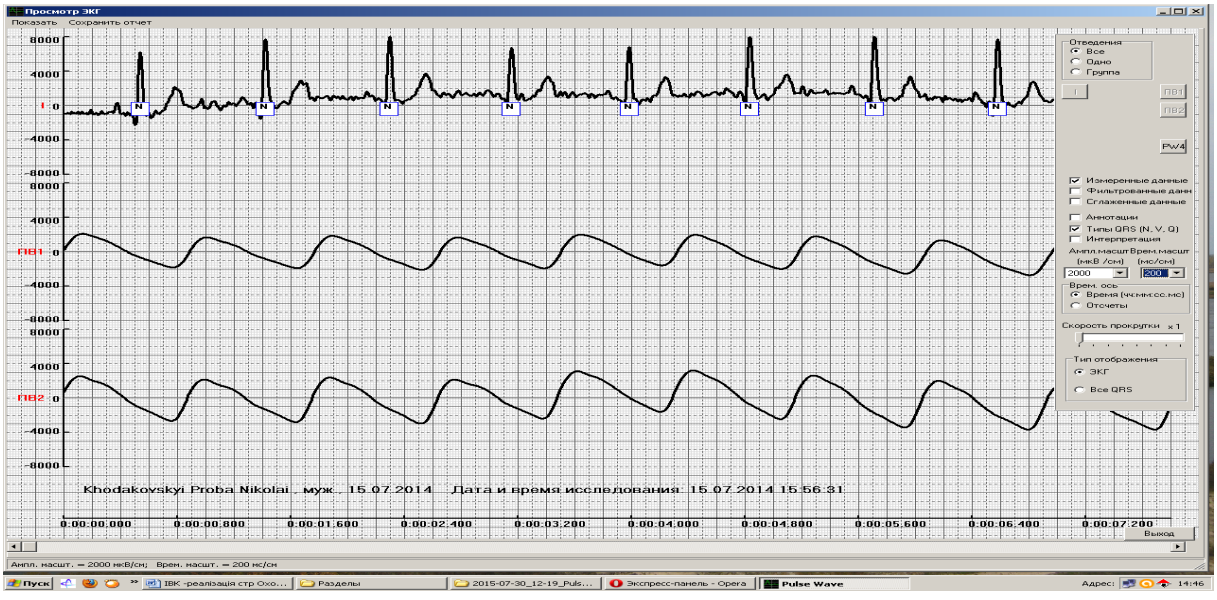


Fig. 4. Synchronized recording of cardiogram and pulse wave

Remote healthcare monitor

Remote healthcare monitor is based on two approaches. The first approach uses wireless sensor networks (WSN) to monitor the state of patients on some territory, for example, in hospital. The second one provides using of smartphones to monitor state of patients away from hospitals, for example, in their houses or on the street.

There are many advantages of WSN in different applications. It gives a lot of benefits to users in science, industry, medicine, environment protection, agriculture and so on. WSNs include from tens to hundreds of nodes equipped with micropower battery supplies. State of the art of WSN applications in medicine and healthcare is given in [Ko, 2010].

For designing pilot elements of WSN it was chosen the wireless microcontroller JN5168, manufactured by NXP Company. Microcontroller includes 32-bit processor with 1–32 MHz clock speed, 2,4 GHz IEEE802.15.4 compliant transceiver, 4-input 10-bit ADC and a comprehensive mix of analogue and digital peripherals etc. This microcontroller complies with the network nodes requirements. It supports data acquisition, data storage and data transferring via wireless channel.

The standardization and unification requirements to WSN are determined by 802.15.4 standard, which defines features of creation of networks with low data throughput. In additional ZigBee and JenNet-IP protocols determine requirements to network routing and security. We selected JenNet-IP protocol for our applied tasks. The JenNet-IP protocol combines IEEE802.15.4-based wireless network technology and the Internet Protocol (IP) to achieve integration between the two domains, supporting the wireless "Internet of Things". Due to the nodes of a wireless network are to be controlled both wirelessly and from IP-connected device, such as a smart remotely located phone.

The network node includes a set of medical sensors, placed on patient, and wireless microcontroller JN5168, which gathers data from medical sensors and transfers them to central hospital network coordinator.

On the initial development stages for proposed tasks it was proposed to use simple tree-like topology of WSN. The part of WSN topology, which was created as base fragment (cluster) of network, is shown on Fig. 5.

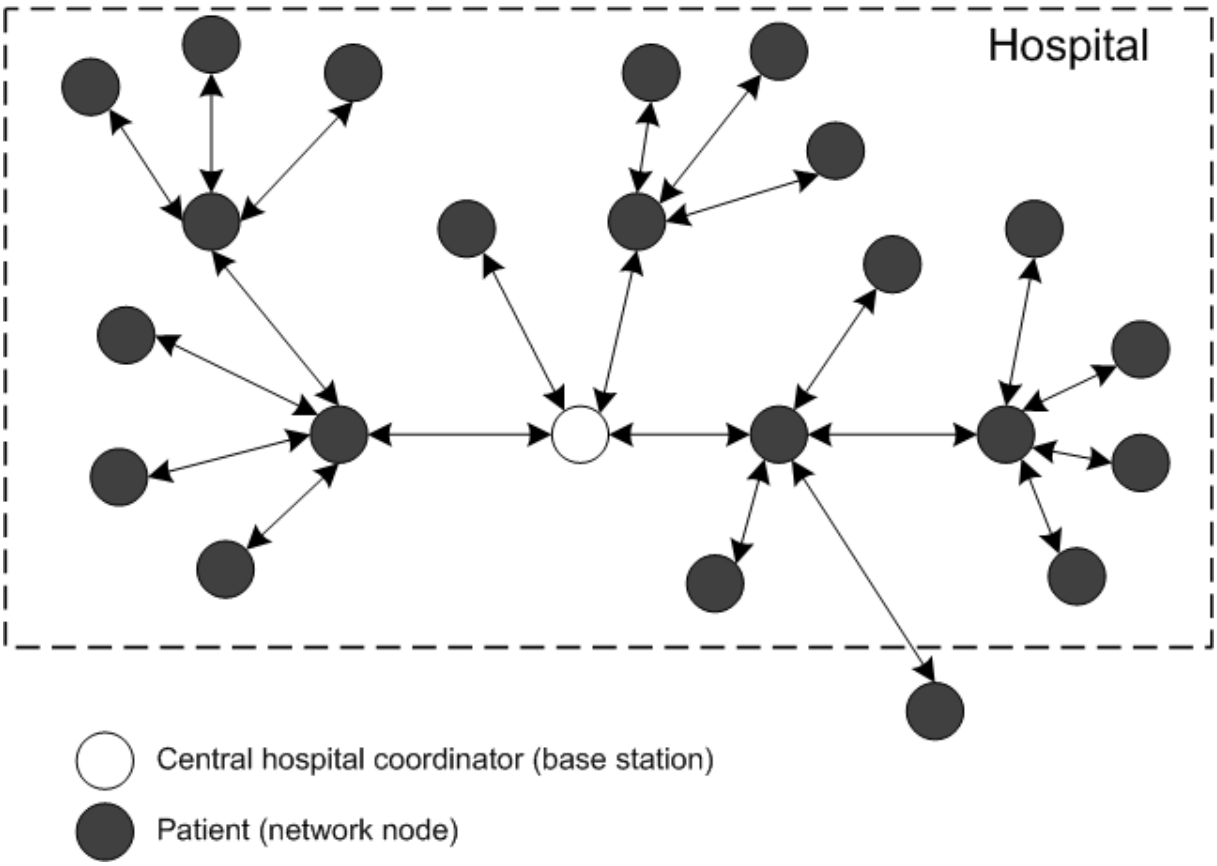


Fig. 5. Structure of remote healthcare monitor (with WSN)

The second approach provides using of smartphones to monitor state of patients away from hospitals, for example, in their houses or on the street. To implement this approach it was decided to develop concentrating unit, which acquires data from medical sensors, placed on the patient body. The set of sensors is defined by risk group, to which the patient belongs: elderly persons, people with cardiac diseases, diabetics, and etc.

For designing the concentrating unit it was used microcontroller LPC4357 with ARM Cortex-M4 (with ARM Cortex-M0 coprocessor) core. For wireless communication between concentrating unit and medical sensors it was used tiny Bluetooth units. In additional for processing data packets from medical sensors the standards of ISO/IEEE 11073 family were taken into consideration. Data packets with information about parameter of patient state are transferred from medical sensor via Bluetooth channel to concentrating unit. Concentrating unit gathers data from all medical sensors, pre-processes, archives and passes this information via smartphone and cloud environment to hospital server. If some patient parameter exceeds determined limits the alert message is formed and quickly transferred to hospital doctor. Then doctor makes a decision and, if it is necessary, calls patient and asks him to carry out some actions to prevent ill health.

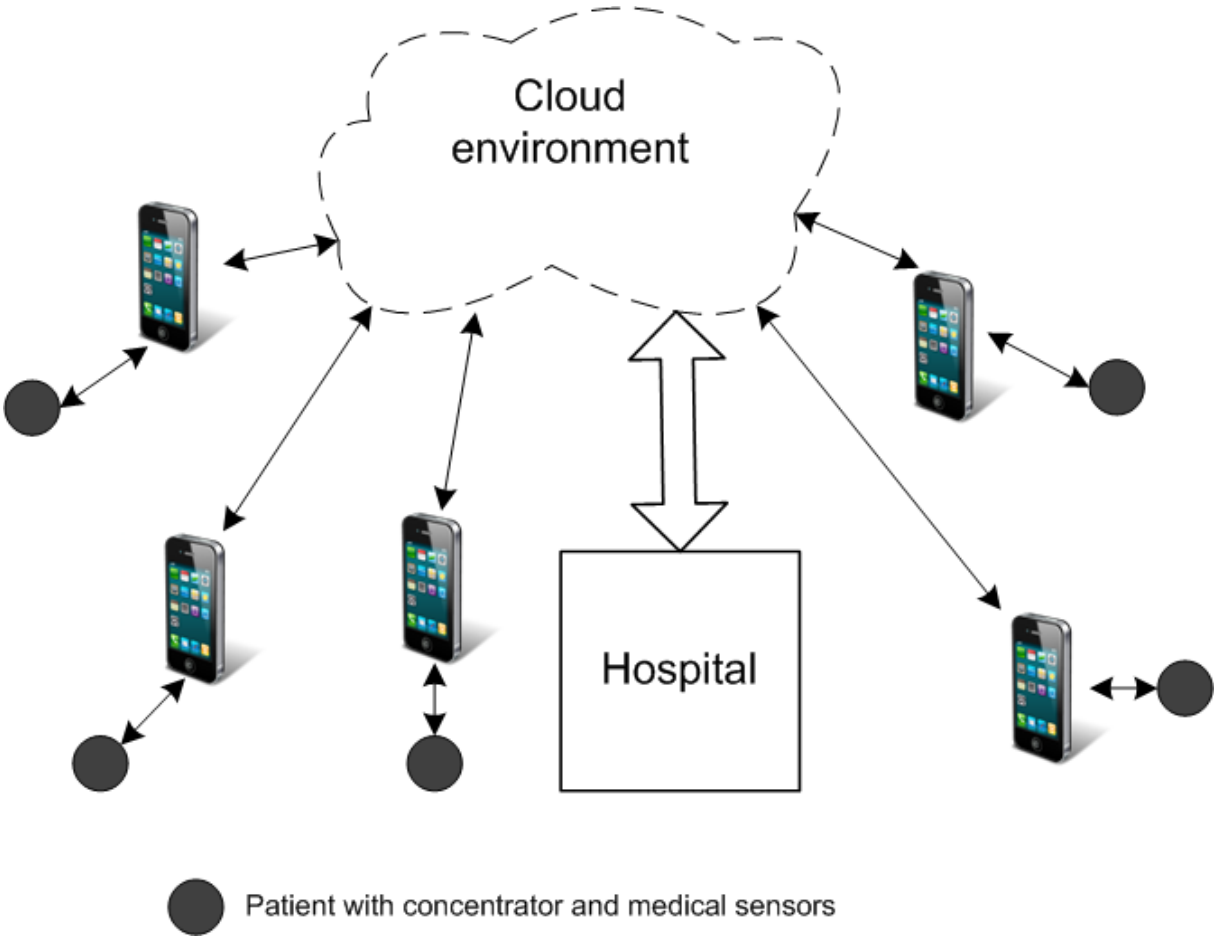


Fig. 6. Structure of remote healthcare monitor (with smartphones)

Application of individual monitoring of risky group patients, for example, elderly people state can reduces mortality, optimize using of medical service, and earlier detect physical impairment. Registration of physical impairment of such remote patients is very important information to make emergency service.

The remote healthcare monitor of any of two approaches consists of following wearable medical sensors (Fig. 7): blood pressure sensor, body temperature sensor, pulse rate sensor, oximeter, EEG and ECG sensors, accelerometer, and GPS unit.

Up to date the wireless units were developed and embedded in medical sensors. Now doctors are developing methods, which allow estimating the state of remote patients using parameters of several medical sensors.

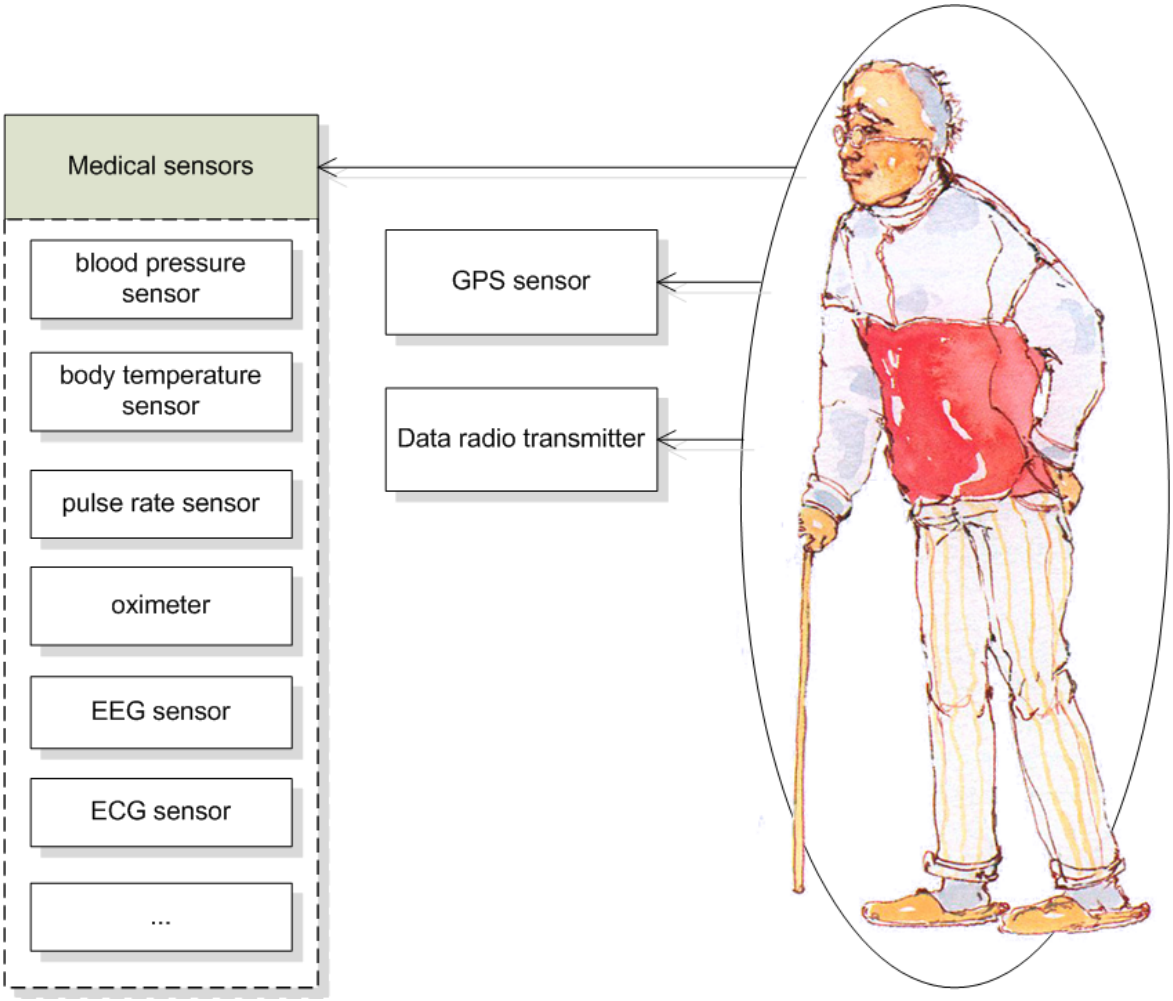


Fig. 7. Sensors of remote healthcare monitor

Conclusion

Development and creation of IT communicator, ECG device with vessels system evaluation, digital phonendoscope, and remote healthcare monitor for monitoring parameters of human state give possibility to improve the quality of life of risk group people and giving timely medical aid to health impairment patients.

Bibliography

[Ko, 2010] Jeonggil Ko, Chenyang Lu, Mani B. Srivastava, John A. Stankovic, Andreas Terzis, Matt Welsh. "Wireless Sensor Networks for Healthcare", in Proceedings of the IEEE. Vol. 98, No. 11, November 2010, pp. 1947–1960.

[Mintser, 2014] O. Mintser, V. Romanov, I. Galelyuka, O. Voronenko "IT-communicators for medicine", in International Journal "Information Models and Analyses", Vol. 3, Number 2, Sofia, Bulgaria, 2014, pp. 169–174.

[Sergienko, 2013] Patent of Ukraine №80490 “Method of communication with person with voice and auditory limitation” / Sergienko I.V., Voronenko O.V., Galelyuka I.B., Romanov V.O., 2013.

Acknowledgement

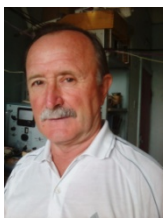
The paper is published with partial support by the project ITHEA XXI of the ITHEA ISS (www.ithea.org) and the ADUIS (www.aduis.com.ua).

Authors' Information

Volodymyr Romanov *– Head of department of V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine, Doctor of engineering sciences, professor; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: VRomanov@i.ua; website: http://www.dasd.com.ua*

Ozar Mintser *– Head of department of P.L. Shupyk National Medical Academy of Postgraduate Education, Doctor of technical sciences, professor;.Dorogozhytska str., 9, Kiev, 04112, Ukraine; e-mail: ozar7@yandex.ru*

Igor Galelyuka *– Senior research fellow of V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine; Candidate of engineering sciences; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine; e-mail: galelyuka@gmail.com; website: http://www.dasd.com.ua*



Viktor Degtjaruk *– Research fellow of V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine*

Viktor Chernetsky *– Senior research fellow of V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine; Candidate of engineering sciences; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine*

Illya Chaikovsky *– Senior research fellow of V.M.Glushkov Institute of Cybernetics of National Academy of Sciences of Ukraine; Candidate of medical sciences; Prospect Akademika Glushkova 40, Kiev–187, 03680, Ukraine*

IMPROVING PERFORMANCE IN NEURAL NETWORKS USING FEATURE SELECTION BASED ON RANDOM FORESTS: APPLICATION TO AUTOMATED MEDICAL DIAGNOSIS

Smaranda Belciug, Florin Gorunescu

Abstract*: Neural networks are well-known for their effectiveness as classifiers. In this context, feature selection has become a usefulness technique to make the classifier faster, cost-effective, and more accurate. The effect of a random forests-based feature selection on the classification accuracy of a multi-layer perceptron has been explored in this paper. Up to 6% improvement in classification accuracy and 40% improvement in computation speed has been observed when using the tandem model on five real-world publically available medical datasets regarding colon cancer, breast cancer, diabetes, thyroid and fetal heartbeat databases.*

Keywords*: multi-layer perceptron, feature selection, random forests, automated medical diagnosis*

ACM Classification Keywords*: F.1 Computation by abstract devices – Self-modifying machines, H.2.8 Database Applications – Data mining*

Introduction

The use of machine learning (ML) tools in processing medical data represents nowadays an important trend [Yang, 2010]. Intelligent medical decision systems, built using state-of-the-art ML algorithms, are particularly conceived for automatic diagnosis, and this approach became a *fruitful niche* within the vast Bioinformatics field.

Within the most powerful ML algorithms one can mention, in particular, the neural networks (NNs). NNs represent "massively parallel distributed processors" [Haykin, 1999], and are used to model arbitrary non-linear data. The computational metaphor behind them is their unique property of mimicking the human brain knowledge-processing paradigm. The most used NN architecture is represented by the multi-layer perceptron (MLP), which stores knowledge in its synaptic weights. The back-propagation (BP) algorithm [Bishop, 2004] is the most popular learning algorithm for MLP. It is based on the minimization of the sum-of-squares error (SSE) using the gradient descent technique.

NNs have been successfully used in automated medical diagnosis [Amato et al., 2013], [Belciug and Gorunescu, 2014], breast cancer detection and recurrence [Kalteh et al., 2013], [Belciug and

Gorunescu, 2013], cardiovascular diseases [Nakajima et al., 2015], diabetes [Nguyen et al., 2014], colon cancer [Spelt et al., 2013], Alzheimer's disease [Coppedè et al., 2013], etc.

Different approaches to improve the NNs performance have been proposed. Various methods underlie them, such as nearest neighbor [Alejo et al., 2006], clustering [Madasamy and Tamilselvi, 2013], hybridization with genetic algorithms [Belciug and Gorunescu, 2013], Bayesian learning [Belciug and Gorunescu, 2014], statistical tools (discriminant function analysis, regression analysis, correlation) along with NN sensitivity analysis [Gorunescu et al., 2012], swarm optimization [Dheeba and Selvi, 2012], etc.

An effective way to improve the NN classification accuracy, regardless of the use of different algorithms to strengthen their performance, resides in using the feature selection (FS). The direct use of medical databases without previous analysis and filtering is often counterproductive, even when using efficient algorithms such as MLP. FS represents an efficient way to solve this issue by choosing the most relevant attributes from data. There are different approaches regarding FS [Liu and Motoda, 2007]. One can mention association rules [Karabatak and Ince, 2009], particle swarm optimization [Shen et al., 2009], genetic algorithms [Marinakis et al., 2009], hill climbing [Stoean et al., 2011], statistical tools [Gorunescu et al., 2012], etc.

Different from other approaches trying to improve the MLP classification accuracy, the current work proposes a FS-based technique obtained using Random Forests (RFs). RFs are among the most popular ML algorithms due to their relatively good accuracy and robustness [Breiman, 2001]. They are used for classification and regression, and it is noteworthy that they provide an efficient method for FS. In this paper, RFs are applied in the beginning on the whole dataset using the method of mean decrease impurity. The feature (predictor) importance is then computed, and all the features are ranked accordingly. Next, the features with the rank higher than a default threshold are chosen as input variables (independent predictors) for MLP. In this way, the most important features will decide the class (label) for a certain item. The statistical comparison indicates that the novel tandem FS-MLP outperforms the conventional MLP algorithm regarding both the decision accuracy and the computation speed.

The remainder of this paper is organized in four sections. Section 2 briefly presents the MLP, FS, and RF techniques, focusing on the detailed presentation of the tandem RF-MLP model, along with the five publically available datasets. Section 3 presents the experimental results in terms of performance analysis and performance assessment, while Section 4 deals with the conclusions and future work.

Methodology

This section briefly outline the principles of MLP, usefulness of using FS along with MLP, and the RF-based FS procedure used in this paper. A detailed description of the novel tandem FS-MLP model is provided at the end of the section.

Multi-layer perceptron

The key paradigm underlying NNs is the information processing system consisting of a large number of highly interconnected processing units called *neurons*, organized in a layered parallel structure called *network*. The *learning* process is typically achieved through progressive adjustments of synaptic weights in order to attain a desired design objective. A neuron consists of a set of inputs x_i, weighted by the corresponding synaptic weights w_i, added together and passed onto an activation function which bounds the acceptable amplitude range of the neuron output to some finite value. The MLP model, which is the most commonly used NN architecture, is based on computational units, which compute a non-linear function of the scalar product of the input vector and the synaptic weight vector. The MLP architecture involves some critical hyper-parameters in its design, such as the network depth, the number of hidden neurons per layer, the initial learning rate, and the momentum. A key observation in the practical use of MLP, due to a theorem by Kolmogorov, is that MLP with only one hidden layers, i.e. a *three-layer perceptron* (3-MLP), is theoretically sufficient to model almost any real-life problem [Bishop, 2004], [Haykin, 1999].

There is a two-fold information process consisting of the *training phase* and the *testing phase*. In the training phase, a training dataset is used to determine the weight parameters w_i that define the neural model. Next, the trained neural model is used to process (unknown) test patterns, yielding the true classification performance. To summarize, MLP is firstly trained to associate outputs with input patterns, and, secondly, it provides the output that corresponds to a taught input pattern that is least different from the given pattern.

Feature selection

Based on some filtering criteria, such as statistical tools, ML algorithms, etc., FS techniques choose the most relevant attributes from the original dataset and remove unimportant or redundant features. FS may be considered a key factor for pattern recognition, since even the best classifiers will perform poorly if the features are not chosen appropriately.

The feature selection methods may be broadly categorized as filters and wrappers [Yuan et al., 1999]. Thus, a filter method selects features based only on the intrinsic characteristic of the data, while a wrapper method uses a search model as part of evaluation of the relevance of a feature. Since RFs are a popular method for feature ranking, we took advantage of this ability to select good features, and

used the Data mining module (Random Forests for Regression and Classification) within Statistica 7 package, StatSoft. Inc. to choose the most appropriate features for the classification process.

Random Forests

RFs are “a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest” [Breiman, 2001]. Each node in the tree predictors is a condition on a single feature, splitting the dataset into two subsets, such that similar response values reach eventually the same set. The usual impurity measures used for classification are either Gini impurity or information gain/entropy. For classification problems, the tree response takes the form of a class membership, which associates a set of independent (predictor) variables with one of the categories present in the dependent variable.

In the tree training process, it can be computed how much each feature decreases the weighted impurity. The feature (predictor) importance is computed by summing the drop in node impurity over all nodes in the tree(s), and expressing these sums relative to the largest sum found over all predictors. It is noteworthy that this is different from the notion of predictor or variable importance [Breiman et al., 1984].

In the FS process, we have considered the *predictor importance* representing the importance of each feature in the classification process, so that we can distinguish the features that make the major contributions to the prediction of the decision class. The predictor importance is ranking on a 0-100 scale basis for each predictor variable in the analysis.

Tandem RF-MLP model

The novel decision model consisting of the tandem RF-MLP is based on the following algorithm.

Algorithm RF-MLP

1. Run the RF algorithm on the dataset using the Statistica 7 module “Random Forests for Regression and Classification” with the following settings:

- Equal misclassification costs;

- Estimated prior probabilities (the likelihood that a case will fall into one of the classes is proportional to the dependent variable class sizes);

- Hold out cross-validation (2/3 was considered as training set, and 1/3 as testing set);

2. Rank features on a 0-100 scale through the “Predictor importance” RF feature, with the default threshold 50.

3. Keep the most influential features (i.e., rank higher than 50) for the subsequent MLP use.

4. Apply the MLP classifier to the selected features as network inputs.

Remark. It is worth noting that the issue of establishing the optimal threshold, keeping the most influential attributes, remains an open issue, depending on the concrete decision problem to solve. The number of MLP hidden units has been chosen heuristically for each dataset, in order to obtain optimal classification performance at the same time with the network simplicity. Concretely, we have 8 for CC, 10 for BCWD, 7 for PID, 13 for THY, and 15 for CARD.

Datasets

Five different real-world datasets have been used to evaluate the usefulness of the proposed approach. They are publically available, enabling the evaluation of the novel methodology and the comparison against other state-of-the-art techniques. Their sizes range from 368 to 7200, the number of input variables, both numerical and categorical, ranges from 7 to 21, and the number of output variables ranges from 2 to 10. Table 1 provides the details of each dataset.

Table 1. Datasets

Dataset	No. of instances	Input variables	Output variables (classes)
Colon Cancer (CC)	368	8 (numerical, categorical)	3 (short stay, medium stay, long stay)
Breast Cancer Wisconsin Diagnostic (BCWD)	569	10 (numerical)	2 (benign, malign)
Pima Indian Diabetes (PID)	768	7 (numerical, categorical)	2 (negative, positive)
Thyroid (THY)	7200	21 (numerical, categorical)	3 (degree: 1, 2, and 3)
Cardiotocography (CARD)	2126	21 (numerical, categorical)	10 (calm sleep, REM sleep, calm vigilance, active vigilance, shift pattern, accelerative/decelerative pattern, largely decelerative pattern, flat-sinusoidal pattern, suspect pattern)

The datasets originate from the University of Medicine and Pharmacy of Craiova, Romania, and from the UCI Machine Learning Repository. Details of the envisaged datasets: https://sites.google.com/site/imediatreat/data-sets, http://archive.ics.uci.edu/ml/datasets.html

Remark. The above datasets are not balanced. It is well-known that the success of ML classifiers is quite limited when they are applied on imbalanced datasets. To overcome this situation, we have considered the prior probabilities of each decision class, reflecting the degree of belief in class memberships of instances (i.e., whether an instance (input variable) belongs to a certain class (output variable)). In this regard, we have considered the prior probability as the percentage of each class (category) in the dataset, as a reasonable approach.

Results

Experimental results on real-world medical datasets are presented in this section. They are directly compared with results obtained by using six well-known state-of-the-art ML algorithms on the same datasets, in order to highlight the effectiveness of the proposed model.

Experimental results

The performance of the proposed tandem RF-MLP model has been assessed by using two important statistical measures: the average (testing) accuracy (ACC), and the corresponding standard deviation (SD) obtained by (independently) running the model 100 times in a complete hold out cross-validation cycle. The purpose of using these two statistical parameters was to highlight both the level of classification performance achieved by the model (ACC), and the 'stability' of the classification accuracy obtained in multiple independent computer runs (SD), because the proposed algorithm is of stochastic nature and ACC differs from a computer run to another. Thus, smaller SD indicates evidence of a more stable model with respect to different computer runs. For binary classification, we have also considered the area under the ROC curve (AUC).

Taking advantage of the RF ability to select the most important features for the classification process, we have considered for each dataset the features with the rank higher than a default threshold equaling 50 to a 0-100 scale, although any other choice being possible (user-defined). The experimental results regarding RF-MLP are displayed in Table 2 in terms of ACC and SD, along with the corresponding selected features and the running time (RF-MLP vs. 3-MLP) obtained in 100 complete hold out cross-validation cycles.

Table 2. Experimental results

Datasets	CC	BCWD	PID	THY	CARD
No. of selected features	5 out of 8	6 out of 10	6 out of 8	2 out of 21	12 out of 21
Performance measures (%)	ACC/SD	ACC/SD	ACC/SD	ACC/SD	ACC/SD
	71.17/3.72	91.14/1.54	75.06/3.81	95.69/0.30	84.08/0.85
Running time RF-MLP/3-MLP	2'04"/2'14"	2'57"/3'12"	3'42"/3'47"	16'28"/27'19"	9'47"/7'12"

As expected, the classification performance (i.e., testing accuracy) strongly depends on each dataset, ranging from 71% (CC dataset) to 96% (THY dataset). On the other hand, the 'stability' of the algorithm, measured through SD, also depends on each dataset, ranging from 0.30% (THY dataset) to 3.81%

(PID dataset). It is noteworthy that the best classification accuracy corresponds to the greatest stability, obtained in the case of the THY dataset (95.69%, 0.30%), although this is the largest dataset (7200 instances). Surprisingly in this case is that FS kept only 2 attributes (i.e., thyroid stimulating hormone, and free thyroxin index) out of 21. The use only of the two attributes provided about 96% classification accuracy, proving once again the significant role of FS in classification issues.

To synthetically illustrate the 'stability' of the model during different computer runs, a "box-and-whisker" plot displaying the ACC variability over the 100 independent computer runs is depicted in Fig. 1.

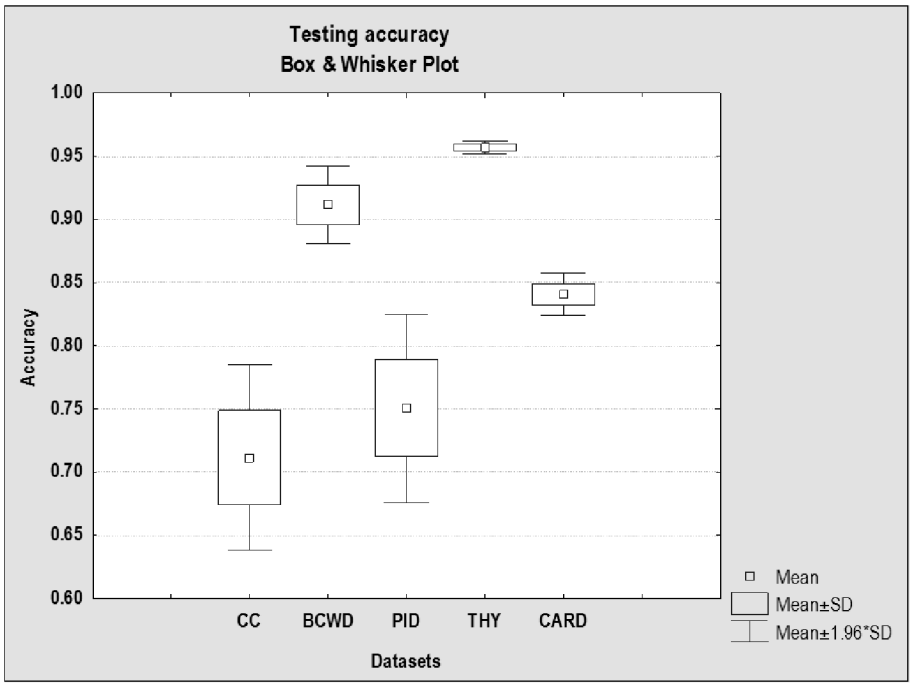


Fig. 1. Box and whisker plot (performance variability illustration)

The "box-and-whisker" plot summarizes the central tendency (central line), the variability around the central tendency (box), and the range of the variable (whiskers around the box). Low SD values (narrow box and close whiskers) mean high model 'stability, in other words, the classification accuracy does not depend significantly on the computer run. Thus, the best 'stability' along with the highest accuracy have been obtained for THY dataset, while the lowest performance has been obtained for CC and PID datasets.

Benchmark: RF-MLP performance evaluation

Table 3 presents the experimental results, in terms of ACC and corresponding AUC, obtained by each ML algorithm during 100 independent computer runs in complete hold out cross-validation cycle.

Table 3. Experimental results: Acc (%), AUC

Model		Classification performance ACC/AUC)			
	CC	BCWD	PID	THY	CARD
	Acc -	Acc AUC	Acc AUC	Acc -	Acc -
RF-MLP	71.17 -	91.14 0.987	75.06 0.859	95.69 -	84.08 -
3-MLP	68.34 -	90.39 0.984	70.73 0.816	95.67 -	81.74 -
RBF	63.57 -	87.03 0.921	71.44 0.819	93.75 -	75.10
PNN	62.63 -	70.92 0.769	65.75 0.796	93.12 -	72.46
SVM	58.69	94.55	76.56	92.58	72.93
n-B	63.04	92.30	75.52	95.27	57.33
k-NN	63.14	93.78	68.75	94.57	75.37

From Table 3 one can see that:

- Overall, the performance of all classifiers strongly depends on the specific dataset used, as expected;

- The use of FS through RF improved the accuracy of the conventional MLP;

- As compared with other state-of-the-art ML algorithms, the tandem RF-MLP model provided a better performance.

In conjunction with Table 3, the visual comparison of the classification performance of the five ML competitors, illustrated in Fig. 2., synthetically reveals the difference between them also depending on dataset.

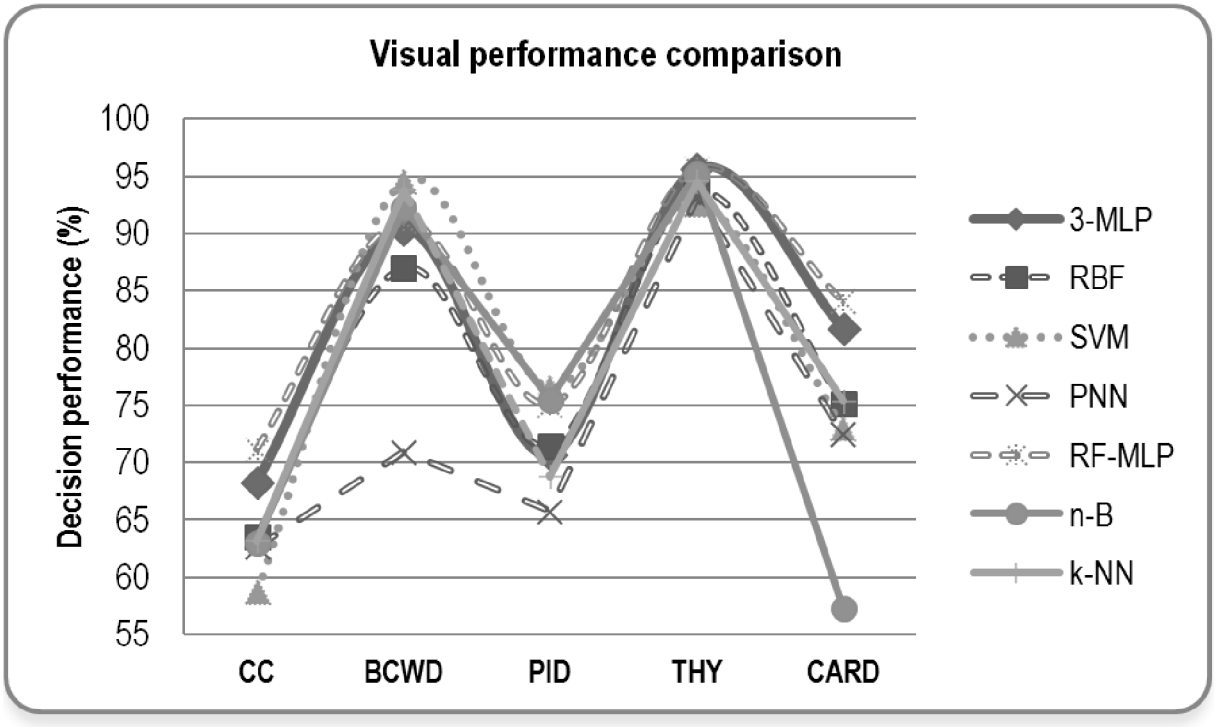


Fig. 2. Visual comparison of the ML algorithms performance

Next, the one-way ANOVA technique along with the Tukey's honestly significant difference (Tukey HSD) *post-hoc* test have been used to statistically address the difference in classification accuracy of the ML algorithms involved in the benchmarking process. The one-way ANOVA along with the Tukey HSD test were performed using the IBM SPSS 21.0 package. The ANOVA output, consisting of (combined) sums of squares (SS), degrees of freedom (df), mean squares (MS), F-value, and *p*-level (contrasts: quadratic polynomial), is presented in Table 4.

Table 4. One-way ANOVA comparison for mean testing accuracy

Dataset	SS	df	MS	F-value	*p*-level
CC	4866.08	3	1622.03	53.63	0.00
BCWD	26895.24	3	8965.08	115.87	0.00
PID	4397.19	3	1465.73	132.97	0.00
THY	522.52	3	174.17	35.19	0.00
CARD	8939.06	3	2979.69	55.99	0.00

The *post-hoc* Tukey HSD test has revealed statistically significant differences in classification performance (p-level < 0.05) regarding the following ML algorithms/datasets:

- RF-MLP vs. 3-MLP (mean diff. = 2.83, std. err. = 0.78), RF-MLP vs.RBF (mean diff. = 7.59, std. err. = 0.78),
RF-MLP vs. PNN (mean diff. = 8.54, std. err. = 0.78) on CC dataset;

- RF-MLP vs.RBF (mean diff. = 4.11, std. err. = 1.24), RF-MLP vs. PNN (mean diff. = 20.22, std. err. = 1.24) on BCWD dataset;

- RF-MLP vs. 3-MLP (mean diff. = 4.32, std. err. = 0.47), RF-MLP vs.RBF (mean diff. = 3.61, std. err. = 0.47),
RF-MLP vs. PNN (mean diff. = 9.30, std. err. = 0.47) on PID dataset;

- RF-MLP vs.RBF (mean diff. = 1.93, std. err. = 0.31), RF-MLP vs. PNN (mean diff. = 2.57, std. err. = 0.31) on THY dataset;

- RF-MLP vs.RBF (mean diff. = 8.97, std. err. = 1.03), RF-MLP vs. PNN (mean diff. = 11.61, std. err. = 1.03) on CARD dataset.

The comparison regarding classification performance between RF-MLP and the non-neural models (SVM, n-B, *k*-NN) has been assessed using the follow-up two-sided *z*-test. Statistically significant differences have been disclosed on CARD dataset, i.e., RF-MLP vs. SVM (p-level = 0.05), and RF-MLP vs. *k*-NN (p-level = 0.0001).

To complete the benchmarking process, we have presented in Table 5, the comparison between the novel tandem model RF-MLP and other well-established ML classifiers, i.e., optimal discriminant plane (ODP), regularized discriminant analysis (RDA), hierarchical pyramid neural network (HPNN), Cox regression (CR), partially connected neural network (PCNN), hybrid MLP/genetic algorithm (MLP/GA), decision trees (DT), logit (LOG), and random forest (RF), with results reported in literature [Belciug and Gorunescu, 2013], [Stoean et al. 2015], [Gorunescu and Belciug, 2014], [Aeberhard et al., 1994], [Wilson and Martinez, 1997], [Kinney, 1988], [Sahin and Subasi, 2015].

Remark. It should be noted that a direct comparison with RF-MLP is not very suitable since in these studies the ML algorithms taken into considerations were not applied to all the datasets, and the results have not been obtained using the same experimental settings. Let us also notice more or less important differences between the performances obtained by simulation using Statistica 7 package and displayed in Table 3, and those reported in literature, obtained by ML algorithms specially designed by authors.

Table 5. BPSS-MLP performance compared to other ML algorithms

Model		Dataset-Acc (%)			
	CC	BCWD	PID	THY	CARD
RF-MLP	71.17	91.14	75.06	95.69	84.08
RBF	-	87.42	70.83	-	-
PNN	-	71.08	65.62	-	-
MLP	71.2	81.39	74.47	92.85	97.78
k-NN	-	94.12	71.90	-	98.40
ODP	-	-	-	-	-
RDA	-	-	-	-	82.10
HPNN	-	-	-	-	-
CR	-	-	-	-	-
SVM	73	96.92	76.30	-	-
PCNN	-	81.07	-	-	-
MLP/GA	-	93.58	-	-	-
DT	72.32	-	-	-	86.36
LOG	60	-	-	-	-
RF	-	-	-	-	99.18

Conclusion

In the automated medical diagnosis, hybridizing in a collaborative manner state-of-the-art ML algorithms and latest medical approaches has become a very important interdisciplinary technology. The effectiveness of improving the performance of NNs, based on FS in tandem with the traditional MLP, was investigated on the task of providing a reliable intelligent decision support system for the automated diagnosis of various diseases. The performance of the novel model equaled or exceeded the results reported in literature on publicly available colon cancer, breast cancer, diabetes, thyroid, and cardiotocography.

Future research has to investigate at least the use of FS provided by RFs in tandem with other state-of-the-art ML algorithms, or hybrid algorithms.

Acknowledgements

The authors acknowledge the support of the research grant no. 26/2014, code PN-II-PT-PCCA-2013-4-1153, entitled IMEDIATREAT - Intelligent Medical Information System for the Diagnosis and Monitoring of the Treatment of Patients with Colorectal Neoplasm - financed by the Romanian Ministry of National Education (MEN) - Research and the Executive Agency for Higher Education Research Development and Innovation Funding (UEFISCDI). The CC database was part of this support. The BCWD database was obtained from the University of Wisconsin Hospitals, Madison. Thanks go to Dr. William H. Wolberg, W. Nick Street and Olvi L. Mangasarian for providing the data. For the PID database thanks go to Vincent Sigillito, John Hopkins University, the donor of the data. For the THY database thanks go to the Garavan Institute, Sydney, Australia, and R. Quinlan (documentation). For the CARD database thanks go to Marques de Sa, J. Bernades, and A. de Campos, Portugal.

Bibliography

[Aeberhard et al., 1994] S.Aeberhard, D.Coomans, O.de Vel. The performance of statistical pattern recognition methods in high dimensional settings. IEEE Signal Processing Workshop on Higher Order Statistics. Ceasarea. 1994;4:14-16.

[Alejo et al., 2006] R.Alejo, V.Garcia, J.M.Sotoca, R.A.Mollineda, J.S.Sanchez. **Improving the classification accuracy of RBF and MLP neural networks trained with imbalanced samples. Proc.** Proceedings of the 7th international conference on Intelligent Data Engineering and Automated Learning -IDEAL'06, 464-471, Springer, 2006.

[Amato et al., 2013] F.Amato, A.Lopez, E.M.Pena-Mendez, et al. Artificial neural networks in medical diagnosis. J Appl Biomed 2013;11:47–58.

[Belciug and Gorunescu, 2013] S.Belciug, F.Gorunescu. A hybrid neural network/genetic algorithm system applied to the breast cancer detection and recurrence. Expert Syst 2013;30(3):243-254.

[Belciug and Gorunescu, 2014] Belciug S, Gorunescu F. Error-correction learning for artificial neural networks using the Bayesian paradigm. Application to automated medical diagnosis. J Biomed Inform. 2014 Dec;52:329-337.

[Bishop, 2004] C.Bishop. Neural Networks for Pattern Recognition. Oxford University Press, 2004.

[Breiman et al., 1984] L.Breiman, J.H.Friedman, R.A.Olshen, C.J.Stone. Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.

[Breiman, 2001] L.Breiman. Random Forests. Machine Learning. 2001;45(1):5-32.

[Coppedè et al., 2013] F.Coppedè, E.Grossi, M.Buscema, L.Migliore. Application of artificial neural networks to investigate one-carbon metabolism in Alzheimer's disease and healthy matched individuals. PLoS One. 2013; 8(8): e74012.

[Dheeba and Selvi, 2012] J.Dheeba, S.T.Selvi. A swarm optimized neural network system for classification of microcalcification in mammograms. J Med Syst 2012;36(5):3051-3061.

[Gorunescu et al., 2012] F.Gorunescu., S.Belciug, M.Gorunescu, R.Badea. Intelligent decision-making for liver fibrosis stadialization based on tandem feature selection and evolutionary-driven neural network. Expert Syst Appl. 2012;39: 12824-12833.

[Gorunescu and Belciug, 2014] F.Gorunescu, S.Belciug. Evolutionary strategy to develop learning-based decision systems. Application to breast cancer and liver fibrosis stadialization. J Biomed Inform 2014;49:112-118.

[Haykin, 1999] S.Haykin. Neural Networks. A comprehensive foundation (2nd Ed.). Prentice Hall International. 1999.

[Kalteh et al., 2013] A.A.Kalteh, P.Zarbakhsh, M.Jirabadi, J.Addeh. A research about breast cancer detection using different neural networks and K-MICA algorithm. J Cancer Res Ther 2013;9(3):456-466.

[Karabatak and Ince, 2009] M.Karabatak, M.C.Ince. An expert system for detection of breast cancer based on association rules and neural network. Expert Syst Appl. *2009;*36:3465-3469.

[Kinney, 1998] E.Kinney. Cox Regression application communication Report with TR-10-88. Harvard University, Center for Research in computing Technology, Aiken Computation Laboratory, 1988.

[Liu and Motoda, 2007] H.Liu, H.Motoda. (Eds.), In: Computational methods of feature selection. Ed. H.Liu and H.Motoda. Chapman and Hall/CRC, 2007.

[Madasamy and Tamilselvi, 2013] B.Madasamy, J.Tamilselvi. Improving classification Accuracy of Neural Network through Clustering Algorithms. Intl J of Computer Trends and Technology. 2013;4(9):3242-3246.

[Marinakis et al., 2009] Y.Marinakis, G.Dounias, L.Jantzen. Pap smear diagnosis using a hybrid intelligent scheme focusing on genetic algorithm based feature selection and nearest neighbor classification. *Comput. Biol. Med*. 2009;39 (1):69–78.

[Nakajima et al., 2015] K.Nakajima, S.Matsuo, H.Wakabayashi et al. Diagnostic Performance of Artificial Neural Network for Detecting Ischemia in Myocardial Perfusion Imaging. Circ J. 2015;79(7):1549-56.

[Nguyen et al., 2014] L.L.Nguyen, S.Su, H.T.Nguyen. Neural network approach for non-invasive detection of hyperglycemia using electrocardiographic signals. Proc. IEEE Conf Eng Med Biol Soc. 2014:4475-4478.

[Sahin and Subasi, 2015] H.Sahin, A.Subasi. Classification of the cardiotocogram data for anticipation of fetal risks using machine learning techniques. J. Applied Soft Computing 2015;33(C):231-238.

[Shen et al., 2009] Q.Shen, Z.Meia, B-X.Ye. Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification. Comput. Biol. Med. 2009;39(7):646–649.

[Spelt et al., 2013] L.Spelt, J.Nilsson, R.Andersson, B.Andersson. Artificial neural networks - a method for prediction of survival following liver resection for colorectal cancer metastases. Eur J Surg Oncol. 2013;39(6):648-54.

[Stoean et al., 2011] C.Stoean, R.Stoean, M.Lupsor, H.Stefanescu, R.Badea. Feature selection for a cooperative coevolutionary classifier in liver fibrosis diagnosis. Comput. Biol Med. 2011;41:238-246.

[Stoean et al., 2015] R.Stoean, C.Stoean, A.Sandita et al. *Ensemble of Classifiers for Length of Stay Prediction in Colorectal Cancer.* In: Intl. Work-Conference on Artificial Neural Networks (IWANN 2015), Advances in Computational Intelligence, Lecture Notes in Computer Science 2015;9094:444-457.

[Wilson and Martinez, 1997] D.R.Wilson, T.R.Martinez. Improved Center Point Selection for Probabilistic Neural Networks. In: Proc. Intl. Conference on Artificial Neural Networks and Genetic Algorithms (ICANNGA'97). 1997:514-517.

[Yang, 2010] Z.R.Yang. Machine learning approaches to Bioinformatics. World Scientific, Science, Engineering, and Biology Informatics, 2010.

[Yuan et al., 1999] H.Yuan, S-S.Tseng, W.Gangshan, Z.Fuyan. A two-phase feature selection method using both filter and wrapper, Proc. IEEE Conference on Systems, Man, and Cybernetics. 1999;2:132-136.

Authors' Information

Smaranda Belciug *–Lecturer, PhD, Department of Computer Science, Faculty of Sciences, University of Craiova, Craiova, Romania, e-mail: sbelciug@inf.ucv.ro; H-index (ISI Web of Knowledge = 6, SCOPUS = 6, Google Scholar = 9)*

Major Fields of Scientific Research: Data Mining, Artificial Intelligence, Machine Learning, Statistics

Florin Gorunescu *– Professor of Biostatistics and Computer Science, University of Medicine and Pharmacy of Craiova, Craiova Romania; e-mail: gorunef@gmail.com; H-index (ISI Web of Knowledge = 12, SCOPUS = 14, Google Scholar = 18)*

Major Fields of Scientific Research: Data Mining, Artificial Intelligence, Machine Learning, Statistics, Operational Research, Mathematics

COMPARATIVE ANALYSIS OF VARIOUS FILTERS FOR NOISE REDUCTION IN MRI ABDOMINAL IMAGES

Antoniya Mihailova, Veska Georgieva

Abstract*: Magnetic resonance imaging MRI is an imaging technique that is primarily used in medical diagnostics for the visualization of the structures and functions of tissues and organs in the body. It physically based on the principles of nuclear magnetic resonance (NMR), and in particular the field gradient NMR, and is therefore also known as magnetic resonance imaging (sometimes colloquially shortened to MRI). Compared to CT occur artifacts (distorted image) by the MRI diagnostic are more frequently and interfere usually more with the image quality. A wide variety of artifacts is routinely encountered on MR images. In this paper, various filtering algorithms are discussed and compared.*

Keywords*: MRI, noise, artifacts, filtration methods*

ACM Classification Keywords*:*

Introduction

The medical visual diagnostic is very important and widely used in the analyses of the human health and diseases. Nowadays it is almost unthinkable for doctors to make diagnose without the help of some kind of medical visualization technique. However it is still imposable for them to count on this visualization to be on hundred present accurate. The main reason of that is the unavoidable presence of noise and a variety of artifacts in the medical images. This justifies our research and observation of the filtration's methods of one of the widely used medical imaging - MRI.

In fact it has a various advantageous features, such as high-resolution capability, the ability to produce an arbitrary anatomic cross-sectional imaging, and high tissue contrast. Unfortunately, there are many potential sources of image artifacts associated with the technology of MRI. Many MR artifacts are neither obvious nor understandable from previous experience with conventional types of imaging. While some MR artifacts are machine specific, the majority are inherent in the imaging method itself. Many artifacts can be considered as noise.

The term noise in MR can have different meanings depending on the context. For example, it has been applied to degradation sources such as physiological and respiratory distortions in some MR applications and acquisitions schemes [Fern´andez and Vega, Kruger and Glover,2001], [Petridou and

all,2009]. Even acoustic sources (the sound produced by the pulse sequences in the magnet) are sometimes referred to as noise [Fern´andez and Vega, Counter and all, 2000]. The presence of noise over the acquired MR signal is a problem that affects not only the visual quality of the images, but also may interfere with further processing techniques such as segmentation, registration or fMRI analysis [Fern´andez and Vega, McGibney and Smith, 1993, Gudbjartsson and all, 1995], [Aja-Fern'andez and all, 2008].

In the signal processing literature, many of the popular denoising algorithms suggested are based on wavelet thresholding [Sarode, Deshmukh, 2010-11]. These approaches attempt to separate significant features/signals from noise in the frequency domain and simultaneously preserve them while removing noise [Sarode, Deshmukh, 2010-11], [Chang and all, 2006]. If the wavelet transform is applied on MR magnitude data directly, both the wavelet and the scaling coefficients of a noisy MRI image are biased estimates of their noise-free counterparts [Sarode, Deshmukh, 2010-11],[Chang and all, 2006]. The difficulty with wavelet or anisotropic diffusion algorithms is again the risk of over- smoothing fine details, particularly in low SNR images [Sarode, Deshmukh, 2010-11, Bernsteinan all, 1989].

In this paper, various filtering algorithms are used to remove noise from MRI images as good as it is possible and to preserve the quality of them. These filtering algorithms have various advantages and disadvantages. There are different filters and none of them overcome others in all situations in respect to computation cost, noise removing and quality of denoised image. That is why noise removal method can be improved and it can be still an open research area.

MRI Imaging and Artifacts

Magnetic resonance imaging (MRI) is primarily used in medicine (radiology) to visualize detailed internal structure and limited function of the body. By MRI contrast between the different soft tissues of the body is much greater than those by computed tomography (CT). This makes it especially useful in neurological (brain), musculoskeletal, cardiovascular, and ontological (cancer) imaging. In MRI there is no ionizing radiation, but uses a powerful magnetic field to align the nuclear magnetization of (usually) hydrogen atoms in water in the body. Every MRI scanner has a powerful radio transmitter to generate the electromagnetic field which excites the spins. If the body absorbs the energy, heating occurs. For this reason, the transmitter rate at which energy is absorbed by the body has to be limited. It has been claimed that tattoos made with iron containing dyes can lead to burns on the subject's body.

It works as a radio frequency transmitter is briefly turned on, producing an electromagnetic field. In simple terms, the photons of this field have just the right energy, known as the resonance frequency, to flip the spin of the aligned protons. As the intensity and duration of the field increases, more aligned spins are affected. After the field is turned off, the protons decay to the original spin-down state and the

difference in energy between the two states is released as a photon. It is these photons that produce the signal which can be detected by the scanner. An image can be constructed because the protons in different tissues return to their equilibrium state at different rates. Contrast agents may be injected intravenously to enhance the appearance of blood vessels, tumours or inflammation. Contrast agents may also be directly injected into a joint in the case of arthrograms, MRI images of joints.

MRI diagnostic is considered as generally very safe procedure. Nonetheless the strong magnetic fields and radio pulses can affect metal implants, including cochlear implants and cardiac pacemakers. In the case of cardiac pacemakers, the results can sometimes be lethal. MRI is used to image every part of the body, and is particularly useful for tissues with many hydrogen nuclei and little density contrast, such as the brain, muscle, connective tissue and most tumours. In clinical practice, MRI is used to distinguish pathologic tissue (such as a brain tumour) from normal tissue. One advantage of an MRI scan is that it is believed to be harmless to the patient. It uses strong magnetic fields and non-ionizing radiation in the radio frequency range. It can be used also during pregnancy. However, as a precaution, current guidelines recommend that pregnant women undergo MRI only when essential. MRI is rapidly growing in importance as a way of diagnosing and monitoring congenital defects of the fetus because it can provide more diagnostic information than ultrasound and it lacks the ionizing radiation of CT.

The artifacts in MR imaging can be grouped into two general categories. First, there are artifacts that are hardware related. These artifacts are relatively uncommon—fortunately, because they are often difficult to diagnose and usually require service personnel to correct. The second category consists of artifacts related to the patient or under operator control. This category is encountered much more commonly and may often be easily prevented or corrected once they are recognized.[Ruan]

In the literature there are a large amount of artifacts' groups, as follow: Motion Artifacts, Susceptibility Artifacts, Chemical Shift Artifacts, Wrap Around Artifacts, Partial Volume Artifacts, Gibbs Ringing Artifacts, Zebra Stripes, Slice-overlap Artifacts, RF Overflow Artifacts, Entry Slice Phenomenon, Zipper Artifacts, Cross-Excitation and Shading [Mirowitz, 1999]. In this study some of them are shown in the experimental part.

Motion is the most prevalent source of MR imaging artifacts. As the name implies, motion artifacts are caused by motion of the imaged object or a part of the imaged object during the imaging sequence.[Ruan, 2011] There are different reasons of motion artifacts and they can be grouped in: Respiratory motion, Cardiac motion and Vascular pulsation.

- Respiratory motion results in ghosting artifacts and blurring that can obscure or simulate lesions. A variety of methods have been used to reduce the effect of respiratory motion artifacts. Mechanical methods, such as use of an abdominal or thoracic binder or taking images with the patient in a prone position, are intended to restrict the amplitude of respiratory motion.

- Cardiac motion produces a series of ghost artifacts along the phase-encoding direction of the image, in addition to blurring and signal loss of cardiac and juxtacardiac structures [Ruan, Huber and all, 2001].
- Vascular pulsation artifacts are recognized by their alignment with the responsible vessel along the phase-encoding direction of the image. These artifacts reproduce the cross-sectional size and shape of the responsible vessel, but not necessarily its signal intensity.

Three different parameters—spin density ρ, spin-lattice relaxation *T*1, and spin–spin relaxation *T*2—determine the resonance signal. *T*1 and *T*2 time constants cannot be measured directly because signal strength is always

influenced by proton density and because field inhomogeneity hide the *T*2 effect. *T*2-enhanced images can be generated by the *spin echo* sequence. Hence, different sequences can be developed for enhancing either of the parameters a more detailed treatment can be found. It changes the appearance of different tissues in images (e.g., water and fat is bright in *T*2 images and tissue is darker while the opposite is true for a *T*1 image) [Toennies, 2012].

The single echo will be taken as the image. The time between the 90° impulse and the echo impulse is called *echo time TE*. The time between two measurements is called *repetition time TR*. Short *TE* (20 msec) and long *TR* (2000 msec) will produce a proton-density-weighted image. Using a shorter repetition time (*TR* = 300–600 msec) will produce a *T*1-weighted image because *T*1 relaxation is generally longer than 200–600 msec. A long *TE (*> 60 msec) and a long *TR* (2000 msec) produces a *T*2-weighted image.

Noise in MR images

In many cases the complex effect of the influence of some different artifacts can be presented as kind of noise. In Magnetic Resonance Images, raw data is intrinsically complex valued and corrupted with zero mean Gaussian distributed noise with equal variance.

A. Gaussian white Noise

Gaussian noise is statistical noise and it has a probability density function of the normal distribution (also known as Gaussian distribution). The probability density function *p* of a Gaussian random variable *z* is given by:

$$p_G(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} \quad (1)$$

where *z* represents the grey level, μ the mean value and σ the standard deviation.

It is widely used as additive white noise to yield additive white Gaussian noise (AWGN). Gaussian noise is properly defined as the noise with a Gaussian amplitude distribution. There is no information of the correlation of the noise in time or of the spectral density of the noise. Gaussian noise is labeled as 'white' because of the correlation of the noise.

After Fourier transformation, the real and imaginary images are still Gaussian distributed given the orthogonality and linearity of the Fourier transform. MR magnitude images are formed by taking the square-root of the sum of the square of the two independent Gaussian random variables (real and imaginary images) pixel by pixel. So the MR magnitude data can be shown to be Rician distributed. When SNR is high (SNR>2), the Rician distribution approaches a Gaussian; when SNR approaches 0, the Rician distribution becomes Rayleigh distributed.

B. Rician Noise

The image intensity in magnetic resonance magnitude images in the presence of noise is to be governed by a Rician distribution [Sarode, Deshmukh, 2010-11].

Rician noise is not additive noise, but is instead data-dependent [Coup´e and all, 2010], [Getreuer and all, 2011]. Consider a set of random numbers, which we take to be the intensity values of a noise-free MR image A defined on a discrete grid M so that $A = \{ai, i \in M\}$. Let σ be the standard deviation of Gaussian noise. There are two sets of Gaussian distributed random numbers $X = \{xi, i \in M\}$ and $Y = \{yi, i \in M\}$ with zero mean and identical standard deviations .Then the following are Rician distributed. [Sarode, Deshmukh, 2010-11]

$$mi = \sqrt{(ai + xi)^2 + y} \quad (2)$$

"Rician noise" depends on the data itself, it is not additive, so to "add" Rician noise to data, what we really mean is make the data Rician distributed [Sarode, Deshmukh, 2010-11].

However, we can still use the two basic models of noise for the MRI images, too:

- additive;
- multiplicative

By the additive model the function that describes the noise doesn't correlate the function of the image:

$$g(x,y) = f(x,y) + \eta(x,y) \quad (3)$$

where *g(x,y)* is the real output image, *f(x,y)* is the ideal not noised image and *η(x,y)* is the noise.
The multiplicative model can be presented with:

$$g(x,y) = f(x,y) \times \eta(x,y) \quad (4)$$

Denoising

Denoising can be described as a process of removing noise from a signal, but in fact is more complicated. The methods of noise reduction are conceptually very similar regardless of the signal being processed. It is interesting that knowledge of the characteristics of an expected signal can mean the implementations of these techniques vary greatly depending on the type of signal [Tisdall and Atkins, 2005]. It is important, when we have a model for the degradation process, to be possible the inverse process to be applied to the image to restore it back to the original form. Denoising is a very important part of preprocessing in medical imaging where the physical requirements for 2 high quality imaging are needed for analyzing images of unique events, in regard to obtain better quality of medical images and more precise diagnostic of disease. Noise can be random or white noise with no coherence or coherent noise introduced by the devices mechanism or processing algorithm [Sarode and Deshmukh, 2010]. In magnetic tape, the larger the grains of the magnetic particles, the more prone the medium is to noise[Sarode and Deshmukh, 2010].

One of the most direct approaches to cope with acquisition noise in MRI (of course, not the only one) is signal estimation via noise removal. Traditionally, noise filtering techniques in different fields have been based on a well-defined prior statistical model of data, usually a Gaussian model. [Aja-Fern'andez and all, 2008]

There are different methods for filtration, which can be applied for noise reduction in MR images. For evaluation of noise reduction some parameters such as: peak signal-to noise ratio (PSNR), the effectiveness of filtration (Eff), which is equal to the difference of signal-to noise ratio in filtered image (SNRF) and signal-to noise ratio in noised image (SNRN), and noise reduction ratio (NRR) are used.

The visualization of result from filtration is another criterion, which can give visually information of the effectiveness of the applied method. This criterion is very important for the doctors. For evaluation of the methods is needed to obtain in the same time high values of PSNR and Eff and low value of NRR. These conditions can be implemented in the algorithm for calculation of the criteria. The main flow diagram of the algorithm is presented in Fig.1..

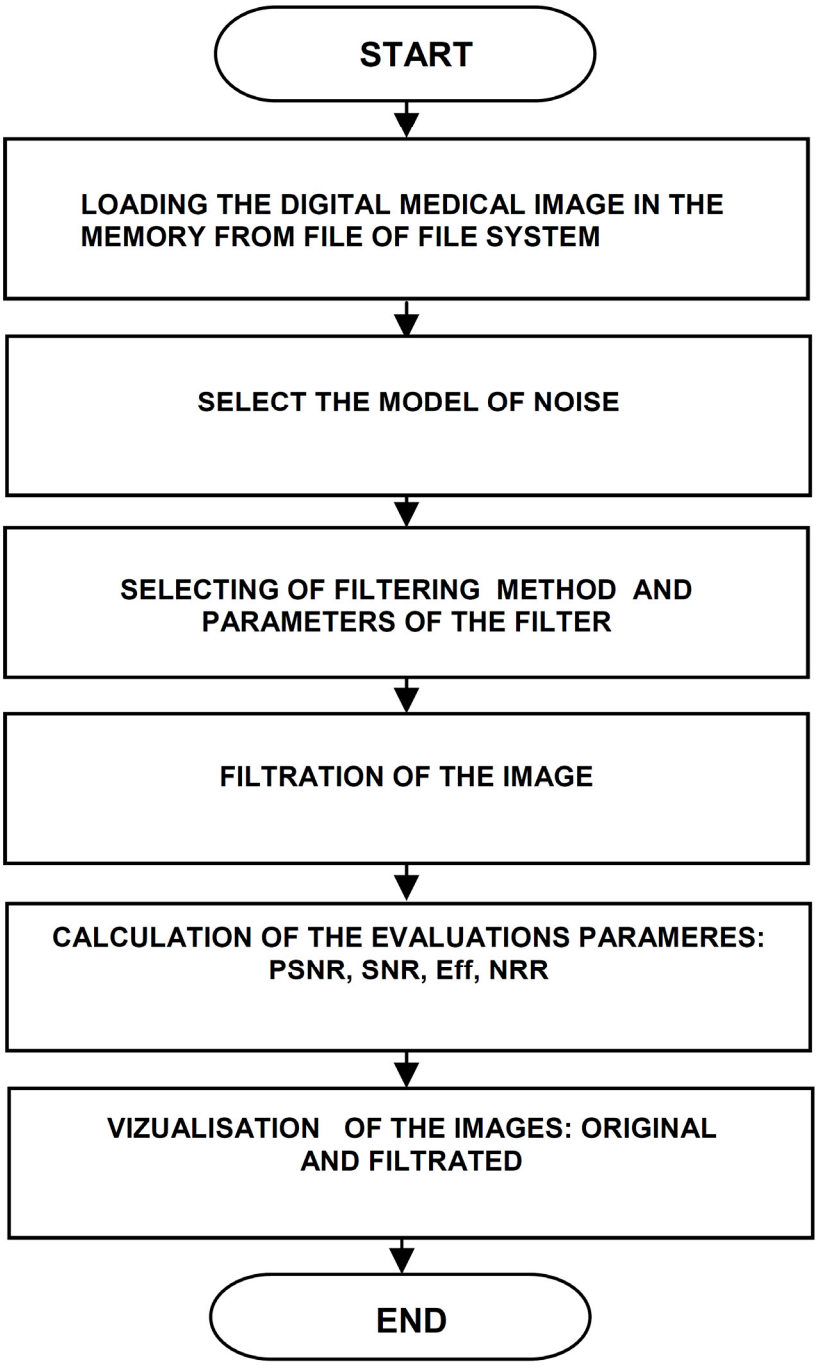


Fig.1 The flowchart of the algorithm for noise reduction

Filtration's methods

The filtration is applied in the spatial and frequency domain.

The filtration in the spatial domain represents an operation between the pixels that are located in the neighborhood. In this case, the neighbors of each pixel in the image are analyzed. The filtration function is represented as a matrix and is called mask of the filter. It has often the following size: 3x3, 5x5, etc., but in any case is smaller than the size of the image. Generally there are two types of filters in the spatial domain: linear (e.g. Gaussian filter) and nonlinear (e.g. Median filter).

The methods of treatment in the frequency domain are based on manipulation of the orthogonal transformation of the image, not on the image itself. The improvement of the image in this domain consists of several basic stages:

- Transformation from the spatial to the frequency domain using strait 2D discrete transformation of the image, e.g. two-dimensional discrete Fourier transformation (2D DFT), cosine-transformation, Hartley-Adamar transformation, Wavelet transformation, etc.(the choice depends on the application);
- Manipulation of transformation factors with operator M(multiplication with the function, which describes the function);
- Performance of reverse discrete transformation from the frequency to the spatial domain.

A. Gaussian Filtration

Significantly improvement of filtration can be achieved, when the mask of the filter coincides with the function that describes Gaussian distribution. In discrete form the Gaussian distribution can be approximated by binominal distribution (binominal Gauss filter) and so the filter becomes particularly suitable for filtration in the spatial domain. Because of the possibility for separately filtration in horizontal direction the sharpness of the image is retained. Another advantage of this filtration is the fast operation.

B. Median Filtration

Median filter is known as nonlinear method that is used to remove noise from MRI images. It is very effective at removing salt and pepper noise. The algorithm of median filter works by moving through the image pixel by pixel, replacing each value with the median value *N* of neighboring pixels. The pixel is calculated by the following order: first the entire pixel values from the pattern of neighbors are sorted into numerical order, and then the pixel being considered is replaced with median pixel value. Median filter is better able to remove noise without reducing the sharpness of the image.

C. Wiener Filtration

An optimization between inverse filtering and noise smoothing is the Wiener filter (nonlinear). This filter removes additive noise and deblurring concurrently. As prove for optimization is the reducing the overall Mean Square Error (MSE). There are two important parts of the operation: inverse filtering and noise smoothing. Wiener filters belong to a kind of optimum linear filters that have the noisy data as input which involves the calculation of difference between the desired output sequences from the actual output. Measurement of the performance can be shown using Minimum Mean-Square Error.

There is also Wiener2 filter that is a 2-D adaptive noise removal filter. This function works as applying a wiener filter which is a type of linear filter to an image adaptively, tailoring itself to local image variance. Wiener2 performs little smoothing by large variance. By small one, wiener2 performs more smoothing. That way leads often to better result than linear filtering. As comparison the adaptive filter is more selective than a comparable linear filter, preserving edges and other high frequency parts of an image. There are no design tasks, the wiener2 function handles all preliminary computations, and implements the filter for preliminary computations, and implements the filter for an input image. Wiener2 filter is best suitable to remove Gaussian noise.

D. Wavelet Filtration

Many of the popular de-noising algorithms suggested are based on wavelet thresholding [Overton and Weymouth, 1979], [Weaver, and all, 1991], [Nowak, 1999]. These approaches attempt to separate significant features from noise in the frequency domain and simultaneously preserve them while removing noise [Pizurica and all, 2006]. The wavelet function can be viewed as a high pass filter, which approximates a data set (a signal or time series). The result of the wavelet function is the difference between value calculated by the wavelet function and the actual data. The scaling function calculates a smoothed version of the data, which becomes the input for the next iteration of the wavelet function. In the context of filtering, an ideal wavelet/scaling function pair would exactly split the spectrum.

The difficulty with wavelet or anisotropic diffusion algorithms is the risk of over-smoothing fine details particularly in low SNR images [Plonka and Ma, 2008].

The general wavelet–based method for denoising and nonparametric function estimation is to transform the data into the wavelet domain, threshold the wavelet coefficients, and invert the transform. We can summarize these steps as:

1. Decompose

Choose a wavelet and a level N. Compute the wavelet decomposition of the image down to level N.

2. Threshold detail coefficients

For each level from 1 to N, threshold the detail coefficients. Hard and soft thresholding are examples of shrinkage rules. After we have determined the threshold, we have to decide how to apply that threshold to our data. The simplest scheme is hard thresholding. The hard thresholding preserves the wavelet coefficients whose absolute values are larger than the threshold, otherwise they are set to zero. In soft thresholding, wavelet coefficients whose absolute values are lower than the threshold are set to zero, otherwise this method shrinks them toward zero.

3. Reconstruct

Compute wavelet reconstruction using the original approximation coefficients of level N and the modified detail coefficients of levels from 1 to N.

E. Homomorphic Wavelet Filtration

The homomorphic filtering technique works in frequency domain. However, before the transformation is taking place, logarithm function has been used to change the multiplication operation in Eq.(4) into addition operation. In the Fourier transform of traditional homomorphic filtering, spatial resolution is lower, and local contrast of image is not increased obviously. Lowpass filtering could reduce noise by smoothing, but the border of image will become to more indistinct. Highpass filtering could enhance the edge of image, but the noise of background will be increased. The standard homomorphic filtering schema can be presented in Fig.2, where DFT is 2D Discrete Fourier Transform, H(u,v) is a filter function, $(DFT)^{-1}$ is 2D Inverse Discrete Fourier Transform and *g(x,y)* is the output image. Using Wavelet Transform, block DFT will be changed with DWT and block $(DFT)^{-1}$ with IDWT.

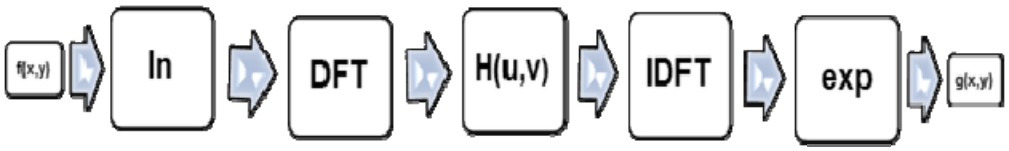


Fig.2 The flowchart of standard homomorphic filtering

The filtration function H(u,v) is based on wavelet decomposition and thresholding of wavelet coefficients. Because the noise of wavelet transform usually concentrate on the state of high resolution, the method is useful to eliminate the noise. This method can be applied in MRI to decry a varying of the intensity in different tissues, which exists because of inhomogeneity of radio pulses.

Experimental part

For the experiments we have used a sequence of 22 MR-images in axial plane, in which the abdominal organs can be well seen (one of them is the spleen). The images are from a study of health in Pomerania (East Germany).The original images were in DICOM format and they were converted to BMP for visualization and experiments. The size of the images is 256x176 pixels and they are originally grayscale. The experiments for noise reduction are made by computer simulation in MATLAB, version 8.1 environment by using of IMAGE PROCESSING and WAVELET TOOLBOXES.

The best results after several filters used for the MRI sequence are achieved with four of them: Wiener filter, Median filter, Wavelet filter and Homomorphic Wavelet filter, which have been already described in the previous section. The best results are obtained by the following conditions: the median filter is with [3x3] neighborhood mask; the wavelet filter and the homomorphic wavelet filters are made on the base of the wavelet decomposition on level 1, using orthogonal wavelets and adaptive threshold of the transformed MR image.

For evaluation of noise reduction by different filters some parameters such as: peak signal-to noise ratio (PSNR), the effectiveness of filtration (Eff), which is equal to the difference of signal-to noise ratio in filtered image (SNR_F) and signal-to noise ratio in noised image (SNR_N), and noise reduction ratio (NRR) are used. The obtained averaging results from simulation are presented in Table 1.

Table 1

Filter	**NRR**	**PSNR** [dB]	**SNR_N** [dB]	**SNR_F** [dB]	**E_{ff}** [dB]
Gaussian	0,671	27,824	21,6530	22,0194	0,3664
Wiener	0,395	31,737	21,6530	22,4101	0,7571
Median	0,463	29,653	21,6530	22,3285	0,6755
Wavelet	0,334	34,676	21,6530	22,6257	0,9727
Homomorphic Wavelet	0,232	36,822	21,6530	23,0404	1,3874

The graphical interpretation of PSNR, E_{ff} and NRR are given in Fig.3.

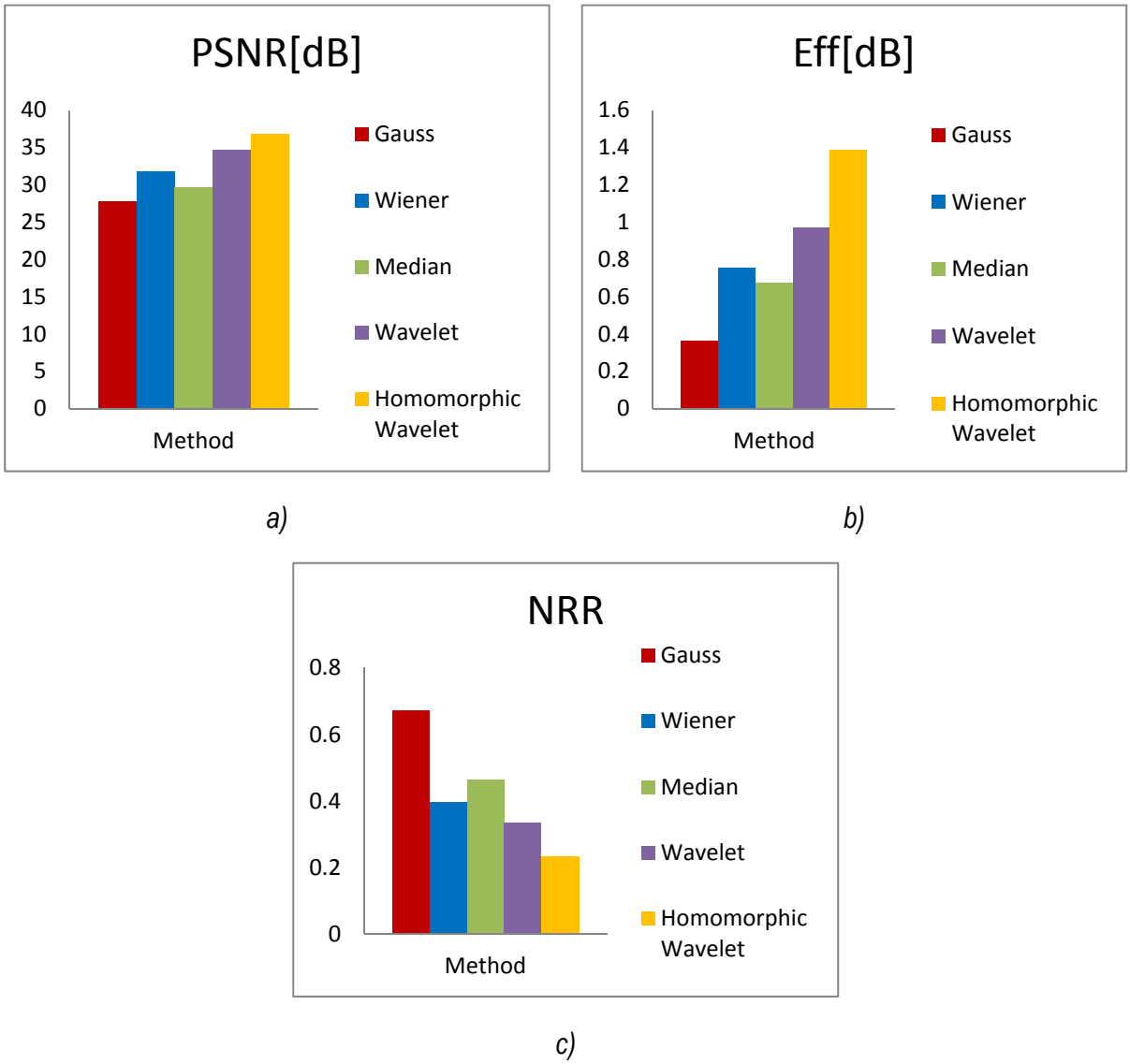
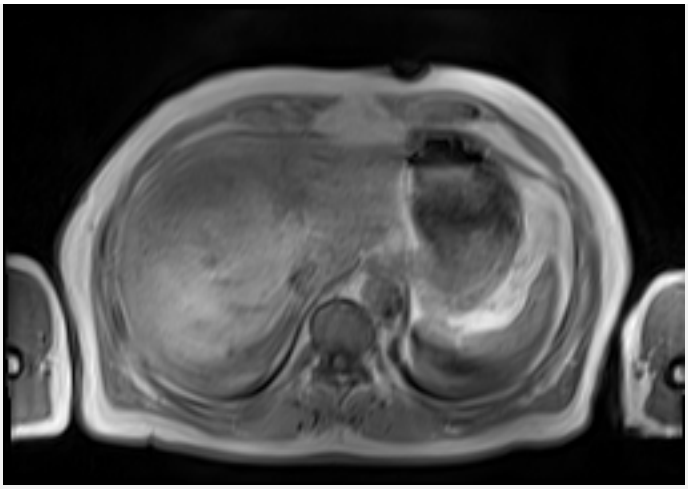
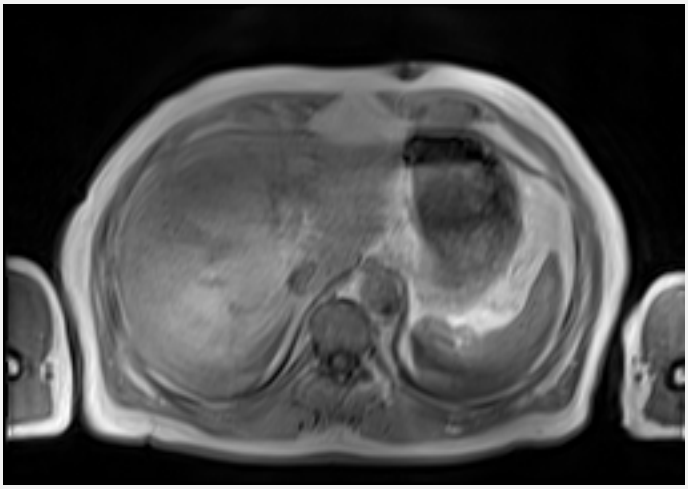


Fig.3 Graphical presentation of the results obtained by investigated methods of filtration for: a) PSNR; b) Effectiveness of filtration (E_{ff}); c) NRR

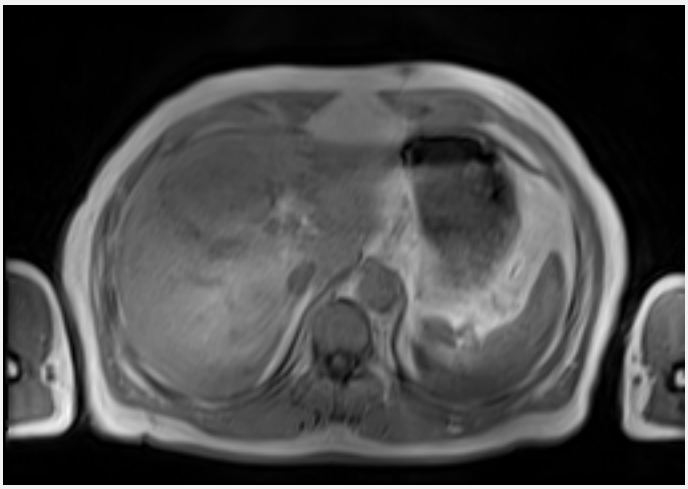
For visualization of the best results were chosen three consecutive images from the sequence. They are given in Fig.4. These images represent as wholeness and in the best manner for observation the human spleen. The original image 1 and its modifications obtained after different filtering methods are given in Fig.5.



Original MRI 1

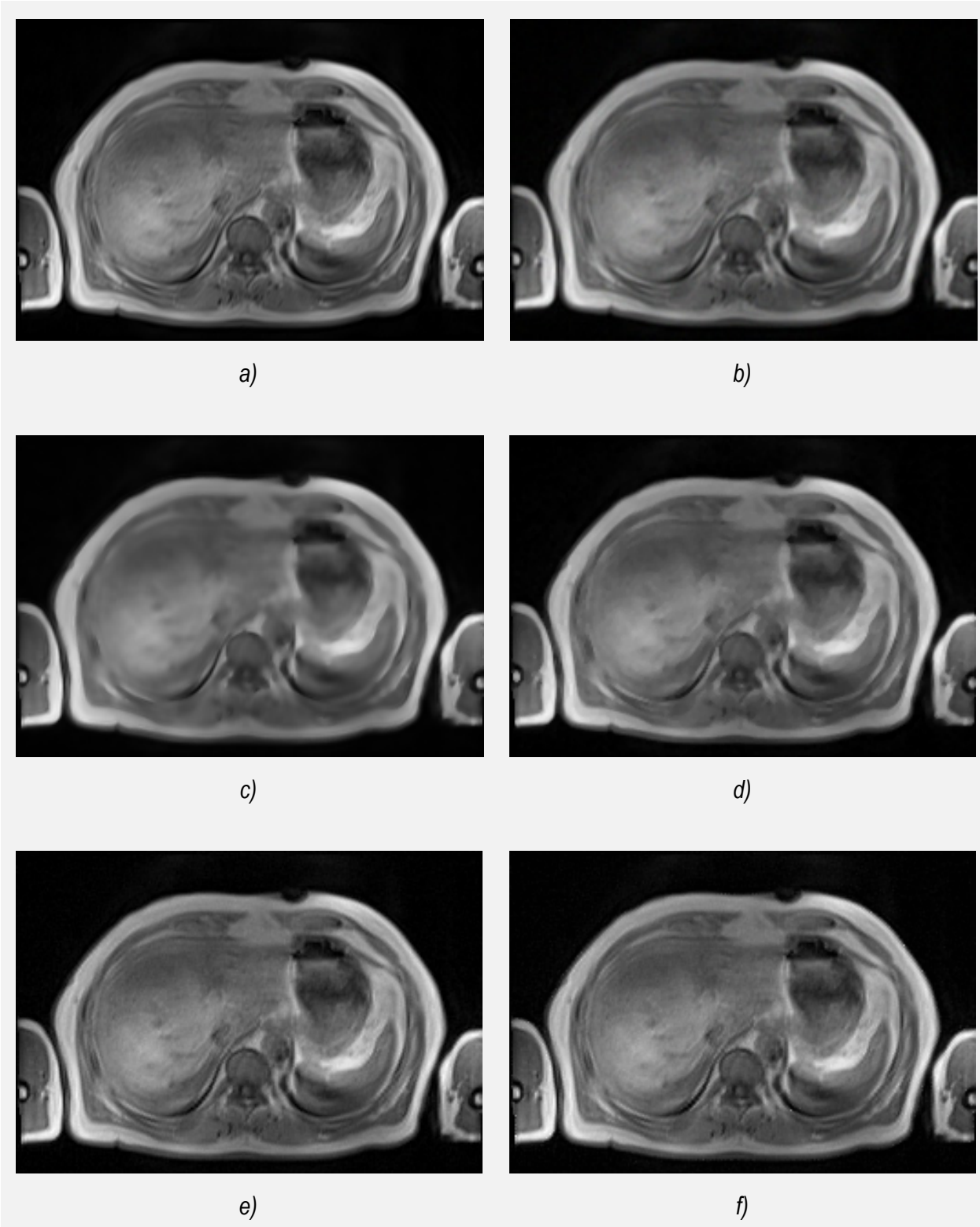


Original MRI 2



Original MRI 3

Fig.4 The original consecutive MR-images from the sequence representing the spleen in axial plane



a) *b)*

c) *d)*

e) *f)*

Fig.5 MRI 1 images obtained by noise reduction with investigated filters: a) original image; b) Gaussian filter; c)Wiener filter; d)Median filter; e)Wavelet filter; f)Homomorphic wavelet filter

The original image 2 and its modifications obtained after different filtering methods are given in Fig.6. The original image 3 and its modifications obtained after investigated filtering methods are presented respectively in Fig.7.

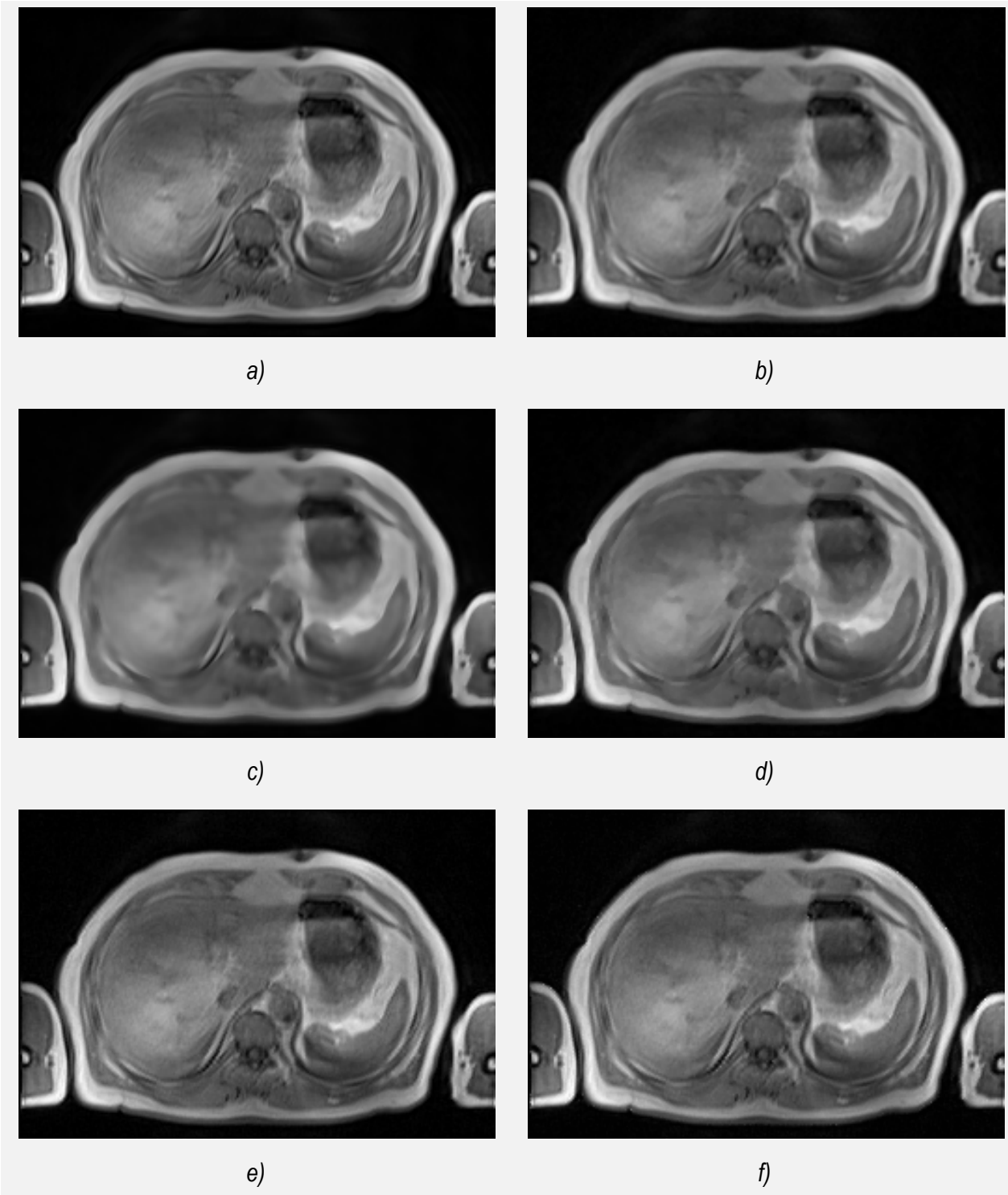


Fig.6 MRI 2 images obtained by noise reduction with investigated filters: a) original image; b) Gaussian filter; c)Wiener filter; d)median filter; e)Wavelet filter; f)homomorphic wavelet filter

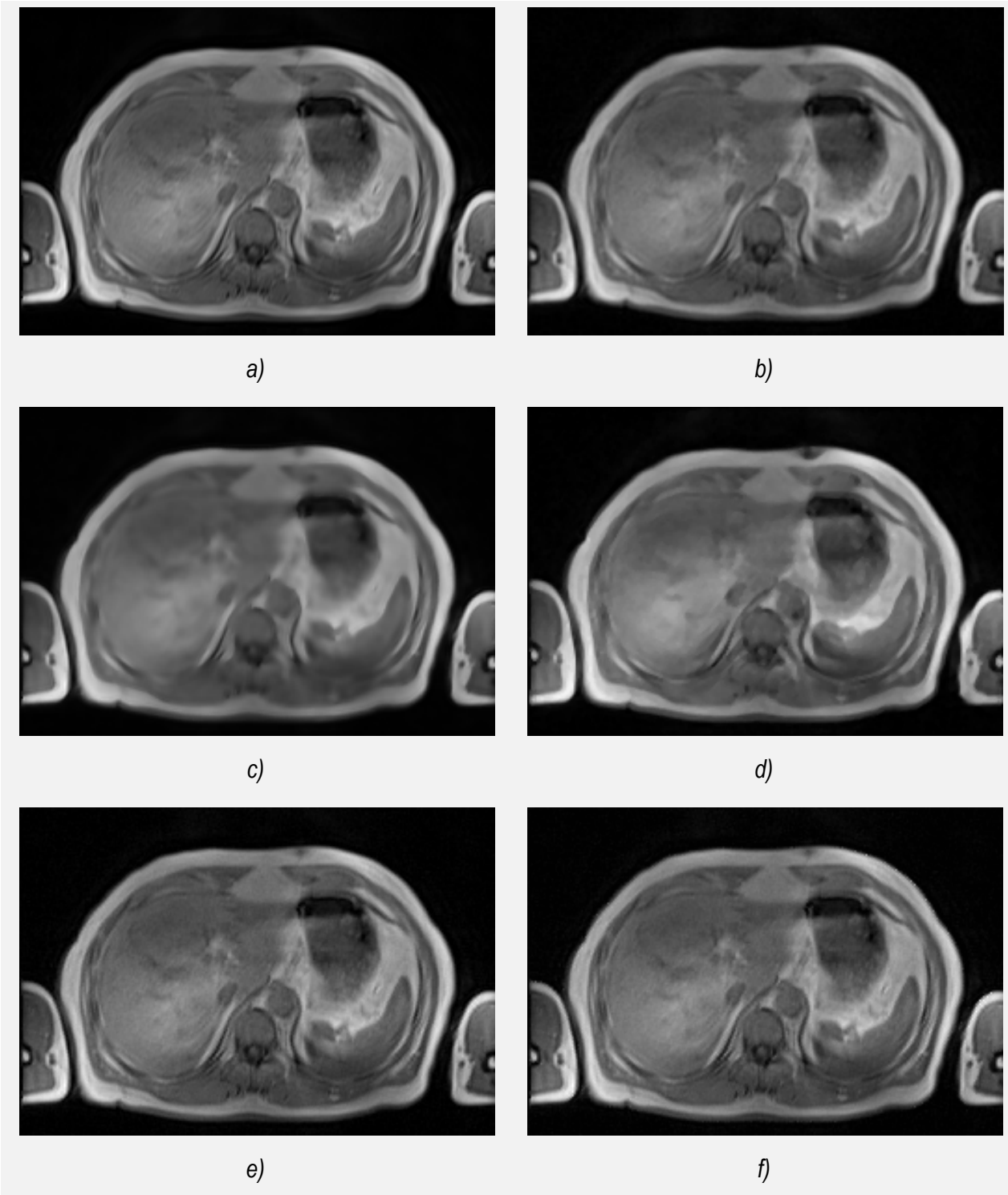


Fig.7 MRI 3 images obtained by noise reduction with investigated filters: a) original image; b) Gaussian filter; c)Wiener filter; d)median filter; e)Wavelet filter; f)homomorphic wavelet filter

The implemented investigations and the obtained results from simulation have shown that by using of homomorphic filter the values of PSNR and the effectiveness of filtration are greater compering to the

other filters. The average value of NRR is around 0.2 and shows that the noise is five times reduced. By noise reduction on the base of wavelet transformation and wiener filter the values of NRR are respectively 0.3 and 0.46. It shows that in these cases the noise is reduced respectively three times and two times. The best results from the homomorphic wavelet filter are obtained by wavelet decomposition on level 1, using orthogonal wavelets coiflet and adaptive threshold of the transformed MR abdominal images.

Conclusion

In this paper a comparative study of noise reducing techniques and various filtering algorithms are implemented on MRI series of abdominal images. MRI images when captured usually have Gaussian noise and Rician noise. Two basic models of noise are applied for investigation of the process of noise reduction for the specific case of the abdominal organs. To reduce the noise filtering algorithms are introduced. The results are analyzed and evaluated on the base of objective estimations parameters and visualization criterion. For evaluation of the methods we propose to be automatically analyzed and compared values of PSNR and Eff, which must be higher and the value of NRR, which must be low in the same time. This simple algorithm is applied and used for optimal choice of parameters of the filters. Median filter performs better result in compare to Gaussian filter. The Wiener filter works better, but more significant results we obtain by wavelet and especially by homomorphic wavelet filter.

In this case the boundaries of the organs are better preserved.

The implemented comparative study and obtained results can be applied:

- for future segmentation of the abdominal organs;
- for clinical diagnosis such as tissue classification;
- restoring textures;
- reconstruction by 3D model of the organs;

Our future investigation will be concentrated in improving of the investigated region of interest (ROI) and choice for suitable algorithm for segmentation of specific abdominal organs, especially of spleen.

Acknowledgement

The paper was supported by the PhD Students Research Project “Algorithms for automatic segmentation of large-scale medical image data, NIS № 162 PD0001-07/2016.

Bibliography

[Anisha and Krishnan, 2015] Anisha, S.R PG Scholar,Dr J Venugopala Krishnan. In: Comparison of Various Filters for Noise

Removal in MRI Brain Image, International Conference on Futuristic Trends in Computing and Communication (ICFTCC-2015) . Department of ECE, Jeppiaar Engineering College, Chennai-119.

[Aja-Fern'andez and all, 2008] S. Aja-Fern´andez, M. Niethammer, M. Kubicki, M. E. Shenton, C.-F. Westin, In: Restoration of DWI data using a Rician LMMSE estimator, IEEE Trans. Med. Imaging 27 (10) (2008) 1389–1403.

[Bernsteinan and all, 1989]M. A. Bernstein, D. M. Thomasson, and W. H. Perman, In: Improved detectability in low signal-to-noise ratio magnetic resonance images by means of phase-corrected real construction. Med. Phys., vol. 16,no. 5, pp. 813–817, 1989.

[Chang and all, 2006] S. G. Chang, Y. Bin, and M. Vetterli, In: Adaptive wavelet thresholding for image denoising and compression, IEEE Trans. On Image Processing, vol. 9, No. 9. pp. 1532-1546, Sept. 2006.

[Counter and all, 2000]A. Counter, A. Olofsson, E. Borg, B. Bjelke, A. H¨aggstr¨om, H. Grahn. In: Analysis of magnetic resonance imaging acoustic noise generated by a 4.7 t experimental system, Acta Oto-Laryngologica 120 (6) (2000) 739–743.

[Coup´e and all, 2010] P. Coup´e, J. V. Manj´on, E. Gedamu, D. Arnold, M. Robles, D. L. Collins, Robust Rician noise estimation for MR images, Med. Imag. Anal.14 (4) (2010) 483–493.

[Fern´andez and Vega] Santiago Aja-Fern´andez and Antonio Trist´an-Vega. In: A review on statistical noise models for Magnetic Resonance Imaging. LPI, ETSI Telecomunicaci´on, Universidad de Valladolid, Spain

[Georgieva,2015] Veska Georgieva, Technical University Sofia. In: Medical Image Processing. TU-Sofia, Bulgaria

[Getreuer and all, 2011] P. Getreuer, M. Tong, L. Vese, A variational model for the restoration of MR images corrupted by blur and rician noise, in: G. Bebis, R. Boyle,B. Parvin, D. Koracin, S. Wang, K. Kyungnam, B. Benes, K. Moreland, C. Borst, S. DiVerdi, C. Yi-Jen, J. Ming (Eds.), Advances in Visual Computing, Vol. 6938 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, pp. 686–698.

[Gudbjartsson and all, 1995] H. Gudbjartsson, S. Patz, In: The Rician distribution of noisy MRI data, Magn. Reson. Med. 34 (6) (1995) 910–914.

[Huber and all, 2001]Huber ME, Hengesbach D and Botnar RM, " Motion artifact reduction and vessel enhancement for

free-breathing navigator-gated coronary MRA using 3D k-space reordering," Magn Reson Med 2001 Apr;45(4):645-52.

[Kruger and Glover,2001], G. Kruger, G. H. Glover. In: Physiological noise in oxygenation-sensitive magnetic resonance imaging, Magn. Reson. Med. 46 (2001) 1631–637.

[Lin and Willson, 1988] H. Lin and A. N.Willson Jr., In: Median filters with adaptive length. IEEE Trans. Circuits Syst., vol. 35, no. 6, pp.675–690, Jun. 1988.

[McGibney and Smith, 1993] G. McGibney, M. Smith. In: Unbiased signal-to-noise ratio measure for magnetic resonance images, Med. Phys. 20 (4) (1993) 1077–1078.

[Mirowitz, 1999] Mirowitz SA, "MR imaging artifacts. Challenges and solutions," Magn Reson Imaging Clin N Am 1999 Nov;7(4):717-32

[Nowak, 1999], R. D. Nowak. Wavelet-based Rician noise removal for magnetic resonance imaging. IEEE Trans Image Process 1999; 8:10:1408–1419.

[Overton and Weymouth, 1979] K. J. Overton and T. E. Weymouth,In: A noise reducing preprocessing algorithm. Proc. IEEE Computer Science Conf. Pattern Recognition and Image Processing, Chicago, IL, 1979, pp.498–507.

[Pizurica and all, 2006] A. Pizurica, A. M. Wink, E. Vansteenkiste, W. Philips, and J. B. T. M. Roerdink. A review of wavelet denoising in MRI and ultrasound brain imaging. Current Med Imag Rev 2006; 2:2:247–260

[Plonka and Ma, 2008] Gerlind Plonka and Jianwei Ma. Nonlinear Regularised Reaction-Diffusion Filter for Denoising of Images with Textures. IEEE Trans. Image Processing 2008; 17:8:1283–1294

[Ruan] Chun Ruan. In: MRI Artifacts: Mechanism and Control

[Sarode, Deshmukh, 2010-11] Milindkumar V. Sarode, Dr.Prashant R. Deshmukh. In: Performance Evaluation of Rician Noise Reduction Algorithm in Magnetic Resonance Images, Journal of Emerging Trends in Computing and Information Sciences. Department of Computer Science and Engineering Jawaharlal Darda Institute of Engineering& Technology, Yavatmal (M.S.) India Department of Computer Science & Engineering and Information Technology SIPNA's College of Engineering Amravati (M.S.) India.

[Sun and Neuvo, 1994] T. Sun and Y. Neuvo, In: Detail-preserving median based filters in image processing. Pattern Recognit. Lett., vol. 15, pp. 341–347, Apr. 1994.

[Tisdall and Atkins, 2005] D. Tisdall and M. S. Atkins. MRI denoising via phase error estimation. Proc SPIE Med Imag 2005; 646–654.

[Toennies, 2012] K.D.Toennies, Guide to Medical Image Analysis, Methods and Algorithms, Springer-Verlag London, 2012.

[Weaver, and all, 1991] J. B. Weaver, Y. Xu, D. M. Healy Jr., and L. D. Cromwell. Filtering noise from images with wavelet transforms. Magn Reson Med 1991; 21:2:288–295.

Authors' Information

Antoniya Mihailova *– PhD Student at Technical University Sofia, Bulgaria; e-mail: antonianankova@gmail.com*

Major Fields of Scientific Research: Medical image processing research, Computer and telecommunication technologies

Veska Georgieva *- Assoc. Prof. PhD at the Department of Radio communication and video technology, Faculty of Telecommunications in Technical University Sofia, Bulgaria; e-mail: vesg@tu-sofia.bg*

Major Fields of Scientific Research: Signals and systems, Medical image processing

INVESTMENT IN ICT IN POLAND FROM THE EU FUNDS - OPPORTUNITIES AND CHALLENGES

Małgorzata Dziembała

***Abstract**: Actions are undertaken in the EU in order to improve the competitiveness of this international grouping and the dynamics of economic growth, also via counteracting the digital exclusion. It should upgrade human capital and the innovativeness of the European economy. In the EU financial instruments supporting the digital inclusion are adopted. The goal of the paper is to present the significance of EU funds for supporting the infrastructure and ICT services in the EU and in Poland, as well as, to point out challenges related to the effective implementation of projects counteracting digital exclusion.*

***Keywords**: digital exclusion, ICT, EU structural funds, EU cohesion policy, Digital Economy*

Introduction

Actions are undertaken in the EU in order to improve the competitiveness of this sector and the dynamics of economic growth, also via counteracting the digital exclusion. It should upgrade human capital and the innovativeness of the European economy. In the EU financial instruments are implemented, regulations supporting the digital inclusion are adopted, as well as, through the involvement of the stakeholders in projects aiming at the digital inclusion, it will be possible to reduce the existing inequalities in this field [European Parliament 2015, pp. 6-7]. Europe 2020 strategy points to three priorities: the smart, sustainable and inclusive growth [Komisja Europejska 2010]. The effective implementation of these priorities will be possible thanks to the support of information and communication technologies (ICT), also within the framework of the cohesion policy.

In the EU budget the amount of financial resources for the years 2014-2020 provided for under the structural and investment funds is 454.1 billion EUR, among which 43.3% of the allocation will come from the European Regional Development Fund (ERDF), 21.7% of funds shall be transferred to the European Agricultural Fund for Rural Development (EAFRD), 19% shall be transferred to the European Social Fund and 14% of the total amount of funds shall be transferred to the Cohesion Fund [https://cohesiondata.ec.europa.eu/overview]. These funds become an important investment support instrument of EU policies, in particular, through better access to broadband networks and ICT services [European Commission 2015, p. 2]

The goal of the article is to present the significance of EU funds for supporting the infrastructure and ICT services in the EU and in Poland, as well as, to point out challenges related to the effective implementation of projects counteracting digital exclusion.

ICT Development in Poland in contrast to the EU

Information and communication technologies cover the family "of technologies processing, collecting and forwarding information in electronic form" [GUS 2015, p. 17]. However, not all social groups benefit equally from advanced technology and its dynamic growth. Therefore a digital divide concept emerged which is recognized as a "gap between individuals, households, enterprises, businesses and geographic areas at a different socio-economic levels with regard both to their opportunities to access information and communication technologies […] and to their use of the Internet for a wide variety of activities" [OECD 2001, p. 5]. However, the digital exclusion concerns two spheres: the material and the immaterial one relating to knowledge, motivation and needs [Bednarczyk 2014, pp. 1-2]. On the one hand, there are inequalities between those who have access to the Internet and those who lack the physical access to it. Attention is also drawn to the second level of digital exclusion occurring due to the diverse level of skills of people, bearing in mind their ability to efficiently find information. Therefore, it is essential not only to ensure access to the ICT infrastructure, but it is also necessary to consider investments within the range of trainings and support so that specific skills to use ICT could be acquired [Hargittai 2002].

According to the Digital Economy Index 2016 (DESI 2016) Poland ranked 22nd out of the 28 EU countries and thus found itself in the group of countries lagging behind to which Bulgaria, Cyprus, the Czech Republic, France, Hungary, Poland and Slovakia belong. These countries are improving their results more slowly than the others. Indicators concerning access to the broadband Internet in Poland are not satisfactory. ICT tools are not often used by the public sector. Poland experienced a moderate progress in the scope of the Internet use and the digital technology integration, improved results in terms of the quality of networks. However, results in the scope of the level of public digital services and human capital for the digital economy worsened. With regard to the assessment of consecutive indicators describing the quoted index, Poland is below the EU average in terms of the quality of networks. 86% of households (2015) were covered by the broadband network in Poland and thus our country remained in the last place in the EU. The indicator for the EU was 97%. Only 61% of households have access to high-speed fixed broadband, as compared to 71% in the EU (2015).Only 57% of households have such a fixed Internet connection and Poland ranked 26th. However, there is a particularly high level of development of mobile broadband services. With regard to the range of fixed communications 94 out of 100 Poles choose broadband (fifth place in the EU). This means that it is

necessary to develop broadband infrastructure – increase broadband coverage and boost its further growth. Poland's progress towards improving human capital for the needs of the digital economy is moderate. There are only 65% of network users (24th place in the EU). Only 40% of people (2015) have basic digital skills, as compared to 55% in the EU what qualifies Poland in 26th position. There is only a 3% share of ICT specialists among the employed, what ranks Poland 19th in the EU. Therefore, it is necessary to improve the digital skills of the consecutive social groups in Poland. In terms of the digital technology integration, results achieved by Poland are also below the average. The delays concern the digitization process of Polish enterprises, social media, cloud services are not frequently used. Only 9.6% of enterprises conduct sales via the Internet, as compared to 16% in the EU (2015). "The Strategy for Innovation and Efficiency of the Economy – Dynamic Poland 2020" whose aim is to support digitization was worked out in 2013. In the field of digital public services, indicators relating to Poland are also below the EU average, nevertheless progress in this field is significant. There is a low level of e-Administration. As far as e-Administration use is concerned, electronic forms are submitted by merely 22% of Internet users (21st place in the EU), 46% of people used online banking in Poland (22nd place), as compared to 57% in the EU in 2015. In terms of digitization of enterprises Poland also lags far behind [https://ec.europa.eu/digital-single-market/en/scoreboard/poland; Digital Economy and Society Index 2016; Sprawozdanie okresowe; Kwieciński 2014] (illustration 1).

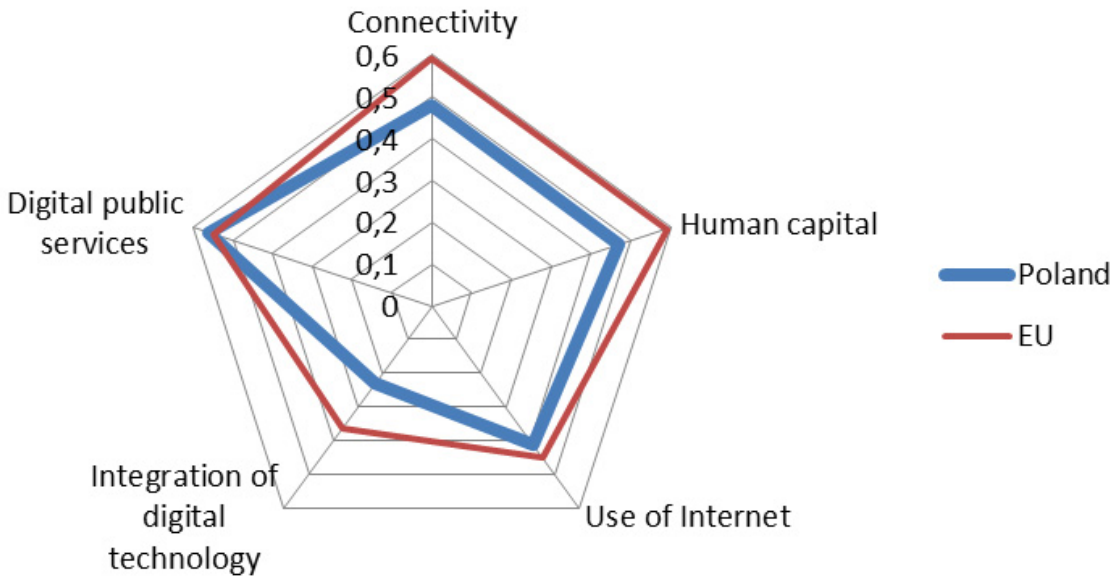


Illustration 1. Position of Poland and the EU in light of the dimensions of the Digital Economy and Society Index 2016

Source: [Digital Economy and Society Index 2016]

Development in Poland in contrast to the ICT support within the EU cohesion policy in the years 2007-2020

Promoting ICT development constitutes an important support guideline not only within the cohesion policy, but also other EU policies, including research and development. In this context a wide range of policies through the implementing instruments promotes ICT development, since they also help achieve the goals set within these policies. New opportunities were created by the instruments outlined in the Juncker plan and implemented from September 2015, ie. Connecting Europe Facility (CEF) supporting the development of projects of common interest in the following sectors: transport, telecommunications, energy (in the scope of infrastructure and services), as well as, the instrument: European Fund for Strategic Investment (EFSI) [Regulation (EU) No 1316/2013]. Within their frameworks new financial instruments are introduced, among them: guarantees or equity [Regulation (EU) No 1316/2013]. 1.141 billion EUR [Art. 5, Regulation (EU) No 1316/2013] will be appropriated from the total CEF budget for the telecommunications sector, out of which about 870 million EUR will be directed to the digital service infrastructure [https://ec.europa.eu/digital-single-market/en/connecting-europe-facility]. ICT support was also foreseen within the consecutive framework programmes and currently the Horizon 2020 programme is being implemented [European Parliament 2015, p. 6]. Further considerations shall be limited only to the support provided within the cohesion policy and the European investment funds in the two consecutive programming periods 2007-2013 and 2014-2020.

Within the cohesion policy in the years 2007-2013, investments concerning ICT or associated projects were, first of all, financed from ERDF and the funds amounted to about 14.6 billion EUR. Consequently access to the broadband network improved for more than 4.7 million EU citizens. Support also came from ESF, also through a better use of ICT, the adaptation of the skills of employees to the needs of employers or through the acquisition of skills related to the use of ICT among older people [http://ec.europa.eu/regional_policy/sources/docgener/informat/2014/fiche_ict_pl.pdf]. The number of supported projects relating to ICT amounted to over 20 thousand. The support also related to investments associated with the broadband implementation and the incurred costs related, among others, to the use of the passive infrastructure, procedures connected with obtaining permits [Polityka spójności, pp. 4-5]. However, the largest number of ICT projects from the EU funds was supported in Spain – more than 10 thousand projects - and then in Hungary and Portugal. However, additionally 1.8 million people were supported in the scope of broadband access (EU countries) [Factsheet 2013, p. 4.]

Despite investments in this field the percentage of people (aged 16-74) in some EU countries who have never used the Internet exceeded 20% and among these countries were: Latvia, Estonia, Hungary, Slovenia, Lithuania, Croatia, Cyprus, Poland (28.1%), Portugal, Italy, Greece, Bulgaria (37.1%) and

Romania (38.6%) in 2014. In Denmark, on the other hand, the percentage of people who have never used the Internet amounted to 2.62% [Polityka spójności, p. 5].

How are the funds from the programmes for the years 2007-2013 actually used for the infrastructure and ICT services? In the light of results of the Strategic Report on the implementation of cohesion policy programmes 2007-2013 [Komisja Europejska 2013; European Commission 2013 a], the biggest support in terms of ICT related to the convergence objective (table 1). ICT support constituted 4.2% of the budget for the cohesion policy [Factsheet 2013, p. 2].

Table 1. Support from structural funds and the Cohesion Fund for IT services and infrastructure in the years 2007-2011. Source: [European Commission 2013 a, pp. 49-50].

	Decided Ops Million € (a)	**% share of total SF per obj**	**% of total all funds**	**Allocated to selected projects 2007-2011 Million € (b)**	**% (c=b/a)**
All Objectives	**346 717.2**		**100.0%**	**246 983.9**	**71.2%**
IT services and infrastructure	14 446.0		4.2%	8 854.8	61.3%
Objective: Convergence	**283 657,7**	**100.0%**	**81.8%**	**198 682.4**	**70.0%**
IT services and infrastructure	11 537.8	4.1%	3.3%	7 158.4	62.0%
Objective: Regional Competitiveness and Employment	**55 154.3**	**100.0%**	**15.9%**	**42 220.1**	**76.5%**
IT services and infrastructure	2 382.3	4.3%	0.7%	1 385.5	58.2%
Objective: European Territorial Cooperation	**7 905.1**	**100.0%**	**2.3%**	**6 081.3**	**76.9%**
IT services and infrastructure	525.8	6.7%	0.2%	310.9	59.1%

In terms of the value of funds appropriated for ICT, the biggest funds were appropriated within the convergence objective – 11.5 billion EUR, what constituted 4.1% of the total allocation of EU funds for this objective. However, the highest share of the allocation of EU funds for ICT was noted within the European Territorial Cooperation objective, although in absolute terms funds allocated for ICT within this objective were the lowest (525.8 million EUR). Progress in the scope of the selection of projects was varied.

As far as the value of planned outlays is concerned, 2.3 billion EUR is planned for the telecommunications infrastructure, including the broadband network, whereas, 5.1 billion EUR was appropriated for services and applications for residents - circa 1/3 of the allocation for ICT. However, the rate of project selection amounted to 61.3%, being below the country average [see: Factsheet 2013]. Delays are also notable in terms of consecutive categories of expenditures (table 2).

Table 2. Categories of actions carried out within the thematic priority: IT services and infrastructure in the years 2007-2011 Source: [Factsheet: 2013, p. 5]

Category	Decided Ops – Million € (a)	% Decided Ops of Total Decided (b)	Allocated to selected projects 2007-2011 – million € ©	Rate for selection 2007-2011 (d=c/a)
Telephone infrastructures **(including broadband networks)**	2 244.6	0.6%	1 300.2	57.9%
Information and communication technologies (…)	3 516.4	1.0%	2 231.9	63.5%
Information and communication technologies (TEN-ICT)	490.8	0.1%	227.0	46.3%
Services and applications for citizens (e-health, e-government, e-learning, e-inclusion, etc.)	5 126.4	1.5%	3 683.9	71.9%
Services and applications for SMEs (e-commerce, education and training, networking, etc.)	1 499.6	0.4%	519.3	34.6%
Other measures for improving access to and efficient use of ICT by SMEs	1 568.1	0.5%	892.5	56.9%
Total IT networks and services	**14 446.0**	**4.2%**	**8854.8**	**61.3%**
Total all themes	**346 717.2**		**246 983.9**	**71.2%**

In the new financing perspective, the funds of the cohesion policy and the European investment funds are to contribute to the achievement of eleven thematic objectives, among which one objective regarding the accessibility improvement, as well as, the degree of using ICT and its quality was distinguished [Regulation (EC) No. 1303/2013]. The ERDF regulation precisely defines the investment priorities relating to ICT: improvement of broadband access and high-speed networks, promoting the adoption of new technologies and networks in favor of the digital economy, promoting the development of products and services relating to ICT and electronic commerce, increase of use [Regulation (EC) No. 1301/2013, Art. 5, item 2; http://ec.europa.eu/regional_policy/pl/policy/themes/ict/]. ERDF with a budget of 196 billion EUR is an important source of support and 13.3 billion EUR, ie. 6.7% of its funds shall be designated for ICT [https://cohesiondata.ec.europa.eu/funds/erdf]. ICT will also be supported by EAFRD. Each member state in the operational programmes implementing this policy designated funds for supporting ICT development.

In connection therewith, in the new financing perspective the cohesion policy in the field of ICT concentrates on infrastructure investments in the European regions, especially in rural areas, in regions of a low level of development and remote regions. Investments are still going to be made in order to ensure access to the broadband networks, which are to improve the efficiency of the enterprises, increase the teleworking opportunities, and promote e-Health more intensively. The development and improvement of ICT instruments, the benefit from using the opportunities of ICT technology by different entities in order to render health services: e-Health, its application in the area of e-Administration, e-Education, e-Integration, e-Culture or improving digital and entrepreneurial skills – are important investment trends promoted in the new perspective 2014-2020. This support trend can also be implemented within other thematic objectives. Priorities relating to ICT investments and priorities connected therewith were formulated within the strategy of smart specializations. As highlighted, about 15% of priorities relating to RIS3 will be associated with ICT. Such a focus on ICT is especially visible in the following countries: Poland, Italy, Spain, Greece and Portugal [http://ec.europa.eu/regional_policy/sources/docgener/informat/2014/fiche_ict_pl.pdf; Polityka spójności, p. 5, p. 7].

Poland, Spain, Italy, the Czech Republic, Greece, Slovakia, Hungary and Romania will receive the biggest funds for ICT in the new financing perspective 2014 –2020 (illustration 2). However, if the percentage of the support of this thematic objective were taken into consideration in relation to the total allocation of funds (within the cohesion and the investment policy) – the biggest share would be in Cyprus (11%), in Ireland (8%), Spain (7.5%), France (7.4%), Italy and in Slovakia (6.1%), in Sweden (5.8%). In Poland, on the other hand, it is planned to appropriate 4.2% of the total allocation for ICT in the new programming period [http://ec.europa.eu/regional_policy/pl/policy/themes/ict/]. It should be

noted that the total funds for the support of this area from EFRR and EAFRD are to amount to over 14 billion EUR, out of which 93.6% will come from EFRR, ie. over 13.3 billion EUR and the remaining amount, ie. 914.5 million EUR shall be allocated from the European Agricultural Fund for Rural Development [https://cohesiondata.ec.europa.eu/themes/2].

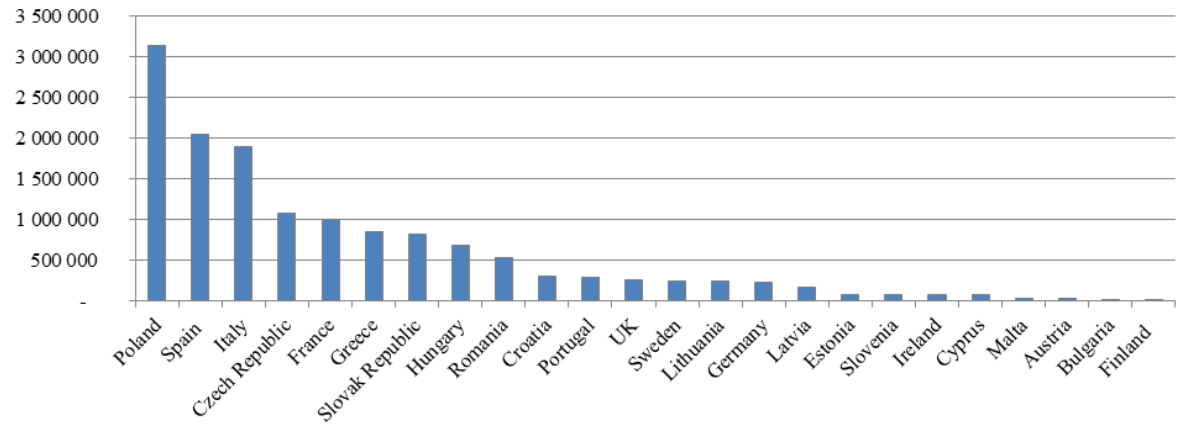


Illustration 2. Planned ICT support from investment funds in the years 2014-2020 in thousand EUR

Source: [https://cohesiondata.ec.europa.eu/themes/2]

The analysis of the main trends of financial support shows that 6 billion EUR will be disbursed from ERDF and the Cohesion Fund for the infrastructure and the digital broadband networks, up to 10 billion EUR will be appropriated for promoting new products and services in the scope of ICT and the e-Commerce market and two billion EUR will relate to the promotion and expansion of various e-Services [Polityka spójności, p. 4]. Objectives were also formulated which are to be achieved thanks to investments from ERDF in favor of ICT: 77 thousand enterprises should receive support, 54 thousand enterprises should receive grants, 5.8 thousand enterprises are to take advantage of financial instruments, consulting services will be provided to13.8 thousand enterprises, over 14 thousand households are to gain access to the broadband networks. In order to use the EU funds, it is necessary to work out the digital development strategy [https://cohesiondata.ec.europa.eu/themes/2].

Challenge of ICT development in Poland within the cohesion policy

In Poland projects connected with ICT are intensively supported from cohesion funds. In the years 2007-2013 nearly 5.2 billion PLN was designated for projects related to the creation of the broadband Internet infrastructure, launched within the regional operational programmes, Eastern Poland Operational Programme and Innovative Economy Operational Programme (measure 8.4) [Sieci

szkieletowo-dystrybucyjne, p. 3]. What effects did the construction of the Regional Broadband Networks referred to as “construction projects of digital motorways” actually bring? An operating network of a length of 23 thousand kilometers was constructed, 2 927 nodes were developed. In the new financing perspective access networks are implemented within the Digital Poland Operational Programme [Sieci szkieletowo-dystrybucyjne, p. 34].

5.4% of funds from the total volume of funds allocated to Poland were used for the ICT infrastructure and services within the cohesion policy 2007-2013. A higher share of the allocation of funds for this objective was merely noted in Denmark (6.5%), Finland (9%), Greece (6.2%), Italy (6%), Sweden (11.8%), Slovakia (8.8%), CB (6.7%) (table 3).

Table 3. Value of the allocation and use of EU funds in Poland in the years 2007-2011

Source: Excel tables showing project selection by Objectives and Member State http://ec.europa.eu/regional_policy/pl/policy/how/stages-step-by-step/strategic-report/

	Decided OPs (a) - in M.€	**% of National SF/CF**	**Allocated to selected projects AIR 2011 (b) - in M.€**	**% (c=b/a)**
Poland	**67 185.5**	**100.0%**	**45 980.5**	**68.4%**
Innovation & RTD	9 309.8	13.9%	6 144.2	66.0%
IT services and infrastructure	3 630.3	5.4%	2 280.3	62.8%
Other SME and Business support	3 473.6	5.2%	2 442.4	70.3%
Energy	2 311.5	3.4%	1 375.1	59.5%
Environment	6 770.7	10.1%	5 326.7	78.7%
Culture, heritage and tourism	1 995.7	3.0%	1 591.8	79.8%
Urban and territorial dimension	1 005.8	1.5%	879.2	87.4%
Rail	5 557.1	8.3%	2 186.5	39.3%
Road	15 741.1	23.4%	12 669.1	80.5%
Other transport	4 138.3	6.2%	2 004.3	48.4%
Labour market	2 833.6	4.2%	1 755.5	62.0%
Social Inclusion	1 158.0	1.7%	711.0	61.4%
Social infrastructure	2 710.8	4.0%	2 197.8	81.1%
Human capital	3 858.5	5.7%	3 079.7	79.8%
Capacity Building	2 690.9	4.0%	1 336.9	49.7%

In the years 2007 –2013, within the Innovative Economy Operational Programme, actions were undertaken in Poland in favor of the broadband Internet development, known as last mile construction projects. Regional programmes whose goal was to construct access networks were supported, however delays occurred in the course of carrying out a few projects. This objective received support within the Regional Operational Programme, Eastern Poland Operational Programme. However, there was little progress in carrying out projects supported from rural funds. Systemic problems also emerged, among them fragmentation of actions and their poor coordination, the delays resulted also, among others, from tender procedures. Irregularities emerged in the digitisation of public services. Information Society projects did not significantly improve the effective governance, either. Nevertheless, the funds contribute to the development of the Information Society, as well as, the ICT industry [Krzyżanowska 2014, Kwieciński 2014].

Special financing opportunities for these types of undertakings were created in the new financing perspective 2014-2020, as Poland is the main beneficiary of the cohesion policy. The funds foreseen in the Partnership Agreement for ICT amounted to 3.08 billion EUR, ie. 3.7% of allocation for Poland. Financing should only be provided by ERDF. Not enough attention was paid to the private sector in terms of supporting Information Society, focus on thematic objective 2 and a weak link with other thematic objectives. These are the weaknesses of solutions adopted in the current perspective [Kwieciński 2014].

Regional operational programmes and 5 horizontal programmes, among them the Digital Poland Operational Programme with a budget of 2 172.5 million EUR solely dedicated to the ICT sector will be executed in the years 2014-2020 [https://www.polskacyfrowa.gov.pl/strony/o-programie/zasady/finansowanie/]. A better coordination and a clear intervention logic distinguishes this Operational Programme [Krzyżanowska 2014]. Projects supported in this programme cover the following three groups: related to the broadband infrastructure – access to the high-speed Internet for all, increase of the availability of public services and projects expanding the use of the Internet and improving and developing the digital competences of the society [https://www.polskacyfrowa.gov.pl/strony/o-programie/zasady/dla-kogo-jest-program/;https://www.polskacyfrowa.gov.pl/strony/o-programie/zasady/co-mozna-zrealizowac/]. The support covers direct support, including infrastructure relating to the broadband Internet, as well as, indirect support, eg. relating to expenditure of different entities on ICT [Kwieciński 2014].

The programme supports investments relating to the expansion of the access to broadband networks, development of products and services based on ICT, an increase in the use of TIK in services, eg. e-Administration. The top priorities were: universal access to high-speed Internet, e-Administration and open office, digital competences of the society. Projects relating to the construction, expansion, as well

as, the reconstruction of the access network or the supplementation of the already existing telecommunications infrastructure will be supported within this first priority axis. Within the second priority axis the formulated objectives are related to the improvement of access and the quality of public e-Services, also the improvement of the functioning of the government administration – the so-called improvement of the digital efficiency of offices, the supported projects will relate to the improvement of the availability of public sector information [on the basis of Fundusze Europejskie 2014, pp. 18-20; Program Operacyjny 2014]. Support will be provided for ICT also within the Regional Operational Programmes, whereas the scope of supported actions is differentiated in regions, among others, by focusing on e-Services, supporting the use of ICT by enterprises or promoting ICT in order to limit the phenomenon of digital exclusion [Krzyżanowska 2014].

Benefits from financing ICT operations do not have to be direct and instantaneous and these types of investments do not always lead to the achievement of cohesion. As pointed out, there is a digital divide between rural and urban areas in the new EU member states, regional variations occur in terms of access to the broadband infrastructure, depending on the existing settlement pattern (lowest access) in rural areas. The varied settlement pattern also leads to the digital divide. The share of the rural population in the new EU member states is significant, however, it seems that the population living in those areas is lagging behind in terms of access to broadband networks, services connected therewith. Therefore different IT strategies, IT policies and regional strategies need to be adopted. It is emphasized that outdated technology is often supported from public funds, those funds are not effectively used for the development of broadband networks. Thus, increasing expenditure from EU funds for ICT may not reduce the existing digital divide. The key element is the support of the development of broadband networks in rural areas and by means of highly advanced technology. An effective national, sectoral and regional policy is also important [Fekó et al. 2011, pp. 152-160].

The use of EU funds is contingent upon the formulation of the strategy. However, as the studies show, the effectiveness of the regional strategies relating to ICT does not have a clear impact on the improvement of the access to the Internet and broadband networks. If appropriate expenditure from ERDF and other expenditure on the Internet infrastructure is dedicated to this objective, a positive effect will occur [Kleibrink et al. 2015]. The analysis of the use of funds from the cohesion policy from 2007-2013 also indicates that regions adopt varied regional digital strategies relating to Information Society. In less developed regions funds are allocated to strategic areas connected with Information Society and e-Services where the regions achieve relatively good results, thus strengthening the existing potential/strengths and not where they display specific weaknesses. Thus, they are not adapted to the requirements of the respective territories and they are not essentially linked to the specific territorial context [Reggi, Scicchitano 2014, pp. 530-538].

Conclusion and Future Work

It can be concluded that actions undertaken in the EU aim at reducing the digital divide and thus gradually help to implement the Digital Agenda. The scale of digital exclusion decreases in Poland, although changes are advancing slowly. In this context support is granted not only for improving the availability of the physical infrastructure, but also for upgrading skills in order to use these technologies competently and effectively. Poland received considerable support for this objective in the perspective 2007-2013 and the share of outlays from these EU backed funds for ICT was relatively high. Nevertheless, weaknesses were disclosed during the implementation of projects. In the new financing perspective ICT is financed, first of all, within the Operational Programme Digital Poland. A wide range of supported undertakings implies that they can prove to be effective. It is therefore essential to carry out the evaluation of the effectiveness of ICT projects supported within this cohesion policy in Poland, also in terms of their adaptation to the needs of the regions. The future research should focus on identification of the types of ICT projects which are needed in each of the Polish voivodeships and which could be supported from the external financial resources.

Bibliography

[Bednarczyk 2014] Bednarczyk D., Przeciwdziałanie cyfrowemu wykluczeniu (e-integracja) w Polsce, Biuletyn EBIB, nr 9 (154)/2014, Gromadzenie i zabezpieczanie danych cyfrowych (http://open.ebib.pl/ojs/index.php/ebib/article/view/297/469).

[Digital Econonom and Society Index] Digital Economy and Society Index 2016, Country Profile, Poland (https://ec.europa.eu/digital-single-market/en/scoreboard/poland).

[European Commission 2013 a] Commission Staff Working Document accompanying the document Report from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Cohesion policy: Strategic report 2013 on programme implementation 2007-2013, Brussel, 18.04.2013, SEC (2013) 129.

[European Commission 2013] Cohesion policy: Strategic report 2013 on programme implementation 2007-2013, Brussels, 18.04.2013, COM (2013) 210 final.

[European Commission 2015] Investing in jobs and growth – maximizing the contribution of European Structural and Investment Funds, Communication from the Commission. Brussels, 14.12.2015, COM (2015) 639 final.

[European Parliament 2015] Bridging the digital divide in the EU, European Parliament 2015, http://www.europarl.europa.eu/RegData/etudes/BRIE/2015/573884/EPRS_BRI%282015%29573884_EN.pdf

[Factsheet 2013] Factsheet: Information and Communication Technologies. Cohesion policy: Strategic Report 2013 (http://ec.europa.eu/regional_policy/sources/how/policy/doc/strategic_report/2013/factsheet2_ict_infrastructure_services.pdf).

[Fekó et al. 2011] Fekó A., Sass M., Nagy G., The role of broadband developments financed from EU structural funds in the enhancement of regional cohesion in the NMS-10, „Studies in Agricultural Economics" 113 (2011).

[Fundusze Europejskie 2014] Fundusze Europejskie 2014-2020 na badania, rozwój i cyfryzację, Warszawa 2014 r.

[GUS 2015],Społeczeństwo informacyjne w Polsce. Wyniki badań statystycznych z lat 2011-2015. GUS, Urząd Statystyczny w Szczecinie, Informacje i Opracowania Statystyczne, Warszawa 2015.

[Hargittai 2002] Hargittai E., Second-level digital divide: differences in people's online skills, „First Monday" 2002, vo. 7, no. 4-1.

[Kleibrink et al. 2015] Kleibrink A., Niehaves B., Palop P., Sörvik J., Thapa B. E.P., Regional ICT innovation in the European Union: Prioritization and performance (2008-2012), "J. Knowl Econ" 2015, 6: 320-333, DOI 10.1007/s13132-015-0240-0.

[Komisja Europejska 2010] Europa 2020. Strategia na rzecz inteligentnego i zrównoważonego rozwoju sprzyjającego włączeniu społecznemu, Komunikat Komisji, KOM 2010 (2020) wersja ostateczna, Bruksela 3.3. 2010.

[Komisja Europejska 2013] Sprawozdanie Komisji dla Parlamentu Europejskiego, Rady. Europejskiego Komitetu Ekonomiczno-Społecznego i Komitetu Regionów. Polityka spójności: sprawozdanie strategiczne 2013 dotyczące realizacji programów na lata 2007-2013, Bruksela, dnia 18.4.2013, COM (2013) 210 final; Commission Staff Working Document accompanying the document Report from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions.

[Krzyżanowska 2014] Kryżanowska A., Cyfrowa Polska 2020: Doświadczenia i wyzwania, Gdańsk, 12-13 czerwca 2014 (https://www.google.pl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwimjYb6pKjNAhXFHpoKHVZvAA4QFgggMAA&url=http%3A%2F%2F18.kmwi.pl%2Fprezentacje%2Fcategory%2F12-v-budzet-cyfrowej-polski.html%3Fdownload%3D28%3Aanna-krzyzanowska-cyfrowa-polska-2020-doswiadczenia-i-wyzwania&usg=AFQjCNHEYgeGd2F3926BZmLe6XdonhJCFg&cad=rja)

[Kwieciński 2014] Kwieciński J., Fundusze unijne na Cyfrową Polskę w starej i nowej perspektywie finansowej- ocena dotychczasowego wsparcia i wyzwania na przyszłość, Europejskie Centrum

Przedsiębiorczości, Gdańsk 2014. , prezentacja (https://www.google.pl/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&ved=0ahUKEwiPsb2KpajNAhUjAZoKHSSmD8QQFggdMAA&url=http%3A%2F%2F18.kmwi.pl%2Fprezentacje%2Fcategory%2F12-v-budzet-cyfrowej-polski.html%3Fdownload%3D30%3Ajerzy-kwiecinski-undusze-unijne-na-cyfrowa-polske-w-starej-i-nowej-perspektywie-finansowej-ocena-dotychczasowego-wsparcia-i-wyzwania-na-przyszlosc&usg=AFQjCNFIOHPizsgC-iebwJw8WrSgxj4-IQ&cad=rja)

[OECD 2001] Understanding the digital divide, OECD 2001.

[Polityka spojności] Polityka spójności w centrum uwagi w kontekście budowy jednolitego cyfrowego rynku Europy, „Panorama inforegio" 2015, nr 53.

[Program Operacyjny 2014] Program Operacyjny Polska Cyfrowa na lata 2014-2020, Wersja zaakceptowana decyzją Komisji Europejskiej z dnia 5 grudnia 2014 r.

[Reggi, Scicchitano 2014], Are EU regional digital strategies evidence-based? An analysis of the allocation of 2007-2013 structural funds, "Telecomunications Policy" 38 (2014),530-538.

[Regulation (EC) No. 1301/2013] Regulation (EC) No. 1301/2013 of the European Parliament and of the Council of 17 December 2013 on the European Regional Development Fund and on specific provisions concerning the Investment for growth and jobs goal and repealing Regulation (EC) No 1080/2006, OJ L 348/347 of 20th December 2013.

[Regulation (EC) No. 1303/2013] Regulation (EC) No. 1303/2013 of the European Parliament and of the Council of 17 December 2013 laying down common provisions on the European Regional Development Fund, the European Social Fund, the Cohesion Fund, the European Agricultural Fund for Rural Development and the European Maritime and Fisheries Fund and laying down general provisions on the European Regional Development Fund, the European Social Fund, the Cohesion Fund, the European Agricultural Fund for Rural Development and the European Maritime and Fisheries Fund and repealing Council Regulation (EC) No 1083/2006, OJ L 347/320 of 20th December 2013.

[Regulation (EU) No 1316/2013] Regulation (EU) No 1316/2013 of the European Parliament and of the Council of 11 December 2013 establishing the Connecting Europe Facility, amending Regulation (EU) No 913/2010 and repealing Regulations (EC) No 680/2007 and (EC) No 67/2010, OJ L 348/129 of 20th December 2013.

[Sieci szkieletowo-dystrybucyjne] Sieci szkieletowo-dystrybucyjne w perspektywie 2007-2013. Raport zamknięcia. Departament Telekomunikacji, Ministerstwo Cyfryzacji (https://mc.gov.pl/files/raport_zamkniecia.pdf).

[Sprawozdanie okresowe] Sprawozdanie okresowe w sprawie digitalizacji Europy (EDPR) 2016 https://ec.europa.eu/digital-single-market/en/scoreboard/poland

http://ec.europa.eu/regional_policy/pl/policy/how/stages-step-by-step/strategic-report/

http://ec.europa.eu/regional_policy/pl/policy/themes/ict/

http://ec.europa.eu/regional_policy/sources/docgener/informat/2014/fiche_ict_pl.pdf

https://cohesiondata.ec.europa.eu/funds/erdf

https://cohesiondata.ec.europa.eu/overview

https://cohesiondata.ec.europa.eu/themes/2

https://ec.europa.eu/digital-single-market/en/connecting-europe-facility

https://ec.europa.eu/digital-single-market/en/scoreboard/poland

https://www.polskacyfrowa.gov.pl/strony/o-programie/zasady/dla-kogo-jest-program/; https://www.polskacyfrowa.gov.pl/strony/o-programie/zasady/co-mozna-zrealizowac/

https://www.polskacyfrowa.gov.pl/strony/o-programie/zasady/finansowanie/

Authors' Information

Małgorzata Dziembała *– dr hab., Professor at the University of Economics in Katowice, 1 Maja 50, Poland*

e-mail: malgorzata.dziembala@ue.katowice.pl

Major Fields of Scientific Research*: International economics, EU cohesion policy, EU economic policies*

ON IMPORTANCE OF TOPOLOGY ON FUNCTIONAL ANNOTATION ON BIOLOGICAL PATHWAYS IN GENE EXPRESSION EXPERIMENTS

Arsen Arakelyan

Abstract: *Nowadays, gene set enrichment analysis (GSEA) is the method of choice for functional annotation in gene expression experiments. However, this approach demonstrates severe inaccuracy, when enrichment of biological pathways is considered. We hypothesize that lack of pathway topology information can be a reason for that. In this paper we evaluated the performance of GSEA and topology-based gene pathway deregulation assessment approach: Pathway Signal Flow (PSF). We demonstrated that PSF outperforms GSEA in pathway deregulation analysis and provided explanations for this observation.*

Keywords: *gene expression analysis, biological pathway, pathway topology, Gene Set Enrichment Analysis, Pathway Signal Flow.*

ACM Classification Keywords: *G.3 Probability and Statistics - Statistical computing, G.3 Probability and Statistics - Statistical software, G.3 Probability and Statistics - Nonparametric statistics*

Introduction

Genome-wide and whole genome transcriptome measurements are usually aimed at the assessment of global changes in gene expression landscape and identification of gene(s) that are differentially expressed under certain condition. Besides the various strategies and statistical approaches developed and used in global gene expression analysis the final output are lists of differentially genes or ranked gene lists. Next, these genes should be functionally annotated in order to better understand the underlying biological processes [Arakelyan et al, 2013].

Nowadays, gene set enrichment analysis (GSEA) is the method of choice for functional annotation in gene expression experiments [Hung, 2013; Subramanian, 2005]. It takes a list of pre-ranked gene list from the user dataset and compares it against known functional gene sets, such as Gene Ontology (GO) categories, microRNA targets, transcription factor binding targets, protein-protein interaction partners, etc. GSEA logic is based on the concept of "small, but concordant" changes [Subramanian, 2005]. While it works well for gene sets joined around molecular function and/or process (such as GO categories, miRNA and TF targets), it has a limited value for gene sets of protein-protein interaction

networks and biological pathways. The reason here is that genes in pathways highly connected to each other with regulatory or physical interactions of opposite directions (activation/inhibition or expression/repression). Thus, overexpression of one gene in pathway may lead to downregulation of others. This scenario has not implemented in GSEA, which searches for gene sets with orchestrated changes of expression towards over- or under-expression [Nersisyan et al., 2016].

Recently, several algorithms were proposed for analysis of pathway involvement based on gene expression data and pathway topology [Draghici et al., 2007; Rahnenführer et al., 2004]. Pathway signal flow (PSF) algorithm developed by us also belongs to this class of algorithms [Nersisyan et al., 2015; Nersisyan et al., 2016]. It allows for the assessment of global pathway involvement, but also identifying specifically deregulated sinks in multibranched pathways. Because PSF takes into account pathway topology and gene expression values; we believe that it should outperform GSEA, when a geneset represents a biological pathway.

In this paper we compared PSF and GSEA in analysis of pathway deregulation in number of diseases characterized by involvement of immune system response.

Data and methods

Data sourses

In this study we analyzed two microarray datasets obtained from Gene Expression Omnibus public repository [Barrett et al., 2011; Barrett and Edgar, 2006] that contain gene expression profiles in disorders characterized by involvement of inflammatory/immune response, such as psoriasis (GSE13355), chronic obstructive pulmonary disease (COPD, GSE42057), multiple sclerosis (MS, GSE13732) and cardioembolic ischemic stroke (IS, GSE58294). From these datasets only samples obtained from untreated patients were used in downstream analyses.

Pathway data used

For comparison purposes we have chosen 24 immune system and cell signaling pathways that were previously shown to be implicated in above mentioned diseased conditions (table 1).

Gene Set Enrichment Analysis

Gene set enrichment analysis (GSEA) was performed using classical algorithm implemented in GSEA Java application by broad institute using default parameters [Subramanian, 2005].

Pathway signal flow analysis

Pathway signal flow (PSF) algorithm evaluates the changes in activity of a given biological pathway depending on the pathway topology and relative gene expression [Arakelyan et al., 2013; Nersisyan et

al., 2015, 2016]. For comparison purposes we have chosen 24 immune system and cell signaling pathways that were previously shown to be implicated in above mentioned diseased conditions (table 1).

Table 1. Immune system and signaling pathways used in analyses.

Pathway names (source KEGG PATHWAY database http://www.kegg.jp/kegg/pathway.html)	
B cell receptor signaling pathway	Hedgehog signaling pathway
Calcium signaling pathway	HIF-1 signaling pathway
Chemokine signaling pathway	Jak-STAT signaling pathway
ErbB signaling pathway	MAPK signaling pathway
Fc epsilon RI signaling pathway	mTOR signaling pathway
Fc gamma R-mediated phagocytosis	Natural killer cell mediated cytotoxicity
FoxO signaling pathway	Notch signaling pathway
NOD-like receptor signaling pathway	PI3K-Akt signaling pathway
TNF signaling pathway	Rap1 signaling pathway
Ras signaling pathway	TGF-beta signaling pathway
RIG-I-like receptor signaling pathway	Toll-like receptor signaling pathway
T cell receptor signaling pathway	VEGF signaling pathway

In order to assess the changes in pathway activities dataset values were converted into natural scale and gene expression values were averaged for each class separately. Then expression fold change (FC) was calculated by division of gene expression in diseased condition to the corresponding values in

healthy controls. FC values for each gene were mapped to corresponding pathway nodes, and were averaged if a node contained more than one gene. After this step an input signal of unity was applied to the pathway source nodes. Then PSF values were calculated at the output nodes. PSF algorithm is calibrated in a way that gene expression of FC=1 at all nodes (normal gene expression) produces PSF=1 values. Values of PSF less than unity refer to pathway de-activation, while PSF>1 indicates pathway activation. Significance of PSF changes at output nodes was assessed by bootstrapping for 200 steps.

Results and discussion

KEGG pathway enrichment in psoriasis

Psoriasis is an immune-mediated, inflammatory and hyperproliferative disease of the skin and joints and conclusive evidence demonstrates that it has a genetic basis [Deng et al., 2016].GSE13355 data series contains log2 transformed gene expression profiles (Affymetrix HG 133 2 Plus platform) for 58 psoriatic patients and 64 normal healthy controls [Nair et al., 2009]. Here we have used GSEA and PSF to evaluate differential gene expression and biological pathway deregulations between skin biopsies taken from healthy patients and at lesion sites from patients. We expected to observe deregulations in immune system related pathways and signaling pathways regulating the immune response.

GSEA analysis identified 5 up-regulated pathways, while PSF detected pathway deregulations in all 24 pathways (Figure 1).

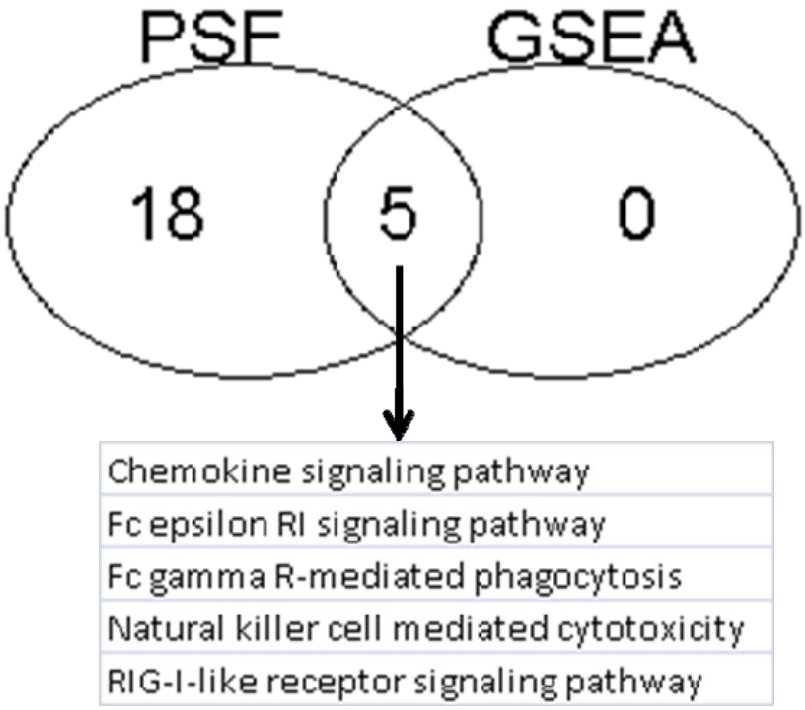


Figure 1. Performance of GSEA and PSF for the psoriasis dataset.

KEGG pathway enrichment in chronic obstructive pulmonary disease

Chronic obstructive pulmonary disease (COPD) is a leading global cause of mortality that is characterized by dysfunction of small airways with cellular inflammation and structural remodeling [Chung and Adcock, 2008]. The GSE42057 dataset contains normalized with robust multi-array averaging method [Irizarry et al., 2003] log2 transformed gene expression profiles of 94 subjects with varying severity of COPD and 45 healthy controls [Bahr et al., 2013]. Gene expression was measured using Affymetrix HG 133 2 Plus platform. GSEA analysis failed to detect pathways associated any pathway associated with inflammation and fibrosis, while PSF correctly identified deregulations in TGF-beta signaling pathway, VEGF signaling pathway as pathways involved in tissue remodeling during COPD and Fc-epsilon RI signaling pathway, T cell signaling pathway, TNF signaling pathway, RIG-I-like receptor signaling pathway and Toll-like receptor signaling pathway as involved in immune response.

KEGG pathway enrichment in multiple sclerosis

Multiple sclerosis (MS) is an organ-specific autoimmune disease caused by an inflammatory demyelinating insult in the central nervous system [Corvol et al., 2008].

We evaluated pathway deregulations base on GSE13732 dataset that contains gene expression profiles measured with Affymetrix HG 133 2 Plus platform in naïve CD4+ T cells of patients with MS before treatment and healthy subjects [Corvol et al., 2008]. GSEA identified TGF-beta signaling pathway as being disturbed in MS patients compared to control. In contrast, PSD detected 21 deregulated pathways (figure 2), from which B cell signaling and T cell signaling are being the most important ones in the pathogenesis of this disorder [Blauthet al., 2015; Holley et al., 2014].

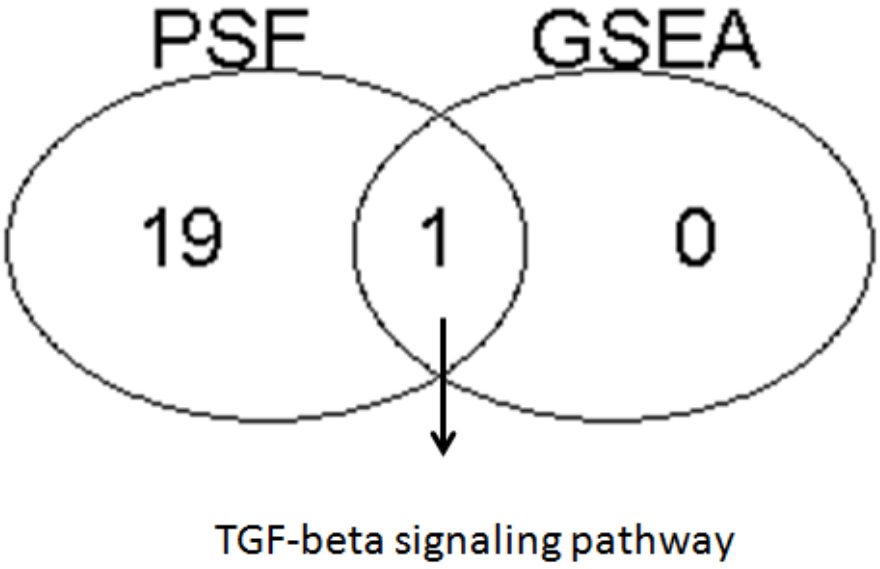


Figure 2. Performance of GSEA and PSF for the multiple sclerosis dataset.

KEGG pathway enrichment in ischemic strokes

Ischemic stroke (IS) accounts for 80% of all strokes, 70% of all acute cerebrovascular diseases, and is the most frequent acute neurological disorder. Its incidence is 14 cases/10000 population/year, and 25% of IS occur in working-age people. Only 1/3 of all stroke patients reach full social and professional reintegration, whereas the remainder die or invalid. The acute phase local and systemic inflammatory response in stroke has shown to play deleterious role along with activated aberrant apoptosis and activation of proinflammatory chemokine signaling [Di Napoli et al., 2006]. In this part of experiments we compared if GSEA and PSF are able to detect the deregulation of inflammation related pathways in stroke patients. We compared gene expression of 24 IS patients (blood collected after 3 hour of insult) with healthy subjects [Stamova et al., 2014]. Again, PSF correctly detects all pathways while GSEA detected only 4 pathways (figure 3).

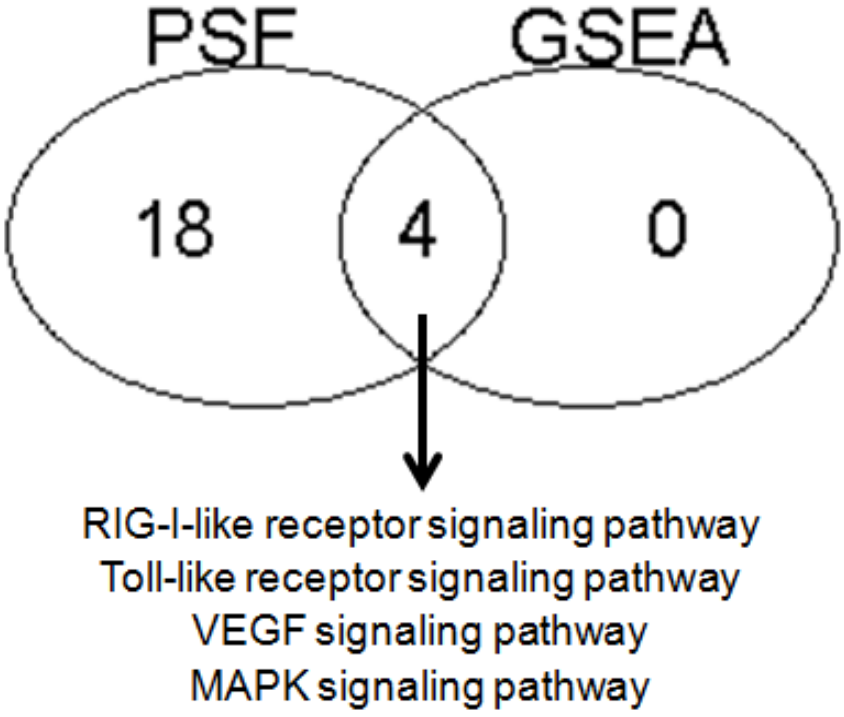


Figure 3. Performance of GSEA and PSF for the ischemic stroke dataset.

Importance of topology for accurate assessment of biological pathway deregulations

The main drawback of the GSEA is its inability to correctly assess the effects of interactions in the pathway. The statistic of GSEA is based on ranking of gene expression while the functional effect of altered gene is neglected. As an example we will analyze TGF signaling pathway deregulation in from multiple sclerosis dataset. The GSEA reported non-significant up-regulation of this pathway ($p = 0.17$, figure 4A). The lack of significance occurred due almost equal number up and down-regulated genes

(figure 4B). However, the closer inspection reveals that the majority of down-regulated genes possess inhibitory action on the on their interacting partners. Downregulation of inhibitory proteins will surely lead to overall up-regulation of this pathway which was actually detected by PSF (table 2). Moreover, PSF allow for evaluation of deregulation at each single branch in multi-branched pathways.

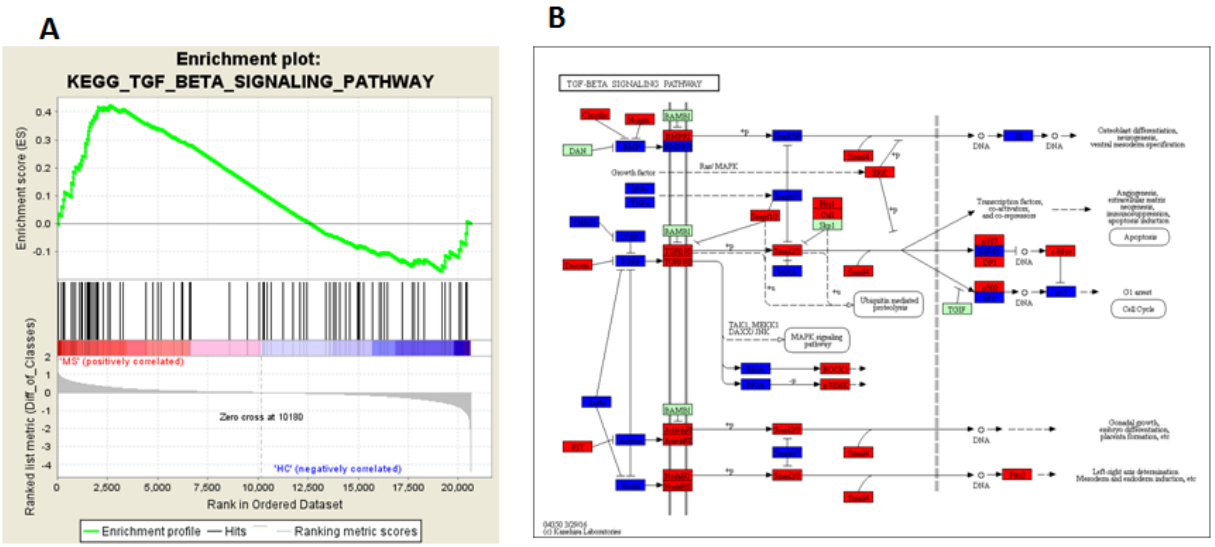


Figure 4. Gene set enrichment statistics score and positions of geneset members on the rank ordered list for TGF-beta signaling pathway (A) and visualization of up- and down-regulated genes in the pathway (B). Note that the most of down-regulated genes inhibit their interaction counterparts.

Table 2. PSF statistic for TGF-beta signaling pathway branches

Pathway	**Branch sink node**	**PSF statistic**	**p value**
TGF-beta signaling pathway	ROCK1	1.361922672	0.045
TGF-beta signaling pathway	RPS6KB1	1.467850941	0.006
TGF-beta signaling pathway	ID1	1.505816732	> 10E-6
TGF-beta signaling pathway	SMAD4	1.590104612	> 10E-6
TGF-beta signaling pathway	SMAD4	1.897146521	> 10E-6
TGF-beta signaling pathway	Cell cycle	0.599811704	> 10E-6
TGF-beta signaling pathway	Apoptosis	1.7290779	> 10E-6

Conclusion

Thus our study demonstrated that different strategy that GSEA should be employed if for analysis of biological pathway deregulations in gene expression studies. While GSEA will work well with sets of independent (i.e. non-interacting) genes, topology should be carefully considered and other methods, such as PSF should be used.

Bibliography

[Arakelyan et al, 2013] A.Arakelyan, L.Aslanyan, A.Boyajyan. High-throughput Gene Expression Analysis Concepts and Applications. In: Ed. iConcept Press Ltd, *Genomics II - Bacteria, Viruses and Metabolic Pathways,* 2013, pp. 71–95.

[Bahr et al, 2013] T.M.Bahr, G.J.Hughes, M.Armstrong, et al. Peripheral Blood Mononuclear Cell Gene Expression in Chronic Obstructive Pulmonary Disease. *American Journal of Respiratory Cell and Molecular Biology*, 2013, *49*(2), 316–323.

[Barrett et al, 2011] T.Barrett, D.B.Troup, S.E.Wilhite, et al. NCBI GEO: archive for functional genomics data sets--10 years on. *Nucleic Acids Research*, 2011, *39*(Database issue), D1005–1010.

[Barrett and Edgar, 2006] T.Barret, R.Edgar. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymology*, 2006, *411, 352-69*.

[Blauth et al, 2015] K.Blauth, G.P. Owens, J.L.Bennett. The Ins and Outs of B Cells in Multiple Sclerosis. *Frontiers in Immunology*, 2015, *6*, 565.

[Chung and Adcock, 2008] K.F.Chung, I.M.Adcock. Multifaceted mechanisms in COPD: inflammation, immunity, and tissue repair and destruction. *The European Respiratory Journal*, 2008, *31*(6), 1334–56.

[Corvol et al, 2008] J.-C.Corvol, D.Pelletier, R.G.Henry, et al. Abrogation of T cell quiescence characterizes patients at high risk for multiple sclerosis after the initial neurological event. *Proceedings of the National Academy of Sciences of the United States of America*, 2008, *105*(33), 11839–44.

[Deng et al, 2016] Y.Deng, C.Chang, Q.Lu. The Inflammatory Response in Psoriasis: a Comprehensive Review. *Clinical Reviews in Allergy & Immunology, 2016*, 1-13.

[Di Napoli et al, 2006] M.Di Napoli, A.Arakelyan, A.Boyajyan, A. (2006). The Acute Phase Inflammatory Response in Stroke: Systemic Inflammation and Neuroinflammation. In: Progress of Inflammation Research. Ed. J.A.Pitzer, New York: Nova Biomedical Books, 2006, 95–145

[Draghici et al, 2007] S.Draghici, P.Khatri, A.L.Tarca, et al. A systems biology approach for pathway level analysis. *Genome Research*, 2007, *17*(10), 1537–45.

[Holley et al, 2007] J.E.Holley, E.Bremer, A.C.Kendall, et al. CD20+inflammatory T-cells are present in blood and brain of multiple sclerosis patients and can be selectively targeted for apoptotic elimination. *Multiple Sclerosis and Related Disorders*, *3*(5), 2007, 650–658.

[Hung 2013] J.H.Hung. Gene set/pathway enrichment analysis. *Methods in Molecular Biology*, 2013, *939*, 201–213.

[Izzary et al, 2013] R.A.Irizarry, B.M.Bolstad, F.Collin. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Research*, 2003, *31*(4), e15.

[Nair et al, 2009] R.P.Nair, K.C.Duffin, C.Helms, et al. Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nature Genetics*, 2009, *41*(2), 199–204.

[Nersisyan et al, 2015] L.Nersisyan, G.Johnson, M.Riel-Mehan, et al. PSFC: a Pathway Signal Flow Calculator App for Cytoscape. *F1000Research,* 2015, *4*.

[Nersisyan et al, 2016] L.Nersisyan, H.Löffler-Wirth, A. Arakelyan, H.Binder. Gene Set- and Pathway-Centered Knowledge Discovery Assigns Transcriptional Activation Patterns in Brain, Blood, and Colon Cancer: *International Journal of Knowledge Discovery in Bioinformatics*, 2016, *4*(2), 46–69.

[Rahnenführer et al, 2004] J.Rahnenführer, F.S.Domingues, J.Maydt, T.Lengauer. Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical Applications in Genetics and Molecular Biology*, 2004, *3*, Article16.

[Stamova et al, 2014] B.Stamova, G.C.Jickling, B.P.Ander et al. Gene expression in peripheral immune cells following cardioembolic stroke is sexually dimorphic. *PloS One*, 2014, *9*(7), e102550.

[Subramanian et al, 2005] A.Subramanian, P.Tamayo, V.K.Mootha, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 2005, *102*(43), 15545–15550.

Author's Information

Arsen Arakelyan *–Bioinformatics group of the Institute of Molecular Biology NAS RA, Chair of Bioinformatics, Bioengineering and Molecular Biology, Russian-Armenian (Slavonic) University, 7 Hasratyan Str., 0014, Yerevan, Armenia, email: aarakelyan@sci.am*

Major Fields of Scientific Research*: Algorithm development foe gene expression analysis, gene network analysis and modeling*

SYSTEM ANALYSIS OF FORMATION OF EMERGENCY RESCUING EQUIPMENT SETS

Vasyl Kryshtal, Vitaliy Snytyuk

Abstract: *Aspects of four scientific paradigms are analyzed: system approach, system analysis, system design and system models, which are used in task solving of determining the optimal set of rescue equipment. Received results from the scientific basis for the definition of the set and its support for all stages of the life cycle.*

Keywords: *emergency rescue equipment, system approach, system model, system analysis, optimality.*

ACM Classification Keywords: H.4: Information systems application

Introduction

Rapid dynamics of modern world is one of the main causes of the amount and types of industrial and ecological catastrophes and accidents growths. Such threats and challenges to the mankind are the motives for the formation and development of safe environment for residing.

A task of emergency rescuing technique (ERT) complicating is analyzed according to system positions, which consist of system approach and system analysis. It is known that system approach is a scientific and practice methodology of difficult problems solving, and systematization is a determinative aspect of it, formalization and aim orientation [Tymchenko, 1991]. In turn, system analysis is a scientific and practice methodology of complicated systems investigation, that consists of such stages as system aim formulating, its elements basis clarification, structural peculiarities defining, the amount of internal system parameters and the amount of external functions of definition of system state and indicators of its efficiency forming, system effectiveness criterion identification [Zhurovskiy, 2005]. The process of completing itself is meant as dual to system projecting – a process of getting system project in the basis of system properties, system recourses and structures of vital cycle [Tymchenko, 1991]. The research is based on system model (SM), which is a row of elements with corresponding reflexing [Tymchenko, 1996] . These four constituents are represented constructively.

At first, we point out that the environment Ω has three constituents relatively to our task (fig. 1)

$$\Omega =< P, AS, NS >, \quad (1)$$

where P – human population, AS – artificial systems (manmade), NS – natural systems.

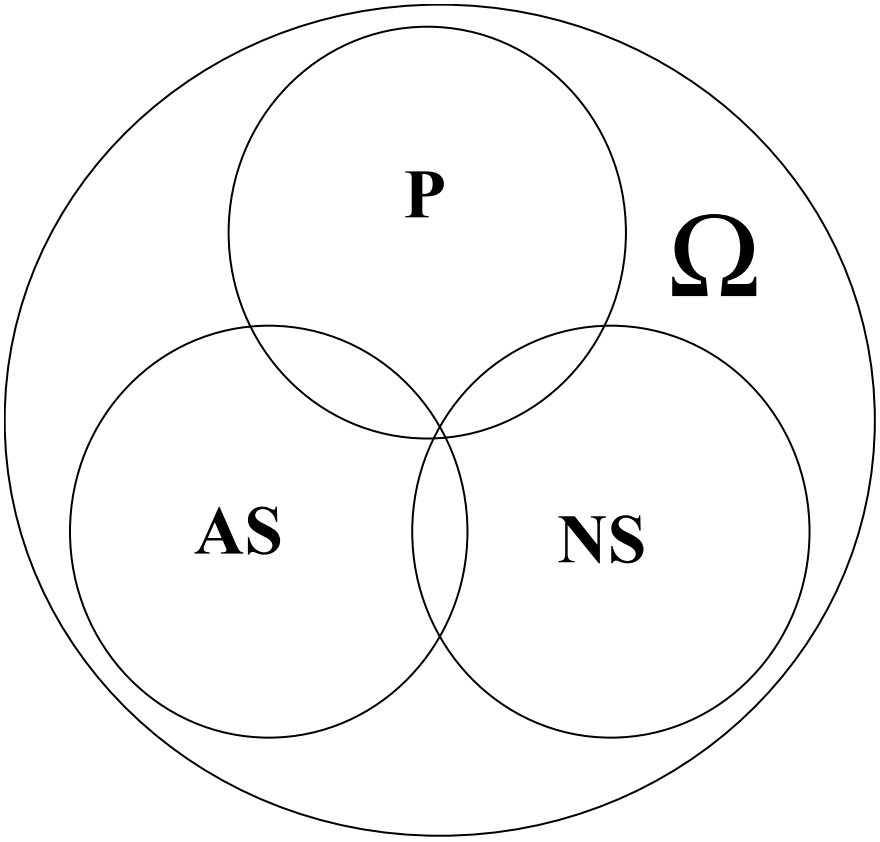


Figure 1. Costituents of environment

Processes of existing and functioning of denoted components is not always independent, in most cases they cross. When are fires, accidents, disasters, emergencies, etc. happen, people suffer, artificial and natural systems get harm in different quantities and proportions. In particular, chemical accidents affect natural systems and people, fires cause damage to people and artificial systems, emergencies hurt people, accidents and disasters of another nature may cause other variants of damage. Exceptions out of all these cases should be noted. It is important that their consequences are negative for people directly or indirectly.

Components of the system model

We regard an area $\Xi \subset \Omega$, which denote territorially determined administrative unit that is frequently a region or a city. We construct a SM of a process of ERT complicating in a region Ξ. It is known [Tymchenko, 1996] that, a SM is a row (fig. 2)

$$SM =< G, M, T, A >, \quad (2)$$

with reflexions:

$$H_1 : G \to M,\ H_2 : M \to T,\ H_3 : T \to A, \quad (3)$$

where G, M, T, A are sets of goals, models, methods, tools, in accordance.

As long as goal is a desired outcome, in our case it is people rescuing (G_{11}), material losses reducing (G_{12}), manmade and environmental disasters preventing (G_{13}), that is:

$$G =< G_0, G_{11}, G_{12}, G_{13}, ..., G_{1n}, G_{11}1, ... > \quad (4)$$

where G_0 is the main aim, which means minimizing of negative effects of accidents, fires etc. G_{ij}, G_{ijk} – goals of the main aim that form a hierarchical graph like structure. Reaching at least one of the G goals or set of goals, or all G goals as elementary structures "and-or" column depends on solution of the set of tasks

$$V =< V_1, V_2, ..., V_m, V_{11}, ... >, \quad (5)$$

where V_i, V_{ij} – verbal task formulations. It is known, a task that have formalized setting can solved, it means that it is necessary to build a reflection of a set $V \to M$. Composition of the set M, justified earlier in [1], includes F_1 – cost of an equipment set, F_2 – its functionality, F_3 – power, F_3 – reliability on macro level.

Then the appropriate tasks are:

$$F_1 \to \min,\ F_2 \to \max, F_3 \to \max, F_4 \to \max . \quad (6)$$

Including limitations on the dimensions of the equipment. Identification structural as well as parametric belongs to tasks and models of lower hierarchy levels. Its features are considered in [2]. Tasks as well as appropriate models also have graph like hierarchical structure.

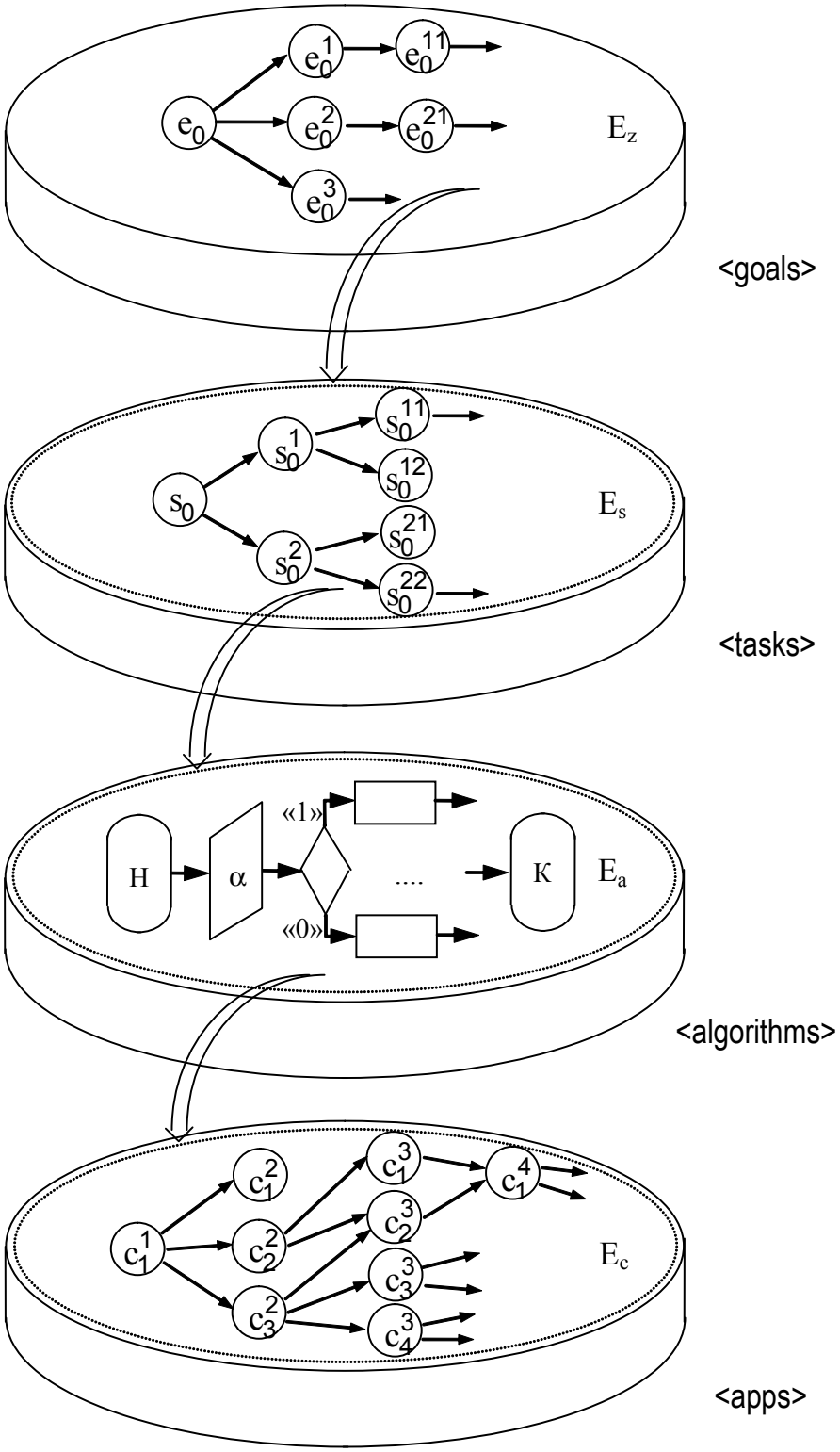


Fig. 2. System model

In order to obtain useful data and knowledge, information set V tasks should be solved. Since certain models of a not empty set (bank) correspond to the objectives, methods (W), also known as ways of task solving, should be used for their processing. It is obvious that a certain class of models and tasks can be solved with methods class. Thus, there is a reflection

$$Q : M \to W, \tag{7}$$

which is not mutually unequivocal and creates a need for optimization, which is to select the best method (determining of the most accurate result in certain resource constraints). One way to implement it is to use an approach with the use of ensemble of methods.

In particular, even and odd multiple linear regression, nonlinear regression, neural networks, polynomial Kolmogorov-Gabor and others belong to the set M. The set of methods includes the least squares method, Brandon method, group method of data handling, neural network learning methods (stochastic, direct, based on back propagation) etc. The majority of the methods mentioned above are too time-consuming, especially for large power source data set.

Whereas solving problems (6) requires significant computational cost, a stage of reflecting of a set of methods (T) on a set of algorithms (U) as formal structures for practical computer implementation is important. As well as methods allow their parametric optimization, algorithm that corresponds to one method can be built a lot of ways. If a set of parametric options of one method is N, the power of a set of appropriate algorithms is M, while $M >> N.$

A problem of execution (implementation) algorithm remains. For the problem of acquisition of ERT, despite the level of technology and computing technologies, regarding combinatorial complexity this problem is urgent. Therefore, the choice of tools (Z) is an important step in its solution and is implemented in reflection $U \to Z$.

In this matter, all components of the system model are denoted. Its construction at the macro level is to implement the set of reflections

$$G \to V \to M \to T \to U \to Z... \to M \to T... \tag{8}$$

Reflection chain corresponds to the implementation of the iterative process, because the requirements of a particular method cannot meet the goals or objectives of completing ERT. It is possible that the effectiveness of tools prevents receiving solution of the problem for existing baseline data, or resources, or the requirements of the process solution. System model (2) and (3) enables vision system features for solving the problem of acquisition of ERT. It is an integrating element of the range of emergency rescue tools, financial resources, organizational and administrative measures.

Levels of implementing a system approach to the problem of recruitment rescue equipment

The need to solve the problem of ERT completion stipulates a systematic study of retrospective features. In particular, Google produces about 105,000 links for key phrases "emergency rescue equipment" and "Acquisition of rescue equipment" in Ukrainian and Russian references about 455000. Analysis of the content of links pointing to research topics of acquisition of ERT in Cherkasy Institute of Fire Safety named after Chernobyl Heroes National University of Civil Defense of Ukraine. Other links let us get acquainted with different samples of ERT, which nomenclature is very broad and therefore there is the problem of forming a set of ERT. As far as this task had been solved only by experts from this institution, as evidenced by Google, we paid attention to the tasks that are relevant to certain aspects of our problem. In particular, the request "Task package" received 789,000 options. There is several dozen varieties of such problem solving. That is a two-dimensional packing, linear packaging, packaging according to weight, cost etc. The feature of such problems is their mono criterion nature [Smirnov, 1991]. Analysis and systematization of technologies for solving the problem of acquisition of ERT shows either little researches of evidence of relevant problems or neglect its multi criterion nature. This conclusion allows us to formalization of ERT completing tasks [Kryshtal, 2015]. Building a model, the purpose of each item as ERT and their set, should be considered.

Stages of system analysis and the process of ERT completing

The initial phase of system analysis is system aim formation; in this case, such aim of ERT set functioning is defined as a set of desired results. In the next stage we identify and order components base, forming row database table

$$< N, ID, Name, TS1, TS2, ..., TSn_N, f_1, f_2, ..., f_{n_N}, s1, s2, s3, \Pr >, \quad (9)$$

where N – is a serial number of a record in the database table, ID – is an identificator or abbreviated name of an equipment item, $Name$ – is a full name of the equipment, $TS1 - TSn_N$ – are equipment specifications, $f_1 - f_{n_N}$ – are functions that equipment performs, $s1, s2, s3$ – are its dimensions, *Pr* – is a price. The data in the table (9) is an initial information for task solving (6). The knowledge of technical specifications and functional features of equipment items is necessary to incorporate structural features while forming optimal set of equipment. In particular, if for two equipment items i's and j's we have $f_k^i = f_k^j \ \forall k \in \overline{1,n}$, the equipment set should be supplied with the item technical specifications value of which are not worse, and at least one is better, that is,

$$\forall k \ TSk^i \geq TSk^j \ \ i \ \ \exists l : TSl^i > TSl^j. \quad (10)$$

There are other situations:

– if the number of functions i's element is larger than of j's ones, we choose j's element in all other equal condition;

– if the number of functions i's element is larger than of j's ones, but at least one important technical specification is better with the meanings not worth of all others, the choice of the element is made either by using technologies of reducing of multicriteria task to monocriterion or of forming of a collective system of preferences using system analysis [Kryshtal, 2016].

The predominant aspect of the choice is different aspectness of items of equipment, that is, ideally there shouldn't be two elements i's and j's in one set, because $\exists k \in \overline{1,l}: f_k^i = f_k^j$. At the same time, the situation is desirable when $\cup f_i = \Phi$, where Φ – is the space of all possible functions that a set of ERT should perform. These features will determine the optimal structure of a potential set of ERT. Note that the price of a ERT set is crucial.

Thus, the initial problem can be reduced to this: to find a ERT set, that

$$\left|\cup f_i\right| \to \max \vee \left|\Phi \setminus \cup f_i\right| \to \min, \tag{11}$$

where $\left|\Theta\right|$ – is the power of a set amount (number of items) Θ, $\min$ or $\max$ are found according to possible sets of equipment subject to limitations on their dimensions. A task (11) can be specified taking into account technical characteristics values

$$\begin{gathered}(\left|\cup f_i\right| \to \max \ \& \sum_k TSk \to \max) \vee (\left|\Phi \setminus \cup f_i\right| \to \min \& \sum_k TSk \to \max), \text{ or, the same} \\ (\left|\cup f_i\right| \to \max \ \vee \ \left|\Phi \setminus \cup f_i\right| \to \min) \& \sum_k TSk \to \max.\end{gathered} \tag{12}$$

Let's consider in more details the system of functions performed by the hardware element of ERT. May f_k^j is the main function of the set of all possible features $\{f_1^j, f_2^j, ..., f_k^j, ..., f_n^j\}$ of j's element equipment set. In addition to the main features j's element can perform a number of utility functions, so our set of features can be ordered as follows:

$$\{f_0^j, f_1^j, ..., f_g^j, f_{g+1}^j, ..., f_{n-1}^j\}, \tag{13}$$

where f_0^j – is the main function of the equipment, f_i^j – are indirect functions, $i = \overline{1,g}$, f_i^j – are functions, that the element j's of ERT does not perform, $i = \overline{g+1, n-1}$. To determine the effectiveness of j's element of ERT we use an indicator of effectiveness j's element its functions performance $\{y_0(f_0^j), y_1(f_1^j), ..., y_g(f_g^j)\}$. Integral estimation of effectiveness is a quantitative characteristic effectiveness indicators,

$$E^j = E^j(y_0(f_0^j), y_1(f_1^j), ..., y_g(f_g^j)) \quad (14)$$

and can be identified using expert opinions analytically. For example, as an additive convolution

$$E^j = \sum_{i=0}^{g} \beta_i \cdot y_i(f_i^j), \ \forall j = \overline{1,m}, \quad (15)$$

where m – is a number of items of ERT, β_i – is i's effectiveness indicator weight performance. Note that the four evaluation criteria of ERT set: price, functionality, capacity and reliability are previously established. The same criteria can be considered for a particular item of equipment. It is also clear that there is a relationship between functionality (F_2) and efficient items of equipment, that is

$$F_2(e^j) = Q(E^j), \quad (16)$$

where e^j – is the piece of equipment. Identification Q requires additional research. Note that F_1, F_3, F_4 as an item of equipment is known. Price is determined by the supplier or manufacturer, reliability is known according to statistics, power – is a rated characteristic of the equipment element.

Rationally, while functionality determining of the equipment item the relevance of all of its functions should be considered. The more frequently emergent situation, where it was necessary to implement a function, arose, the more functional the equipment with this function was. This actualization brings to the construction of dependency that defines the functionality of ERT element

$$F_2(e^j) = Q(E^j) = Q(\sum_{j=0}^{g} \beta_i \cdot y_i(f_i^j)), \ \forall j = \overline{1,m}, \quad (17)$$

where $\beta_i = \gamma_i \cdot \delta_i$, and γ_i – is a weight coefficient of i's functions defined by manufacturer, δ_i – is a weight coefficient, indicating the urgency of i's functions implementing, $i = \overline{0,g}$. Common criterion of ERT set effectiveness is formed on the basis of the features denoted above.

Aspects of system designing

While solving the task of ERT completing, the characteristics that accompany the process of system design are to be considered. In this case financial resources, as well as carrier safety equipment – special vehicles can be included to system resources. System properties of ERT set include: openness – the ability to include new elements to a set or to exclude them, scalability – the ability to use a set of ERT on vehicles of various types, system – the ability to perform the broadest set of features, integrity – the property of collaboration of diverse items of equipment while performing a problem etc. Finally it should be noted the need to implement the requirements of the proposed operation programmable of ERT set, which will provide not only an effective work today but also in the future that may be associated with the need to replace used items to new and perspective.

Conclusion

Consistency in any task solving is a prerequisite for obtaining effective solution and optimization of all stages of its life cycle. System features of acquisition of ERT, investigated in the article, let us obtain the best (affordable) options in acquisition of ERT, which, in turn, will be targeted at strengthening the rescue people, minimization of material losses, the effects of natural and man-made disasters.

In the future, attention should be paid to technology of an integral objective function forming, methods of group result forming by selecting individual preferences, as far as these tasks are parametric and mostly subjective.

Bibliography

[Kryshtal, 2015] V. Kryshtal, A. Sergeev, V. Snytyuk. Determining the optimal variant complete rescue techniques using fuzzy conclusions. Vestnik NTU "KPI", Kyiv, № 49 (1158), 2015.

[Kryshtal, 2016] V. Kryshtal, V. Snytyuk. The evolutionary method of forming an optimal set of rescue equipment. Mathematical Machines and Systems, Kyiv, № 1, 2016.

[Smirnov, 1991] A.V. Smirnov. On the problem of packing in containers. SMS, Moskau, Vol. 46, Issue 4, 1991.

[Tymchenko, 1991] A. Tymchenko, A. Rodionov. Informatics principles of system designing objects of new technics. Naukova Dumka, Kiev, 1991.

[Tymchenko, 1996] A. Tymchenko, S. Aleshnikov, V. Snytyuk. Solutions existence investigation of system designing task of new technics objects. Preprint / NAS of Ukraine. In-t Cybernetics, Kiev, 1996.

[Zhurovskiy, 2005] M.Z. Zhurovskiy, N.D. Pankratova. System analysis. Problems, Methodology, Applications. Naukova Dumka, Kiev, 2005.

Authors' Information

Vasyl Kryshtal *– Senior Lecturer, Cherkassy Institute of Fire Safety named after Heroes of Chornobyl of National University of Civil Defense of Ukraine, 18000, Cherkasy, Ukraine, 8 Onoprienko str.; e-mail: kryshtal.v@ukr.net*

Major Fields of Scientific Research: Fire Safety, Evolutionary Modeling

Vitalii Snytiuk *– Head of Department, Taras Shevchenko National University of Kyiv, 03022, 81a Lomonosov str., Kyiv, Ukraine; e-mail: snytyuk@gmail.com*

Major Fields of Scientific Research: decision making under uncertainty, Evolutionary modeling

TABLE OF CONTENTS

Term Matching Approaches in E-Commerce

Mariam Gawich, Marco Alfonse, Mostafa Aref, Xavier Tannier, Abdel-Badeeh M. Salem 3

Proposed Mobile System for Bone Visualization Using Gpu Raycasting Volume Rendering Technique

Yassmin Abdallah, Abdel-Badeeh M.Salem 12

Computer Devices and Information Technologies for Medicine

Volodymyr Romanov, Ozar Mintser, Igor Galelyuka, Viktor Degtjaruk, Viktor Chernetsky, Illya Chaikovsky 21

Improving Performance in Neural Networks Using Feature Selection Based on Random Forests: Application to Automated Medical Diagnosis

Smaranda Belciug, Florin Gorunescu 33

Comparative Analysis of Various Filters for Noise Reduction in MRI Abdominal Images

Antoniya Mihailova, Veska Georgieva 47

Investment in ICT in Poland from the EU Funds - Opportunities and Challenges

Małgorzata Dziembała 67

On Importance of Topology on Functional Annotation on Biological Pathways in Gene Expression Experiments

Arsen Arakelyan 82

System Analysis of Formation of Emergency Rescuing Equipment Sets

Vasyl Kryshtal, Vitaliy Snytyuk 91

Table of contents 100