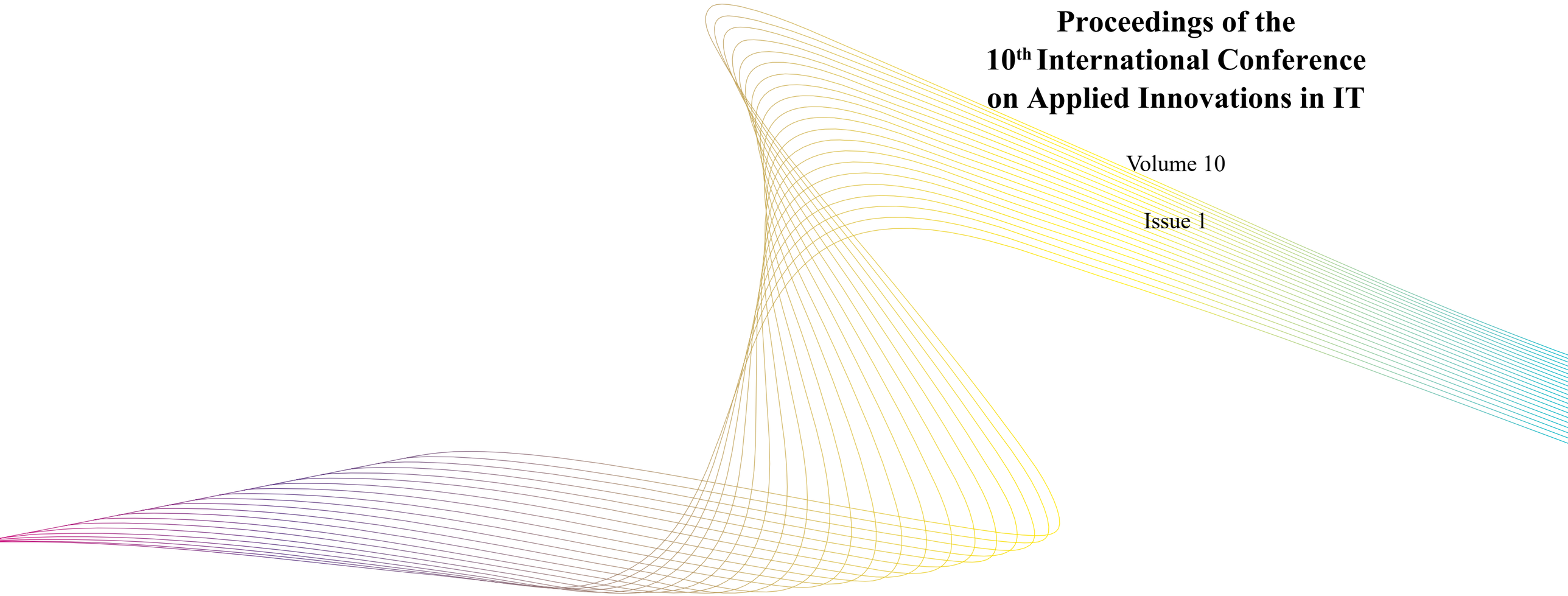
by Anhalt University of Applied Sciences and Perm National Research Polytechnic University

Proceedings of the 10th International Conference on Applied Innovations in IT

Volume 10

Issue 1



EDITION
Hochschule Anhalt

Proceedings of the 10th International Conference on Applied Innovations in IT

Volume 10 | Issue 1

Koethen, Germany
9 March 2022

Editors:

Prof. Dr. Eduard Siemens* (editor in chief)
Assoc. Prof. Dr. Leonid Mylnikov**

(*Anhalt University of Applied Sciences,
** Perm National Research Polytechnic University)

This volume contains publications of the International Conference on Applied Innovations in IT (ICAIIT), which took place in Koethen March 9th 2022. The conference is devoted to problems of applied research in the fields of automation and communication technologies. The research results can be of interest for researchers and development engineers who deal with theoretical base and the application of the knowledge in the respective areas.

ISBN: 978-3-96057-140-7 (Online)
ISSN: 2199-8876

Content

Cost-Effective High Performance Distributed GPU Cluster for Deep Learning Tasks

Kirill Karpov, Dmitry Kachan, Maksim Iushchenko, Ivan Luzianin and Eduard Siemens
Department of Electrical, Mechanical and Industrial Engineering, Anhalt University of Applied Sciences,
55 Bernburger Str., Köthen, Germany
{*kirill.karpov, dmitry.kachan, maksim.iushchenko, ivan.luzianin, eduard.siemens*}*@hs-anhalt.de*

Keywords: Tensor Flow, DNN Training, Performance, Horovod, HPC, High-Performance Computing.

Abstract: The expenses on computational resources for modern Deep Learning computing can be extremely large. However, most of them are spent on the chassis and not on the GPU units themselves. Since modern mass market graphic cards are usually chipper and have huge performance for video games, it was hypothesized, that a low cost cluster, made of several graphic cards, can reach the same performance for computational tasks as ready-made enterprise GPU-server with significantly lower price. The concept of distributed GPU cluster based on mass market GPU units is presented in the article. During the experiments, performance of a cluster with two mass market GPU units was compared with performance of enterprise GPU-server with 8 GPU-units on the Deep Learning bench mark. The results shows benefits and limitations of use proposed distributed cluster. It describes cases, when this solution is up to 7 times more effective than enterprise one in terms of cost savings for chassis itself as well as for, additional equipment and maintenance.

1 INTRODUCTION

The rapid development of GPU-accelerated computing systems makes Deep Learning (DL) algorithms the most powerful Machine Learning solutions. At the same time, DL models require complex and expensive enterprise-level High-Performance Computing (HPC) hardware for training. They require additional equipment for operating such as racks, air conditioning, power supply systems, etc. Moreover, these solutions require high-qualified personnel. Finally, enterprise GPU units are generally more expensive than usual servers.

In opposite, mass-market graphic cards do not require special conditions and maintenance to operate. Modern ones are good enough to perform DL computational tasks. Therefore, potentially, if several mass-market GPU modules are incorporated together in a single cluster, then the overall cluster's performance will be the same as for the enterprise solution but with a lower price. This may be useful if there is a need to create a GPU-accelerated computing cluster that is simple to use and does not require special conditions to operate.

It is possible to create a computing cluster of multiple single nodes using technologies like RDMA [1], MPI [2], and NCCL [3]. Furthermore, the frameworks like Horovod [4, 5] provide a simple interface to their technologies for deep learning tasks.

This work aims to confirm the hypothesis that it is possible to create a cost-effective cluster of multiple mass-market GPU nodes with a higher performance than an enterprise solution has.

It is necessary to perform the following tasks to reach the goal:

- to prepare the testbeds for mass-market and enterprise solutions;
- to provide performance measurements on both testbeds;
- to analyze and conclude the results of the performance measurements.

The remainder of this paper is structured as follows: Section 2 describes the hardware and software equipment of the current work. Section 3 presents the results of the performance measurements of experiments. Finally, Section 4 discusses the results, followed by the conclusion in Section 5.

2 EXPERIMENTAL SETUP

The experiments were carried out on two testbeds: Single-Chassis GPU Infrastructure and PC Cluster-

Based GPU Infrastructure with the same software components.

2.1 Hardware Setup

2.1.1 Single-Chassis GPU Infrastructure

The enterprise-level solution for AI and Deep Learning data centers is represented by the Supermicro SYS-4029GP-TVRT server that has eight Tesla V100 GPUs. The system is shown in the Figure 1. In the current work, this testbed is named as **Single-Chassis GPU Infrastructure**. The server supports Nvidia's Volta V100 SXM2 form factor GPUs that benefit from Nvidia's NVLINK architecture to deliver GPU to CPU data rates of up to 300GB/s compared to using PCIe based GPUs which offer only up to 32GB/s data rates. GPUDirect Remote Direct Memory Access (RDMA) technology allows direct peer-to-peer (P2P) data exchanges between other devices in the network, bypassing the CPU and reducing GPU to GPU latency.

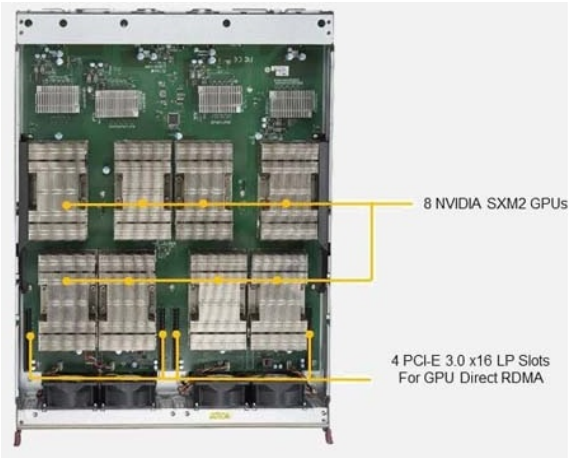


Figure 1: Supermicro SYS-4029GP-TVRT.

The detailed configuration of the Single Chassis GPU Infrastructure testbed is provided in the Table 1.

Table 1: Single-Chassis GPU Infrastructure Configuration.

System Type	Supermicro SYS-4029GP-TVRT
CPU	2x Inel Xeon Platinum 8268, 24-Core, 2.9 GHz, HT, 35,75MB Cache
Motherboard	Dual Socket P (LGA 3647) intel Xeon Scalable, X11DGO-T
Chipset	Intel C621, UPI upt to 10.4 GT/sec
GPU	8x Nvidia Tesla V100, 32 GB CoWoS HBM2, SXM2 - NVLink 2.0, CUDA Cores: 5120, Core Clock: 1455 MHz, FP32 Computing Performance: 15.0 TF, Memory Bandwidth: 900 GB/s, Memory Type: 4096-bit 16 GB HBM2, GPU: GV100 (Volta),
RAM	24x 64 GB DDR4, PC2933, ECC registered
Price	112,000 € (2020)

2.1.2 PC-Cluster Based GPU Infrastructure

The PC-based testbed consists of two nodes. Each node is composed of a small-form-factor computer Intel NUCs Ghost Canyon 9 Extreame (Figure 2) with Intel-i7 CPU and an external GPU module Aorus gaming Box RTX 3090 (Figure 3) connected to the computer via Thunderbolt 3 (TB3).

Figure 2: Intel Ghost Canyon 9 Extreme PC.

Since a NUC 9 computer has an additional PCIe3 16x slot, it is equipped with Mellanox ConnectX-5 100G NIC for high-bandwidth and low-latency network communication that is a crucial part of distributed computing.

Figure 3: Aorus Gaming Box RTX 3090 eGPU.

The scheme of PC Cluster Based GPU Infrastructure setup is shown in the Figure 4. The NUC computers, each equipped with Mellanox 100G network adapter (NIC), connect to their respective external GPU modules via TB3 at the speed up to 40Gbps. In this work the NUCs' network adapters are connected via 100G network switch. However it is possible to connect them directly in point-to-point manner.

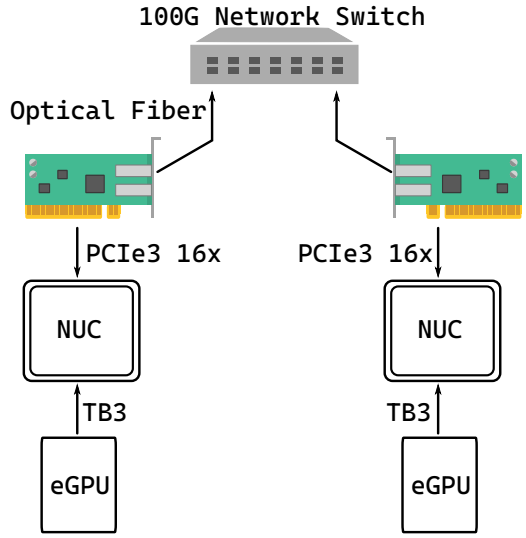


Figure 4: PC-Cluster Based GPU Infrastructure Setup.

The detailed configuration of PC-Cluster Based GPU Infrastructure is provided in the Table 2.

Table 2: PC-Cluster Based GPU Infrastructure Configuration.

PC Model	NUC 9 Extreme Ghost Canyon [6, 7]
CPU	i7-9750H
Motherboard	Dual Socket P (LGA 3647) intel Xeon Scalable, X11DGO-T
Chipset	Intel C621, UPI upt to 10.4 GT/sec
eGPU	Aorus Gaming Box RTX 3090
RAM	16 GB DDR4
NIC	Mellanox ConnectX-5, 2x100GbE, QSFP28 [8]
100G Switch	Extreme Networks x870-32c
Price for a node	3,900 € (2021)
Price for a switch	27,000 € (2020)

2.2 Software Setup

All tests were performed with the same software set for both PC Cluster-Based GPU Infrastructure and Single Chassis GPU Infrastructure testbeds.

The computers are equipped with Ubuntu 20.04 operating system with 5.4.0.97-lowlatency kernel. Low latency kernel contains optimizations, such as Preempt-RT, to achieve the lowest possible latency for applications.

Since the network is the bottleneck that significantly affects the scaling factor [9], the operating system's TCP stack was tuned to achieve the bandwidth closest to 100 Gbps with the following configuration:

```
net.core.rmem_max = 16777216
net.core.wmem_max = 16777216
net.core.rmem_default = 16777216
net.core.wmem_default = 16777216
net.ipv4.tcp_rmem = 4096 87380 16777216
net.ipv4.tcp_wmem = 4096 87380 16777216
net.ipv4.tcp_mem = 1638400 1638400 1638400
net.ipv4.tcp_sack = 0
net.ipv4.tcp_dsack = 0
net.ipv4.tcp_fack = 0
net.ipv4.tcp_slow_start_after_idle = 0
jumbo_frames=yes (default no)
```

In this work, Horovod [4] is used as a distributed learning framework. Horovod supports the Remote Direct Memory Access (RDMA) technology [1] improves its efficiency. RDMA provides access to the memory from one computer to the memory of another computer without involving either computer's operating system. This technology enables high-throughput and low-latency networking with low CPU utilization. The Mellanox ConnectX-5 NICs of the NUCs were configured with `mlnx-en-5.5-1.0.3.2` driver and MLNX OFED version 4.9-4.1.7.0. These NICs make use of RDMA over Converged Ethernet (RoCE) - a network protocol that enables remote direct memory access (RDMA) over Ethernet.

The network performance was tested using qperf [10] tool with the following result for TCP:

```
# qperf nuc2.lab tcp_bw tcp_lat
tcp_bw:
    bw  =  6.6 GB/sec (52.8 Gbps)
tcp_lat:
    latency  =  8.6 us
```

And for RDMA over Converged Ethernet:

```
# qperf -cm1 nuc2.lab rc_bw rc_lat
rc_bw:
    bw  =  12.2 GB/sec (97.6 Gbps)
rc_lat:
    latency  =  5.3 us
```

This network configuration meets the Horovod [5] requirements. Horovod is a distributed deep learning training framework for TensorFlow, Keras, PyTorch, and Apache MXNet. The goal of Horovod is to make distributed deep learning fast and easy to use. The Horovod version 0.22.1 deployed on both NUCs as a Docker [11] container provided by developers. It is configured with the following components: TensorFlow [12], PyTorch, MXNet, MPI, Gloo, NCCL, and CUDA 11.0.

Tensorflow 2.4 [12] is the framework to help develop and run DL-based solutions and is used to estimate the performance of distributed learning. TensorFlow is one of the most popular machine learning frameworks, and it has been used in a wide variety of applications and to conduct AI research.

3 EXPERIMENTAL RESULTS

The performance measurement experiments were provided using Horovod-adopted Tesnsorflow2 syn-

thetic benchmark with ResNet101 model and batch size 64. The estimated metric is Images/sec provided by the Tensorflow2 benchmark. Each run was performed 10 times for each number of GPUs.

Figure 5 shows the performance of the Single-Chassis GPU setup with 8 Tesla V100 GPUs.

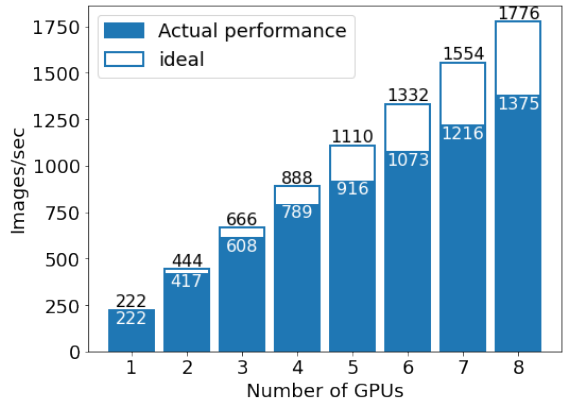


Figure 5: Multi-GPU scaling performance of Single Chassis GPU infrastructure using TensorFlow. The ideal value is the value of a single GPU performance multiplied by the corresponding number of GPUs.

As can be seen, the performance of scaling efficiency drops with each additional GPU. Whereas two parallel GPUs lose only 5 % of the ideal performance, the eight GPUs lose 22 %.

Figure 6 shows the performance of PC Cluster-Based GPU setup with two RTX 3090 GPUs.

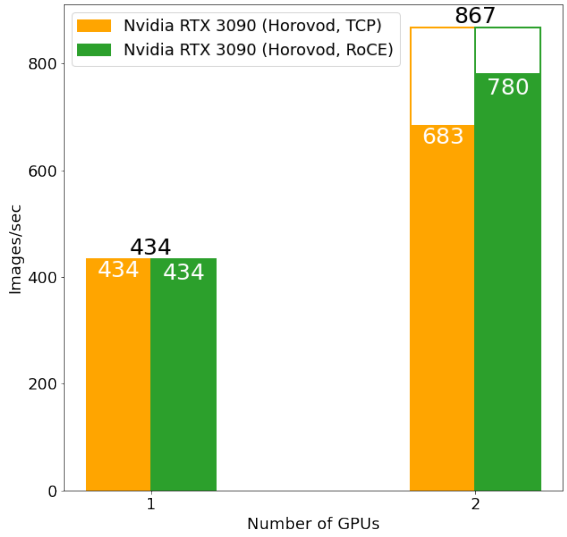


Figure 6: A comparison of the images processed per second by Horovod framework over plain 100GbE TCP and 100GbE RoCE-capable networking.

In this case, the drop of performance scaling efficiency is significantly higher than in the previous experiment. Two GPUs communicated via TCP stack lose 21% of ideal performance. RDMA brings slightly better results, the difference between the ideal value and the actual one is 10%. Nevertheless, the performance of two GPUs in absolute values in the PC Cluster-Based GPU testbed is comparable to the four GPUs in the Single Chassis testbed.

4 THE DISCUSSION OF THE RESULTS

There is a significant difference between the test scenarios in the number of parallel nodes, which makes the comparison quite challenging. Fortunately, there is a method that estimates the performance of scalable computing systems. The Universal Scalability Law [13] (USL) is an extension of Amdahl's Law that corrects the performance concerning communication latency between nodes. The USL is given by the (1).

$$S(n) = \frac{\gamma n}{1+\alpha(n-1)+\beta(n^2-n)} \qquad (1)$$

The coefficient γ represents the slope associated with linear-rising scalability in the case of ideal parallelism, α defines the serial coefficient, and β represents additional delays (see Figure 7).

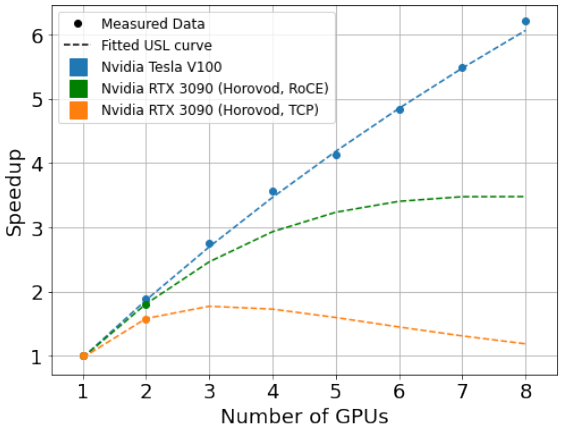


Figure 7: The result of fitting the measured data to the USL formula.

Using the Universal Scalability Law (USL) fit technique described in [14] the following parameters for the equation 1 were obtained: $\alpha = 0.04$, $\beta = 0$, $\gamma = 0.97$.

In order to evaluate the performance of the PC-Cluster Based GPU Infrastructure testbed the only parameter of additional delay was variated to fit the first

two points. The result of estimation is shown in the Figure 8.

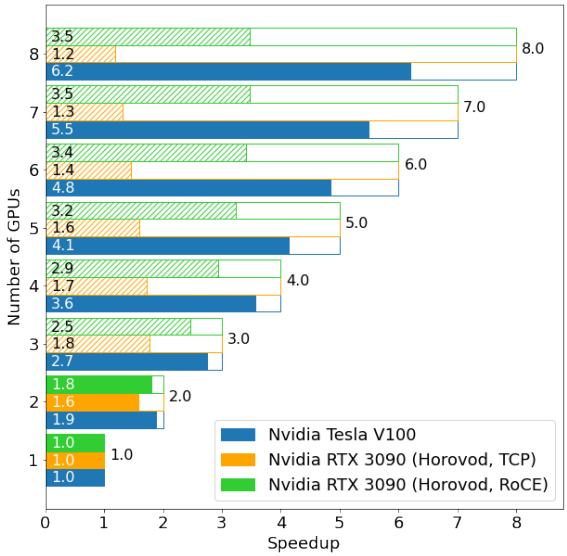


Figure 8: The dependency of speedup coefficient o n the number of GPUs. The colored frame is an ideal speedup value. The stroked column is an extrapolated value of the speedup coefficient.

With known speedup coefficients it is possible to estimate the absolute performance of the PC-based setup. The result of estimation is shown in the Figure 9.

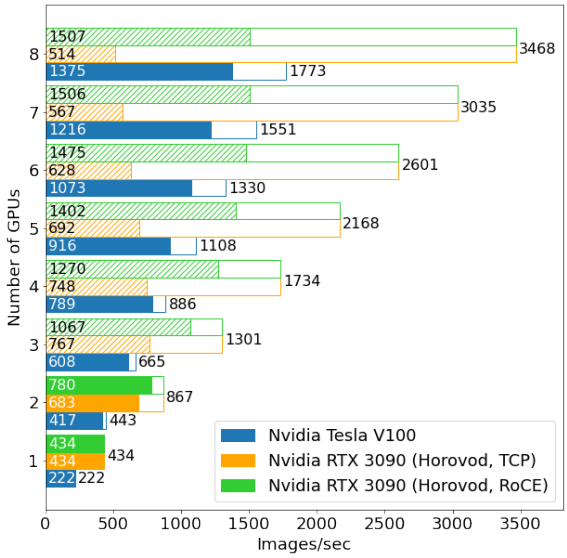


Figure 9: The dependency plot of learning performance on the number of GPUs. The colored frame is an ideal performance value. The stroked column is an extrapolated value of the performance.

Relying on the estimation it is possible to conclude that PC-Cluster Based GPU Infrastructure testbed consisting of five nodes orchestrated by Horovod and communicated via RoCE protocol can surpass the Single Chassis setup in terms of the number of processed images/sec.

5 CONCLUSIONS

The experimental results clearly show that the distributed setup with only two GPUs provides half of the performance of the modern and professional solution that costs ca. 14 times higher. However, the scalability of PC-based solutions is limited. The estimation shows that the performance of PC Cluster-Based solution with 5 GPUs can surpass Single-Chassis, though, the following scaling is unreasonable. In this case, the cost-efficiency drops significantly, since the setup with more than two GPUs requires a 100G network switch. Here the estimated cost win is supposed to be 2.5 times.

Pros and cons of PC-based Cluster infrastructure:

+ The results show that the PC Cluster-Based solution is very cost-efficient within the defined limits.
+ The system is easily upgradable.
+ Flexibility is one of the main advantages of this setup. The PC-based setup can be assembled any time in many ways. It also might be distributed from a location perspective. Since the setup is as compact as a usual desktop, the nodes can be distributed across working tables, offices, or even building, if the optical network infrastructure allows it.
+ This is a multi-purpose solution. Since each node is basically a usual PC, with the corresponding periphery interfaces (USBs, HDMI, Ethernet, etc), it might be used as an office workstation during the day, and as a node of the cluster the rest of the time.
+ Each node or even element of the node (PC, NIC, eGPU) can be easily changed in case of malfunctioning. Also, in the event of an overcurrent in a nodal element, this highly likely will not lead to a malfunction of the node or the entire system.
- The scaling factor drops with the little number of nodes. As it was shown, in this particular case, after five nodes the scaling is meaningless.
- The system is limited in distributed computational tasks. It is necessary to use special frameworks which are able to distribute desired tasks across the nodes.

REFERENCES

[1] "RDMA Aware Networks Programming User Manual," Mellanox Technologies, Tech. Rep. Rev 1.7, 2015. [Online]. Available: https://network.nvidia.com/related-docs/prod_software/RDMA_Aware_Programming_user_manual.pdf

[2] W. Gropp, E. Lusk, and A. Skjellum, Using MPI: Portable Parallel Programming with the Message Passing Interface. The MIT Press, 11 1999. [Online]. Available: https://doi.org/10.7551/mitpress/7056.001.0001

[3] "Nvidia collective communication library (nccl) documentation," Nvidia Corporation, Specification, 2020. [Online]. Available: https://docs.nvidia.com/deeplearning/sdk/nccl-developer-guide/docs/index.html

[4] A. Sergeev and M. Del Balso, "Horovod: fast and easy distributed deep learning in tensorflow," arXiv preprint arXiv:1802.05799, 2018.

[5] Horovod, "Horovod," https://github.com/horovod/horovod, 2017.

[6] "Intel NUC 9 Extreme/Pro Kit, Techincal Product Specification," Intel Corporation, Specification 1, Dec. 2019. [Online]. Available: https://simplynuc.com/wp-content/uploads/2020/01/NUC9QN TechProdSpec.pdf

[7] "Intel NUC 9 Kit, User Guide," Intel Corporation, Specification 1, May 2020. [Online]. Available: https://www.intel.com/content/dam/support/us/en/documents/intel-nuc/nuc-kits/NUC9xyQNX UserGuide.pdf

[8] "ConnectX-5 EN Card," Mellanox Technologies, Tech. Rep., 2020. [Online]. Available: https://www.mellanox.com/files/doc-2020/pb-connectx-5-en-card.pdf

[9] Z. Zhang, C. Chang, H. Lin, Y. Wang, R. Arora, and X. Jin, "Is network the bottleneck of distributed training?" in Proceedings of the Workshop on Network Meets AI & ML, 2020, pp. 8–13.

[10] J. George, "Qperf 0.4.11 (1)-linux man page," Qperf (1): Measure RDMA/IP Performance.

[11] D. Merkel et al., "Docker: lightweight linux containers for consistent development and deployment," Linux journal, vol. 2014, no. 239, p. 2, 2014.

[12] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," arXiv preprint arXiv:1603.04467, 2016.

[13] N. J. Gunther, Guerrilla Capacity Planning: A Tactical Approach to Planning for Highly Scalable Applications and Services, 1st ed. Springer Publishing Company, 2010.

[14] F. Alkhoury, D. Wegener, K.-H. Sylla, and M. Mock, "Communication efficient distributed learning of neu-ral networks in big data environments using spark," in 2021 IEEE International Conference on Big Data (Big Data), 2021, pp. 3871–3877.

Estimation of the Available Bandwidth of the Cellular Network Connection with Android API

Maxim Gering, Nikolai Mareev and Eduard Siemens
Future Internet Lab Anhalt, Anhalt University of Applied Sciences, 57 Bernburger Str., Köthen, Germany
{maksim.gering, nikolai.mareev, eduard.siemens}@hs-anhalt.de

Keywords: TCP, IP, Internet, Wireless Network, LTE, Available Bandwidth, Android API, BBR.

Abstract: The speed at which mobile and wireless networks are currently developed continues to be astounding. The rate of data reception and transmission over a wireless network becomes a major factor for consumers and mobile phone service providers, as service providers compete with each other for leading positions in the quality of mobile communications provided. This paper describes the problem of available bandwidth estimation of the LTE network in dense urban environments under the TCP BBR Congestion Control protocol. For this, the open-source HTWK Signal Harvester app from the Leipzig University of Applied Sciences has been modified to estimate the available bandwidth of a connected 4G LTE network. The test results presented in the article show that the available bandwidth gathered from Android 11 API metrics are close to the results obtained by iPerf3, a stand alone utility for active measuring of maximum achievable bandwidth on IP networks, widely used in PC and server computers.

1 INTRODUCTION

Provision of a maximum possible available bandwidth (AvB) of a wireless network is currently a key issue for mobile communication operators. Currently, several tools are available to analyze available bandwidth on wired and mixed networks, already embedded in devices and allowing observation of changes in internet traffic. However, these applications have a structure hidden from the user, making it difficult to analyze and evaluate the factors affecting the quality of the connection. Several factors are known to affect wireless connection speeds, e.g. different weather conditions, technical (such as equipment workload), and landscape.

Examples of such weather conditions are a forested area [1] or rainy weather [2, 3]. Landscape [4] has also a major influence on the data transport speed, as the radio signal can pass through, be reflected, and be absorbed by obstacles. Also, there is the area of responsibility of the provider. Which includes such parameters as workload and channel capacity, quality of the equipment used, reliability and quality of the cable from the server to the customer's home, the network equipment that is leased to the customer. These are only some of the factors affecting signal quality. The main focus of this paper is on estimation of available bandwidth of the wireless network based on parameters obtained by the Android application on a mobile device.

To estimate the available bandwidth in LTE communication some specific transport parameters have been introduced such as BW, RB, RE, TBS, MCS, CQI, RSSNR. Hereby, BW - frequency range in a given band used for signal transmission. There and below, the concept of available bandwidth will be used to refer to data rates, and the concept of channel bandwidth will be used to define the frequency range in a given frequency band. Resource Block (RB) is a block consisting of a channel resource and includes 12 adjacent subcarriers occupying the 180 kHz band. During transmission, 12 adjacent sub-carriers are multiplied by 7 cyclic prefixes to form 84 RE resource elements (for the 7OFDM cyclic prefix case which is commonly used in LTE) which in turn forms one resource block. Adaptive Modulation and Coding (AMC) is used to increase the network capacity or downlink data rates. AMC stands for different Modulation and Coding Schemes (MCS). MCS defines different types of modulation, such as quadrature phase shift keying (QPSK), quadrature amplitude modulation (QAM). CQI - Channel Quality Index. The CQI channel-state report effect to change the modulation scheme in the subframe. This parameter may vary from manufacturer to manufacturer and is not consistent with each other.

RSSNR - Signal to Noise Ratio is used to determine the RSSNR-CQI relationship.

The main objective of this paper consists of two tasks. The first task is to propose an adaptive available bandwidth estimation scheme using the CQI prediction scheme proposed by the authors Alessandro Chuimento and Mehdi Bennis in [5]. We have implemented this estimation scheme into the android information monitoring application *HTWK Signal Harvester* which allows to perform and record LTE/5G NR measurements using the Android API. The app collects cell information, cell signal strength as well as location information from GPS. In the second task the estimation of trace data is recorded while driving in an urban scenario. In this, a laptop connected to the internet via a smartphone as a tethered device has been used. Work has been performed in Future Internet Lab Anhalt [6].

2 EXTERIMENTAL SETUP

The following equipment was used for tests and measurements here: OnePlus 8T 5G mobile device (Android 11.0, Qualcomm Snapdragon 865, 8RAM +128 ROM), Lenovo ThinkPad T430 laptop (Kernel: GNU/Kubuntu 21. 10 5.13.0-30-generic64bit equipped with Intel® Core i5-3210M CPU 2.5GHz, 8GB of RAM), the server was running on Linux 5.13.0-28-generic Ubuntu 21.10 SMP x86_64 and equipped with Intel® Core Skylake, IBRS 3.8GHz, 4GB of RAM, 128GB NVMe Storage. The testbed, used in this research is presented in Figure 1.

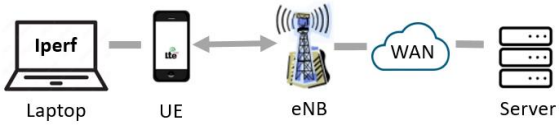


Figure 1: Network setup.

On the laptop - the sender runs the iPerf3 cross-platform sender-receiver console program [7]; on the other side the server opens the receiver side and starts writing data to a JSON file, and the smartphone runs the downlink data received from the connected base station and estimates the available bandwidth.

The main elements here are the OnePlus 8T 5G mobile device and the HTWK Signal Harvester [8] developed at HTWK University Leipzig. The HTWK Signal Harvester application is available as open source code at GitHub for information monitoring and analysis. The experiment was conducted along a closed route of the city of Köthen (Figure 2).

The wireless service providers were the local O2 provider and Alditalk provider as a virtual provider on the infrastructure of O2. The mobile device was connected to a laptop as an access point to the internet connection. The wireless link parameters were evaluated from a single UE served by the next eNB. The available bandwidth of the downlink is estimated at the MAC (Medium Access Control) level. The required TCP traffic load was generated using iPerf3. This was done in order to maximize the use of available bandwidth on the link. At the start of the traffic, both devices UE and Laptop started recording data in parallel. The route has been chosen so that there are a line of sight areas to the base station as well as dense urban areas with blind spots and areas with building obstacles and so with multi-path propagation scenarios.

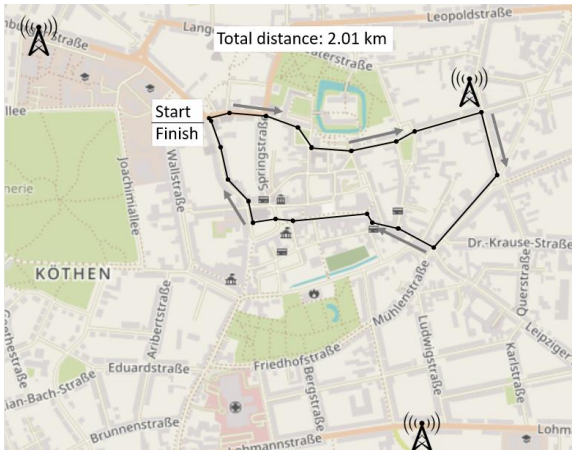


Figure 2: The route of the test scenario.

The mobile device also hands over to several different base stations along the route, which repeatedly changes the width of used LTE band and therefore changes the reception/transmission rate. After the metrics were collected, the results obtained by the mobile device and the iPerf3 throughput testing software were compared for verification of the results.

3 AVAILABLE BANDWIDTH ESTIMATION

The first parameter to look at when estimating the maximum available bandwidth is the CQI. This parameter cannot be calculated or in some other way determined from the information received from any

Android API. The CQI is vendor-specific, so no unambiguous value can be given. During the literature review, it was adopted to use the CQI indices and their interpretations are given from the official 3GPP TS specification [9] for reporting CQI based on QPSK, 16QAM, and 64QAM.

It is mentioned from the introduction that one resource block includes 12 subcarriers and 7 cyclic prefixes. The supported uplink-downlink configurations are listed in official 3GPP TS documentation [10]. Predominantly operators choose a configuration focusing on the downlink since templates with a predominant allocation to the uplink are only in demand by broadcasters or streamers. The resource blocks are allocated according to the appropriate channel bandwidth. That is, 50, 75, and 100 blocks are allocated for channel bandwidth 10, 15 and 20MHz respectively. The index of transport blocks for each specific case must then be calculated. These are determined by relating the CQI to the modulation scheme (MCS) (Table 8.6.1-1 3GPP documentation [9]). Using the transport block index in conjunction with the information already available, it is possible to suppose the number of transport blocks. To do this, it is needed to refer to the relevant section in the official documentation (Table 7.1.7.2.1-1 3GPP documentation [9]) and compare the I_{TBS} with the deferred number of resource blocks horizontally. In this way, we can construct a table of the number of allocated transport blocks size (TBS) for each modulation scheme in both directions (Table 1, Table 2).

Table 1: Downlink TBS allocation.

CQI	MCS	TBS Index	10MHz TBS	15MHz TBS	20MHz TBS
1	0	0	1 384	2 088	2 792
2	0	0	1 384	2 088	2 792
3	2	2	2 216	3 368	4 584
4	5	5	4 392	6 712	8 760
5	7	7	6 200	9 144	12 216
6	9	9	7 992	11 832	15 840
7	12	11	9 912	15 264	19 848
8	14	13	12 960	19 080	25 456
9	16	15	15 264	22 920	30 576
10	20	18	19 848	29 296	32 932
11	23	21	25 456	37 888	51 024
12	25	23	28 336	43 816	57 336
13	27	25	31 704	46 888	63 776
14	28	26	36 696	55 056	75 376
15	28	26	36 696	55 056	75 376

The resulting TBS size determines how much data (in bits) can be transmitted in one Transmission Time Interval (TTI) (=1ms). The last step is to multiply these values by 1000, thus obtaining the data rate in bits per second units. For convenience, the speed was further converted to Mbps. The correlation of the CQI values with RSSNR readings is carried out concerning the article [5].

Table 2: Uplink TBS allocation.

CQI	MCS	TBS Index	10MHz TBS	15MHz TBS	20MHz TBS
1	1	1	1 800	2 728	3 624
2	2	2	2 216	3 368	4 584
3	3	3	2 856	4 392	5 736
4	4	4	3 624	5 352	7 224
5	5	5	4 392	6 712	8 760
6	6	6	5 160	7 736	10 296
7	7	7	6 200	9 144	12 216
8	8	8	6 968	10 680	14 112
9	9	9	7 992	11 832	15 840
10	10	10	8 760	12 960	17 568
11	11	10	8 760	12 960	17 568
12	12	11	9 912	15 264	19 848
13	13	12	11 448	16 992	22 920
14	14	13	12 960	19 080	25 456
15	15	14	14 112	21 384	28 336

After the above steps, the implementation in the Android 11 API was carried out.

4 EXPERIMENTAL RESULTS

The suggested route of the mobile device in the city for taking LTE metrics has already been pointed out in Figure 2.

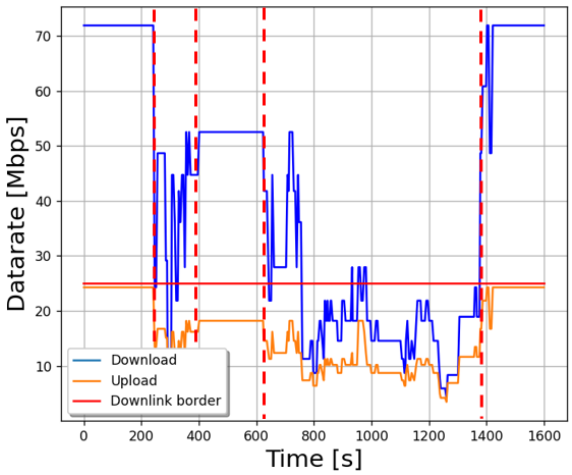


Figure 3: AvB estimation by android app.

The results of the metrics taken from Alditalk operator are presented in Figure 3. It must immediately be noted that the horizontal 'Downlink border' marks the maximum downlink speed set by the operator. As the ALDI operator has a data rate limit of 25Mbps - a comparison of the estimated data rate with the real rate would not give the required results. However, in comparison with the results obtained from the O2 operator, a general overlap of the estimated data rate over time areas separated by vertical lines can be observed. This is due to the fact that the ALDI operator does not have its base stations. The connection is provided by renting the capacity of the O2 operator's base stations. For the O2 operator, a brighter example can be observed (Figure 4), as the upper limit for download speeds is 225Mbps, which has not been achieved.

When defining the current SISO-MIMO scheme, difficulties were encountered in obtaining parameters using the Android 11 API. This parameter will be further estimated on base station locations and RSSNR values. MIMO in 4G LTE networks is primarily used due to spatial multiplexing that improves data rates by using multiple antenna elements that are physically separated in space on a transmitter or receiver.

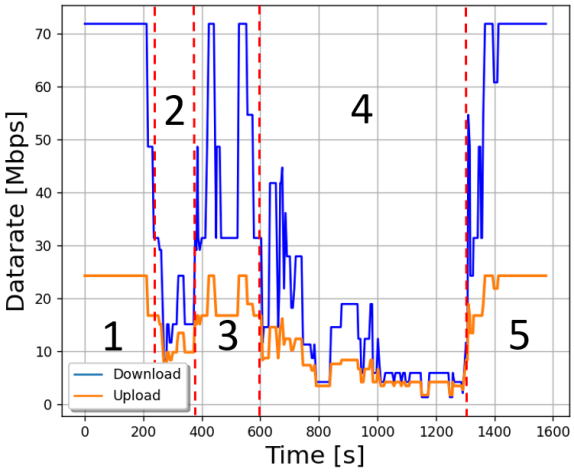


Figure 4: AvB estimation by android app.

To determine the connection scheme there are specialized software programs, e.g. SCAT: Signaling Collection and Analysis Tool or The Qualcomm Extensible Diagnostic Monitor (QXDM). The impact of MIMO on LTE link performance is described in detail in [11] and [12]. In our case, the spatial multiplexing pattern assumptions were determined based on the location of the nearest base stations and RSSNR values. Different MIMO schemes results in different data rates reached by Iperf3 below.

Figure 5 shows the iPerf3 receiver measurement results. The unstable throughput can be explained by measurements under TCP conditions and it's varying performance.

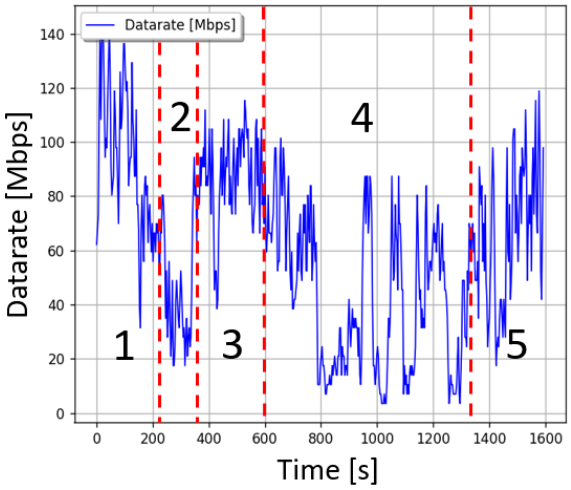


Figure 5: iPerf3 receiver side measurements.

During data analysis, it has been decided to divide the measurement area in Figure 4 and Figure 5 into 5 parts and evaluate them together. Measurements gathered during first time period on the Figure 4 correspond the situation of MIMO2x2 and the location of the base stations, which are determined by the eNB (eNodeB) ID. In the second part, there is a blind spot and the data transmission occurs by the reflection of the signal. The third zone shows a predominantly MIMO2x2 connection with medium signal strength. In the fourth zone, multiple switching between SISO and MIMO2x2 modes is observed due to the dense urban development and the close location of the base stations. In the fifth zone, there is a smooth transition to a stable SISO connection.

It is important to note that the measured values, presented in Figure 4 have been estimated for the SISO (Single Input Single Output) scenario. In a SISO system, a single antenna is used for transmission and reception. The required TCP traffic load with was generated using iPerf3 utility. By analyzing the graphs, it can be noted that the download speed has changed in the same time periods of the measurements. The occurrence of these areas can be explained by the blind spots along the route as well as by the time required for the mobile device to switch between base stations.

Trace was also taken in rainy weather conditions to assess the effect of weather conditions on internet connection speeds. An average deviation of 5-10% of the results against dry weather conditions has been found.

5 CONCLUSIONS AND FURTHER WORK

In this paper, a way has been described to estimate the available bandwidth based on cell information, cellular signal strength obtained by Android API for both uplink and downlink in SISO scenario conditions. Correspondence in the values of the estimated method with those of the trusted utility has been proven. The results gathered by android API are applicable for wireless network emulation in wired equipment for emulation but do not predict the actual available network bandwidth due to the complexity of determining the spatial coding method of the signal. The results and accuracy are expected and sufficient for our study. Subsequently, ideal experimental conditions can be reconstructed in the emulation. To precisely measure real AvB, it is necessary to determine the MIMO scheme, CQI metrics, the operator's network load as well as ideal weather conditions, and the absence of obstacles. The further work is to use this experience to emulate wireless channels in a wired environment to test the performance of the data transfer software in an Apposite Netropy [13] environment as a tool for datasets generation.

REFERENCES

[1] M. Yu Song, Y. H. Lee, and Ng. Boon. “The Effects of Tropical Weather on Radio-Wave Propagation Over Foliage Channel,” IEEE Transactions on Vehicular Technology, vol. 58, issue 8, pp. 4023 – 4030, 2009.

[2] S. Sebin, S. Renimol, D. Abhiram, and B. Premlet, “Effect of rainfall on cellular signal strength: A study on the variation of RSSI at user end of smartphone during rainfall,” 2017 IEEE Region 10 Symposium (TENSYMP), pp. 1-4, Cochin, India, July 2017.

[3] M. O. Alor, D. O. Abonyi, and P. Okafor, “Determination Of The Effect Of Rain On Cellular Signal Receptions,” International Journal Of Advances in Engineering and Management (IJAEM), vol. 2, issue 3, pp. 96-101, March 2015.

[4] F. J. Rida and A. B. Abood, “Survey of Improved Performance Radio Frequency Channels in Wireless Communication Systems,” International Journal of Civil Engineering and Technology (IJCIET), vol. 10, issue 12, pp. 70-83, 2019.

[5] A. Chiumento, M. Bennis, C. Desset, L. Van der Perre, and S. Pollin, “Adaptive CSI and feedback estimation in LTE and beyond: a Gaussian process regression approach,” EURASIP Journal on Wireless Communications and Networking, vol. 1, December 2015.

[6] Future Internet Lab Anhalt. [Online]. Available: https://fila-lab.de/. [Accessed: 9-Feb-2022].

[7] iPerf - the TCP, UDP and SCTP network bandwidth measurement tool. [Online]. Available: https://iperf.fr/. [Accessed: 20-Jan-2022].

[8] I. Kim, HTWK Signal Harvester. [Online]. Available: https://github.com/igorskh/harvester/. [Accessed: 14 - Jan-2022].

[9] Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (3GPP TS 36.213 version 14.2.0 Release 14), 2017.

[10] Evolved Universal Terrestrial Radio Access (E-UTRA); Physical channels and modulation (3GPP TS 36.211 version 14.2.0 Release 14), 2017.

[11] S. Krapovic, S. Mujkic, and S. Muracic, “Enhanced MIMO influence on LTE-Advanced network performances,” ELEKTRONIKA IR ELEKTROTECHNIKA, vol. 22, pp. 81-86, February 2016.

[12] P. Shah, K. Sakhardande, and G. Shah, “Performance Analysis of LTE Network using QAM and MIMO Configuration,” 2018 IEEE International Students’ Conference on Electrical, Electronics and Computer Science, vol. 3, pp. 1-6, February 2018.

[13] Leaders in network emulation and testing. [Online]. Available: https://www.apposite-tech.com/. [Accessed: 11-Feb-2022].

Frequency-Difference Brillouin Reflectometry of Optical Fiber Parameters

Igor Bogachkov[1], Nikolay Gorlov[2], Tatyana Monastyrskaya[3] and Natalya Medvedeva[4]
[1]*Department of Communications and Information Security, Omsk State Technical University, 11 Mira Str., Omsk, Russia*
[2]*Department of Communication Lines, Siberian State University of Telecommunications and Computer Science, 86 Kirov Str., Novosibirsk, Russia*
[3]*Department of Social Communication Technologies, Siberian State University of Telecommunications and Computer Science, 86 Kirov Str., Novosibirsk, Russia*
[4]*Department of Foreign Languages for Technical Faculties, Novosibirsk State Technical University, 20 Karl Marx avenue, Novosibirsk, Russia*
bogachkov@mail.ru, gorlovnik@yandex.ru, t.monastyrskaya@mail.ru, n.medvedeva@corp.nstu.ru

Keywords: Brillouin Reflectometry, Frequency Shift, Frequency Synthesizer, Optical Modulator, Radiation Spectrum.

Abstract: The paper discusses the variety of the Brillouin reflectometry technique proposed by the authors. The distinctive feature of this technique is the isolation of the differential frequency of Brillouin scattering signals which come from the inhomogeneity and the fiber segment adjacent to it. First, a number of components are allocated from the entire signal spectrum, each of which occupies a narrow frequency band. Each of these narrow-band components is further subjected to amplitude detection and averaged with the data of several measurements. The proposed technique allows implement the simplest and most convenient principle of constructing a Brillouin reflectometer, and also it enables to obtain Brillouin reflectograms from conventional Rayleigh reflectometers. This technique can find wide applications both in the field of distributed fiber-optic sensors and for early troubleshooting of telecommunication optical cables.

1 INTRODUCTION

The well-known problem of the high cost of equipment which characterizes the current situation in the field of Brillouin reflectometry creates prerequisites to search for innovative technical solutions. Among the factors determining the complexity of measuring equipment [1], we should mention a sufficiently great significance of the Brillouin frequency shift with its relatively small change depending on the measured parameters (tension and temperature). The devices must contain a tunable frequency synthesizer with the necessary characteristics, as well as an optical modulator and a photodetector with the appropriate speed. We should also mention the high requirements for the radiation spectrum of the laser used.

2 THEORETICAL BACKGROUNDS

Works [2] and [3] describe the optical signal heterodyne oscillation generated by forced Mandelstam-Brillouin scattering (FMBS). Developing this idea we propose a technical solution based on the use for this purpose of a spontaneous Brillouin scattering signal received from an adjacent section of an optical fiber with relation to the measured one.

Figure 1 shows a part of the schematic diagram of the reflectometer working on the described principle of operation.

The scheme contains a laser and an optical modulator, two directional couplers, between which there is a fiber segment that creates the necessary time delay. Optical probing impulses are formed at the modulator output. Radiation from the couplers outputs enters the photodetectors, their spectrum of the output signals has components conditioned by Brillouin scattering. The frequency of one of these

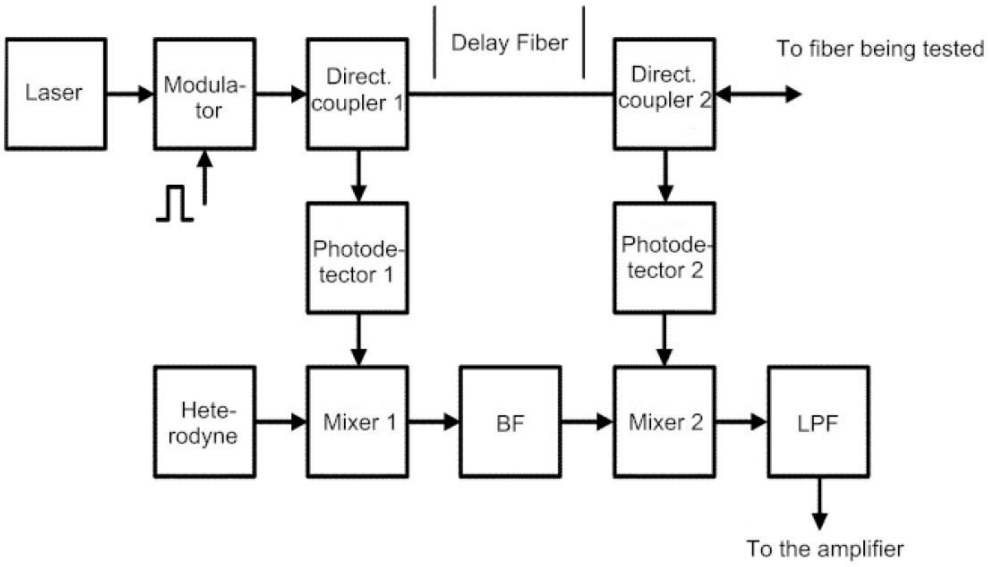


Figure 1: Schematic diagram of the frequency-difference Brillouin reflectometer.

signals is shifted by a mixer and a heterodyne oscillator. The received signal as well as the signal from the output of another photodetector are fed into the second mixer to form a differential frequency. After being filtered by means of a low-pass filter (LPF), the output signal from this mixer is amplified, digitized and processed using a microcomputer to obtain a reflectogram.

According to the scheme, the signal frequency at the output of the LPF placed after the second mixer is determined (at a fixed frequency of the heterodyne oscillator) by the frequency difference of the Brillouin scattering signals in neighboring fiber sections whose beginnings are located at a distance equal to the length of the delay fiber.

Additional frequency conversion is particularly necessary to enable the identification of the direction of change in the tension (or temperature) of the fiber as the wave path propagates along it. This inhomogeneity of the fiber state will be further referred to as Brillouin inhomogeneity. Let us consider the probing impulse and the inhomogeneity, the lengths of which do not exceed the length of the delay fiber l_d, (Figure 2). The figure shows how the frequency difference between the source and delayed signals changes when the packet passes through the inhomogeneity in the absence and the presence of additional frequency conversion.

In the first case, the frequency shift of the Brillouin signal due to the presence of inhomogeneity does not depend on which of the packets (the source one or the delayed one) is inside the inhomogeneity. In both cases, the frequency increment Δf_H is the same. It is also possible that the first inhomogeneity is followed by the second one, in which the frequency shift occurs in the same direction and by the same value. If the position of the source packet corresponds to this second inhomogeneity, and the position of the delayed one corresponds to the first inhomogeneity, then the frequency shift will also be equal to Δf_H and it will not be possible to distinguish this situation from the one shown on the right top in Figure 2.

The presence of additional frequency conversion makes it possible to solve this problem (Figure 2, lower graphs). Additional frequency shift Δf_{add} must satisfy the inequation [4].

$$\Delta f_{add} \geq |\Delta f_{\mathrm{H}\ max}| + \frac{2}{t_{min}},$$

where $\Delta f_{H\ max}$ is the maximum frequency shift due to inhomogeneity;

t_{min} is the minimum duration of the signal element corresponding to the Brillouin inhomogeneity.

The second term takes into account the restriction imposed by Kotelnikov's theorem. It enables us to detect short inhomogeneities, which also have a small effect on the frequency of the scattered signal.

The upper graph corresponds to the absence of an additional frequency conversion, and the lower graph corresponds to the presence of this conversion.

To obtain a tension or temperature distribution curve along the fiber, it is necessary to integrate the dependence of the difference frequency on time (distance), taking as zero the value corresponding to the shift frequency Δf_{add}.

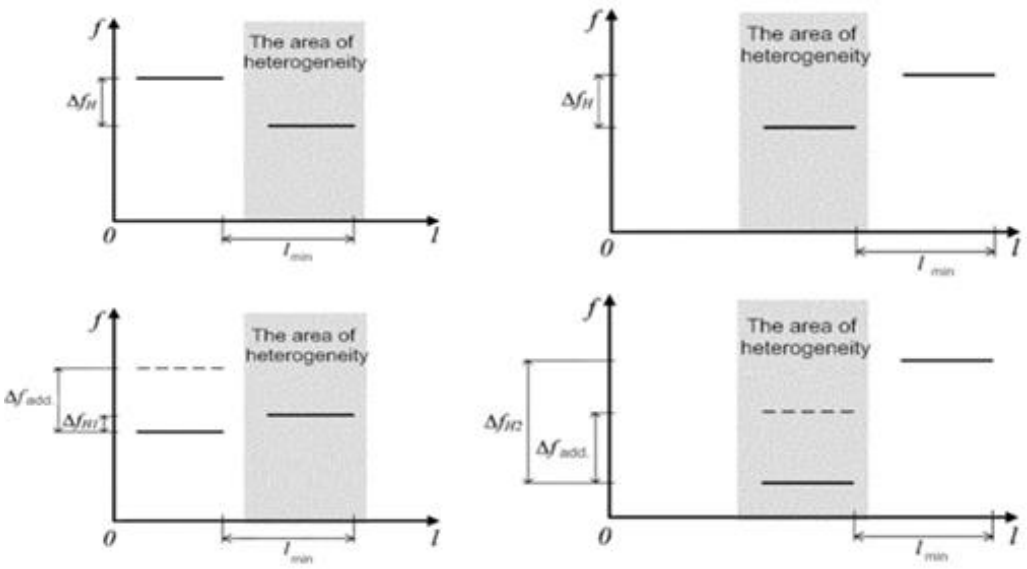


Figure 2: Frequencies of Brillouin scattered signals for a short-length probing impulse.

We should note a number of significant deficiencies of the scheme, which include high requirements for the speed of photodetectors, the presence of microwave mixers, as well as the fact that due to the use of an additional directional coupler, the power of the backscattering signal drops by 3 dB. Therefore, a technical solution providing for the presence of adirectional coupler and one photodetector at the input of which there are both outgoing and delayed optical signals is of considerable practical interest. At the output of such a detector, it is possible to distinguish difference spectral components and this significantly reduces the requirements for its performance. We have to find a method to determine the direction of change of the measured value in the path of propagation of the probing impulse [5].

Obviously, the longer the delay time, the longer inhomogeneities can be determined unambiguously. This trend also takes place concerning the increase in the length of the probing impulse (Figure 3).

Let us note that in the situation shown in Figure 3, with a probing impulse length exceeding the length of the Brillouin inhomogeneity, signals scattered by both this inhomogeneity and adjacent fiber sections will simultaneously occur at the photodetector input. This allows us to formulate a very important conclusion. The signal at the output of the photodetector of a conventional reflectometer can carry information about the presence of Brillouin inhomogeneities in the fiber and theoretically can be detected in digitized data stored in the device memory by special processing.

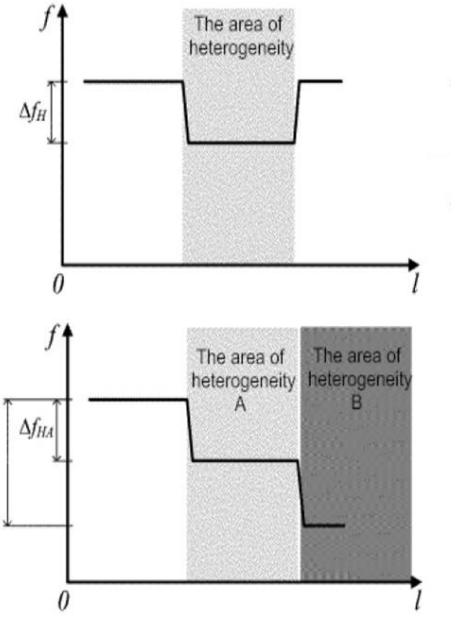


Figure 3: Frequencies of Brillouin scattered signals for a long-length probing impulse.

The increase in the wavelength of the probing impulse is known to decrease the resolution of the OTDR [6]. Therefore, each measurement cycle of this kind should include obtaining a number of reflectograms at different packet lengths. The data generated by the reflectometer and transmitted to further processing stages should not be the result of an averaged set of received reflectograms, since in this case the Brillouin component of the signal would be suppressed.

In order to extract data on Brillouin inhomogeneities, processing of the digital image of the fiber response stored in the memory of the reflectometer can be performed as follows. First, a number of components are allocated from the entire signal spectrum, each of which occupies a narrow frequency band. A set of such adjacent frequency bands should cover the entire spectrum region of the original signal, in which components due to Brillouin scattering may occur. Each of these narrow-band components is further subjected to amplitude detection. This is followed by averaging over the data of several measurements.

The data set formed at the detection stage and averaged carries information about both the frequency of the Brillouin scattering signal and its amplitude. The latter is determined by summing the amplitudes of all components and can be determined only in cases where there is at least one non–zero summand, i.e., in the fiber sections where the frequency of the Brillouin signal changes. For the remaining points the amplitude should be determined by interpolation and extrapolation methods.

Information about the value of the Brillouin frequency shift is contained in the distribution of amplitudes between the components and can be isolated in the manner described below. With each signal envelope obtained as a result of detection, an action is performed that can be called conditional or adaptive inversion. The next step is time integration. The data obtained as a result of it for each narrow frequency band should be multiplied by weight coefficients, and the products should be summed up element-by-element in order to obtain the dependence of the tension (or temperature) of the fiber on its length. The result together with the previously obtained data on the signal amplitude is used to form a three-dimensional Brillouin reflectogram.

The idea of adaptive inversion is that the signal is either inverted or not, depending on certain conditions. The need for inversion is due to the absence of an additional frequency shift Δf_{add}, considered in the first part of the paper (see Figure 1), and, accordingly, because of the fact that the signal directly received by the described technique does not contain information about the signs of change in the measured value as the probing impulse propagates along the fiber. As we have shown earlier, this information can be extracted from the data of several measurements in which packets of different lengths are used.

Thus, during the measurement process, a number of sets of digital data samples must first be obtained, representing the averaged results of the amplitude detection mentioned above. Sets are formed when different probing impulse lengths are specified. Then we make the tables of points in the inversion mode of signals for switching on and off o. Further processing is carried out taking into account these tables.

Let's note also that the absence of an additional frequency shift Δf_{add} excludes the possibility of detecting short inhomogeneities, which are characterized by a slight change in the frequency of the Brillouin scattered signal. This is a significant disadvantage of this variant of the technical implementation of the considered technique.

It should be assumed that not every reflectometer can be used for this kind of measurement. In addition to the ability to save raw data received from the ADC to a file and the possibility of external control from a computer, it is necessary to note the special requirements for its laser.

The requirements for Brillouin reflectometer lasers are significantly more stringent than for conventional ones. The laser should form one narrow spectral line [7]. A significant change in the oscillation frequency of the probing impulse throughout its length (chirp) is not allowed. This phenomenon is typical for the case of applying a modulating signal directly to the laser. Therefore, the reflectometer must have a separate optical modulator. We should emphasize that the considerable width of the spectral line makes it difficult or impossible to identify inhomogeneities characterized by a slow increase and decrease in the frequency of the Brillouin signal.

In addition, it follows from the abovesaid that the duration of the probing impulse should be several times longer than the maximum period of the difference frequency component of the signal involved in this measurement. This means that the bandwidth of the receiving path of the reflectometer should be wider than necessary for conventional reflectometry for a given packet duration. Since the signal-to-noise ratio increases when the bandwidth is narrowed, developers most likely strive to optimize it for each duration of the probing impulse, which is a significant obstacle to the proposed measurement technique [8].

The study of the available reflectometers for their suitability for the considered application has not been carried out. However, the authors consider it to be a very promising area of research, giving a chance to implement the simplest and most affordable installation for Brillouin reflectometry. Although this method cannot be considered a fully-functional replacement for the conventional one, it can be widely used both in the field of distributed fiber-optic sensors and for early diagnostics of telecommunications optical cable malfunctions [9].

3 RESEARCH AND DISCUSSIONS

The authors of the report conducted experimental studies using the Brillouin optical reflectometer of the Swiss company "Omnisens SA" DITEST Interrogator on a single-mode optical fiber of the G.652 standard according to the scheme in Figure 4.

The magnitude of the Brillouin frequency shift from the radius of curvature of the optical fiber winding was investigated. At the same time, two cylinders with diameters of 37.5 mm and 19.9 mm were used. In both cases, 2 meters of optical fiber were wound on the cylinders. The measurements were carried out at different spatial resolution capabilities of the device.

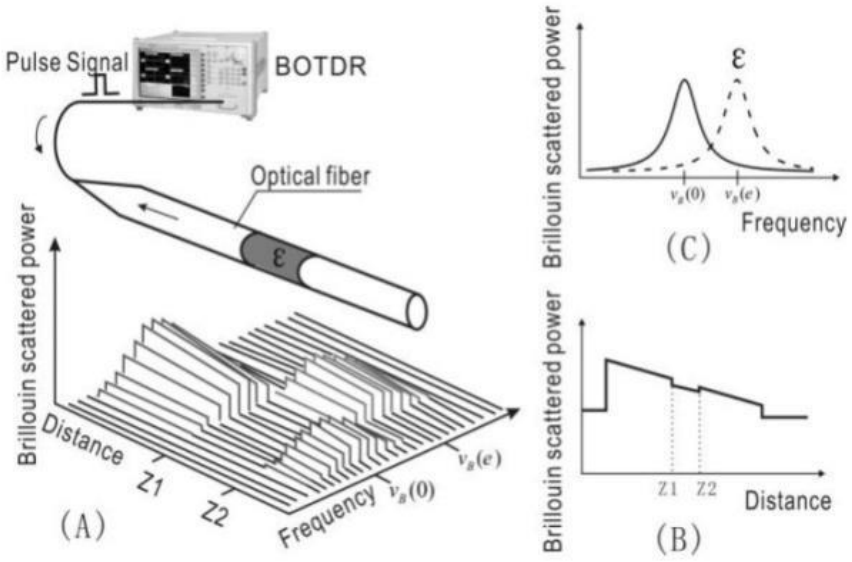


Figure 4: The scheme of experimental studies.

The spectrograms obtained during the experiment are shown in Figures 5-8, where 1 – 17 turns, diameter 37.5 mm, 2 – 32 turns, diameter 19.9 mm.

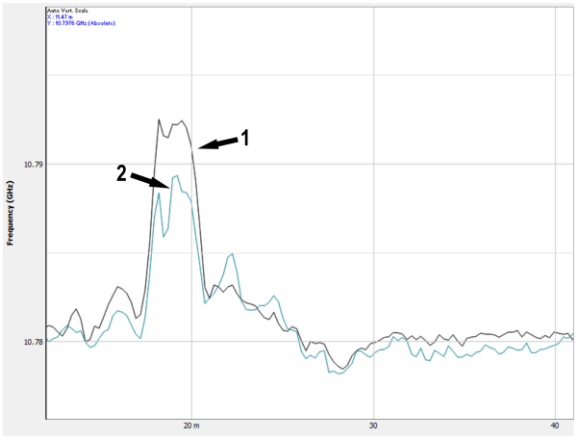


Figure 5: Spectrogram for an optical fiber of the G.652 standard at a resolution of (1-0.25) m.

With an average Brillouin shift without deformation of 10.7802 GHz, the difference was 2.9 MHz.

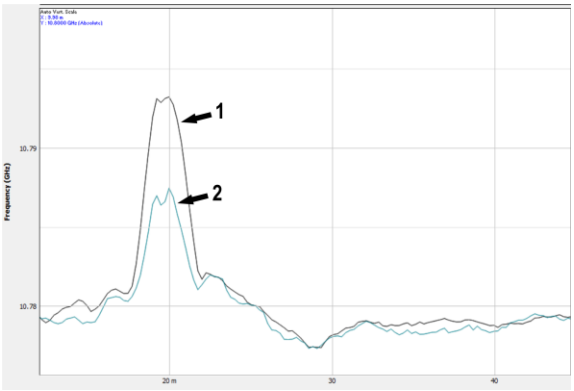


Figure 6: Spectrogram for an optical fiber of the G.652 standard at a resolution of (2-0.25) m.

The additional Brillouin shift was 5.7 Mhz.

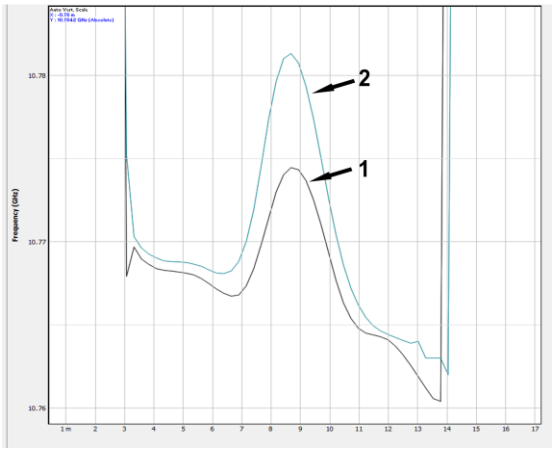


Figure 7: Spectrogram for an optical fiber of the G.657 standard at a resolution of (2-0.25) m.

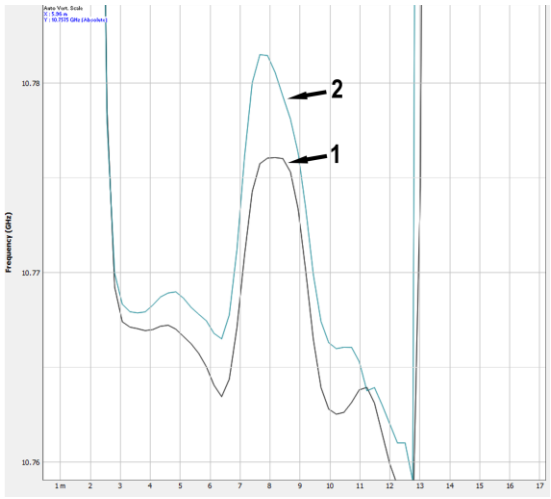


Figure 8: Spectrogram for an optical fiber of the G.657 standard at a resolution of (1-0.25) m.

The results obtained show that the value of the additional Brillouin frequency shift is inversely proportional to the bending radius of the optical fiber.

4 CONCLUSIONS

The proposed model of frequency-difference Brillouin reflectometry enables the improvement of the control quality of the main transfer parameters of optical fibers. In addition, the model is efficient for monitoring during the technical operation of fiber-optic communication lines.

Optical reflectometers based on the Brillouin scattering principle can be widely used in telecommunications systems, mechanical engineering, electric power, construction, aviation and space. They can take a special place in control and automation systems of technological processes and objects. They are widely used to prevent natural disasters. However, there are no generally accepted or well-developed standards for Brillouin reflectometer applications, which has led to the embarrassing fact that they cannot be adopted in some intelligent design and construction structures, especially for safety-related functions. This fact is pointed out in the works of researchers S. Delepine-Lesoille, J. Bertrand, L. Lablonde and X. Phéron [10]. The details of the application should be guided and limited by the relevant industry standards. Proposals and the development of standards or guidelines are necessary to promote the popularization of the use of Brillouin reflectometers.

The objectives of further research will be the metrological analysis of all stages of transformation of measuring information and replenishment of the database of Brillouin spectrograms of optical fibers of various types. Of particular importance is the improvement of algorithms for the automatic processing of spectra in order to expand the functionality of the sensors under study.

REFERENCES

[1] N. H. Zhu, J. H. Ke, and W. Chen, “Wavelength Coded Optical Time-Domain Reflectometry”, Journal Of Lightwave Technology, vol. 28, No 6, March 15, 2010, pp. 972-976.

[2] U. Glombitza, and E. Brinkmeyer, “Coherent Frequency-Domain Reflectometry for Characterization of Single-Mode Integrated Optical waveguides”, Journal Of Lightwave Technology, vol. 11, 1993, pp. 1377-1384.

[3] O. Kamatani, and K. Hotate, “Optical Coherence Domain Reflectometry by Synthesis of Coherence Function with Nonlinearity Compensation in Fequency Modulation of a Laser Diode”, Journal Of Lightwave Technology, vol. 11, 1993, pp. 1854-1862.

[4] H. F. Martins, S. Martin-Lopez, and P. Corredera, “Coherent Noise Reduction in High Visibility Phase-Sensitive Optical Time Domain Reflectometer for Distributed Sensing of Ultrasonic Waves”, Journal Of Lightwave Technology, vol. 31, No 23, December 1, 2013, pp. 3631-3636.

[5] J. Luo, Y. Hao, Q. Ye, and L. Li, “Development of Optical Fiber Sensors Based on Brillouin Scattering and FBG for On-Line Monitoring in Overhead Transmission Lines”, Journal of Lightwave Technology, vol. 31, No 10, May, 2013, pp.1559-1565.

[6] F. Barrios, S. López, A. Sanz, P. Corredera, and J. Castañón, “Distributed Brillouin Fiber Sensor Assisted by First-Order Raman Amplification”, Journal of Lightwave Technology, vol. 28, No 15, August 1, 2010, pp. 2162-2172.

[7] M. Belal, and T. P. Newson, “Experimental Examination of the Variation of the Spontaneous Brillouin Power and Frequency Coefficients Under the Combined Influence of Temperature and Strain”, Journal of Lightwave Technology, vol. 30, No 8, May, 2013, pp. 1250-1255.

[8] Y. Li, X. Bao, Y. Dong, and L. Chen, "A Novel Distributed Brillouin Sensor Based on Optical Differential Parametric Amplification", Journal of Lightwave Technology, vol. 28, No 18, September 15, 2010, pp. 2621-2626.

[9] J. Luo, Y. Hao, Q. Ye, Y. Hao, and L. Li, "Development of Optical Fiber Sensors Based on Brillouin Scattering and FBG for On-Line Monitoring in Overhead Transmission Lines", Journal Of Lightwave Technology, vol. 31, No 10, May 15, 2013, pp. 1559-1565.

[10] S. Delepine-Lesoille, J. Bertrand, L. Lablonde, and X. Phéron, "Distributed Hydrogen Sensing With Brillouin Scattering in Optical Fibers", IEEE Photonics Technology Letters, vol. 24, No 17, September 1, 2012, pp. 1475-1477.

Anomaly Detection with Various Machine Learning Classification Techniques over UNSW-NB15 Dataset

Martina Shushlevska, Danijela Efnusheva, Goran Jakimovski and Zdravko Todorov
Computer Science and Engineering Department, Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, 18 Rugjer Boshkovik Str., Skopje, R. N. Macedonia
martinasuslevska@yahoo.com, {danijela, goranj, todorovz}@feit.ukim.edu.mk

Keywords: Anomaly Detection, Intrusion Detection System, Machine Learning, Network Security, UNSW-NB15 Dataset.

Abstract: The exponential growth of computers and devices connected to the Internet and the variety of commercial services offered creates the need to protect Internet users. As a result, intrusion detection systems (IDS) are becoming an essential part of each computer-communication system, detecting and responding to malicious network traffic and computer abuse. In this paper, an IDS based on the UNSW-NB15 dataset has been implemented. The results obtained indicate F1 Score and Recall values of 76.1% and 85.3% for the Naive Bayes algorithm, 78.2% and 96.1% for Logistic Regression algorithm, 88.3% and 95.4% for Decision Tree classifier, and 89.3% and 98.5% for Random Forest.

1 INTRODUCTION

Network attacks are one of the biggest security problems in the world today. The constant increase in computers, mobile phones, sensors, IoT devices, big data, web applications, server and cloud systems, and more sophisticated computing resources imposes even more significant challenges for keeping network connections under control. Additionally, the enormous increase in network traffic has caused many new approaches to network intrusions to be planned by various hackers and malicious users. Therefore, IDS are a rapidly evolving field aimed at providing detection of malicious behaviour and attacks in the network [1].

The two crucial methods for detecting threats that intrusion detection systems can use are: signature-based and anomaly-based [1]. Signature-based detection is usually applied in identifying known threats, by using a pre-programmed list of them and their indicators of compromise. In fact, an indicator of compromise could be a specific behaviour that generally precedes a malicious network attack, known byte sequences, malicious domains, file hashes, or even the content of email subject headings. On the other hand, an anomaly-based IDS is used to alert a suspicious behaviour that is unknown. An anomaly-based detection system doesn't operate by searching for known threats, but it may utilize machine learning methods for training the detection system to recognize a normalized baseline. This baseline shows what is the system's normal behaviour, and then all network activity is compared to that baseline. Therefore, instead of searching for known indicators of compromise, an anomaly-based IDS identifies any odd behaviour in order to trigger alerts.

Many techniques have been developed to detect anomaly-based intrusions by applying data mining and machine learning methods [2-7]. Mainly, well-known datasets (ex. KDDCUP'99, NLS-KDD, UNSW-NB15) that consist of real-time network traffic with a large number of features are used in anomaly-based intrusion detection [8], [9]. Therefore, in this paper, we implement several ML algorithms over the UNSW-NB15 dataset to analyse and verify that machine learning is very applicable for solving a problem with unauthorized attacks in network traffic. Assuming that KDDCUP'99 and NSL-KDD benchmark datasets were generated a decade ago, in this research, we use the UNSW-NB15 dataset that was published in 2015. This dataset targets more realistic and network traffic and novel types of modern attacks. Indeed, UNSW-NB15 is a network intrusion dataset that contains raw network packets, characterized with 49 features

and organized in 10 categories (9 attack types plus 1 for normal activity) [9].

This paper aims to examine the differences between a Naive Bayes (NB), a Logistic Regression (LR), a Decision Tree (DT), and a Random Forest (RF) ML algorithms in order to determine the strengths and weaknesses of using these methods over the UNSW-NB15 dataset. By evaluating the performance of these algorithms in terms of accuracy, precision, recall, and F1 metrics, we can consider which of the analysed classification methods is the most effective and suitable for detecting anomalies.

The rest of the paper is organized as follows: Section II presents the state of the art in the domain of intrusion detection. Section III provides a brief description of the UNSW-NB15 dataset and explains how the ML model is built. Section IV analyses the results from several classification methods, including Naive Bayes, Logistic Regression, Decision Tree, and Random Forest. Section V concludes the paper and provides directions for future work.

2 CURRENT STATE

As more people use the Internet for personal or business reasons, different cyber-attacks and intrusions grow daily. An IDS is one of the most crucial considerations of cyber-security. This type of system can be software or hardware-based and can recognize successful violations even after they have happened. Generally, an IDS's purpose is to monitor network packets or systems to detect malicious activity and take specific measures [1].

There are many types of IDSs, which are discussed and summarized below.

A host-based IDS (HIDS) monitors and analyzes the internal computing system or system-level activities of a single host such as system configuration, application activity, wireless network traffic (only for that host) or network interface, system logs or audit log, running user or application processes, file access and modification security logs [10]. Examples of some known HIDS systems are Tripwire and OSSEC.

A network-based IDS (NIDS) monitors and analyzes network traffic on specific network segments for suspicious activities detection. This type of IDS is activated when packets enter a particular network from the Internet, and its function is to decide whether to reject or accept the entry packets and pass them to the local network. An example of a known NIDS system is Snort [11].

A protocol-based IDS (PIDS) monitors and checks the specific protocol behavior and its state like HyperText Transfer Protocol (HTTP). It focuses on actions in some particular application by monitoring and analyzing the application log files or measuring their performance. A PIDS approach for detecting jamming attacks in a LoRaWAN network is proposed in [12].

A wireless IDS (WIDS) monitors wireless networks to detect any harmful activity (ex. too many de-authentication packets, too many broadcast requests, analysis of the number of packets sent during a single time window). If malicious behavior from certain users is detected, they forbid them from connecting to the wireless network access point. Examples of some known WIDS systems are Kismet and NetStumbler [13].

Network behavior analysis (NBA) monitors and checks network traffic to detect threats that produce uncommon traffic flows, such as DDOS attacks, malware, and policy violations [14]. It is recommended to be used together with a firewall and other types of IDS systems.

Nowadays, due to increased use of the Internet and company networks, network traffic increases daily. Access to company networks should be given only to authorized users, so, detecting unauthorized entities or intruders is necessary. Machine learning techniques have been used and applied in many studies [2-7], where they have provided solid results in detecting intrusions and protecting the network from sudden attackers. The applicability of ML for intrusion detection systems is due to well-known technologies, such as identification, extraction, classification, regression, and prediction, as well as solid datasets composed of real-time network traffic with many features and their description. For example, the research in [2] and [3] gives an opportunity to review classification techniques and ML models for an IDS application.

There is also a hybrid attack detection system based on SVM (Support Vector Machine) and C5.0 Decision Tree proposed by authors in [4], where using a combination of popular ML algorithms improves the accuracy of attack detection, compared to being used apart. A similar hybrid system in which two algorithms (K-means and NB) are used to group some data and classify it is proposed by authors in [5]. MapReduce is very popular for processing extensive structured and unstructured data placed in key/value pairs. The authors of [6] propose an intrusion detection model that uses

MapReduce. MapReduce relies on using a combination of Fuzzy C-means (FCM) and SVM for classification and generating key pairs/values for attack detection. Furthermore, a survey of different approaches for intrusion detection with deep learning is given in [7].

3 OVERVIEW OF UNSW_NB15 DATASET AND BUILDING ML MODELS

UNSW-NB15 is a network traffic dataset with different categories for normal activities and malicious attacks, generated by the Australian Center for Cyber Security and published in 2015, [9]. This dataset includes 100 GB of raw network traffic (pcap files) generated as a hybrid of real normal activities and synthetic contemporary attack behaviors. Indeed, the traffic is categorized into nine different attacks and a wide range of real normal activities. The complete dataset contains 257,673 records, each represented by 49 features and a class label.

The following text discusses the nine types of attacks that are included in the UNSW-NB15 dataset:

1) Analysis: a type of attack where the attacker listens to the network traffic and then performs analysis of the observed data.
2) Backdoors: a type of attack that provides attackers with unauthorized remote access to a system without the usual authentication process.
3) DoS: a type of attack in which the attacker crashes or floods the services of a target machine, in order to make it overloaded and unavailable for serving further requests.
4) Exploit: a type of attack which utilizes the software vulnerabilities and errors within the networks, operating systems or hardware.
5) Fuzzers: a type of attack in which the attacker tries to stress the application in order to cause unexpected behavior, such as resource leaking or even crashes.
6) Generic: a type of attack that acts against a cryptographical primitive and it tries to break the key of some secure system.
7) Reconnaissance: a type of attack that gathers information about the target computer network in order to bypass its security control. Some examples are: phishing, social engineering port scanning, packet sniffing, etc.
8) Shellcode: a type of malware attack in which the attacker uses a special type of code that is used to exploit a variety of software vulnerabilities, so the attacker could take control over the compromised machine.
9) Worms: a type of malware attack that replicates itself in order to be spreaded to other computers by a computer network.

The most common attacks in the UNSW-NB15 database are Generic and Exploits, with are a total of 40000 and 33393 records, respectively. Additionally, if an analysis of the number of malicious or normal dataset records is made, we get the distribution shown in Figure 1. Here it can be seen that there is a higher prevalence of malicious records (68.06%) compared to the prevalence of normal traffic records (31.94%). Malicious records include the nine types of previously described attacks.

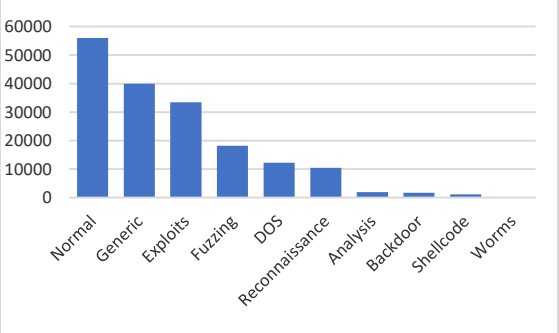


Figure 1: Number of records that represent normal traffic and malicious types of attacks in the UNSW-NB15 dataset.

The UNSW-NB15 dataset is characterized with 49 features shown in Table 1. These features are organized in six groups, discussed below:

1) Flow features (0-5): These features have the identifier attributes between hosts (client-server and vice-versa).
2) Basic features (6-18): These features include the attributes that represent protocols connections.
3) Content features (19-26): These features contain the attributes of TCP/IP and as well some attributes of http services.
4) Time features (27-35): This group contains the attributes of time, such as: start/end packet time, arrival time between packets, and round trip time of TCP protocol.
5) Additional generated features (36-47). This group can be further divided into two groups:

- General purpose features (36-40), whereas each feature of this group has its own purpose, in order to protect the service of protocols.
- Connection features (41-47) are built from the flow of 100 record connections based on the sequential order of the last time feature.

6) Labelled Features (48-49): This category shows the label and attack type of each record.

Table 1: UNSW-NB15 dataset features.

N	Feature	Description
1	srcip	Source IP address
2	sport	Source port number
3	dstip	Destination IP address
4	dsport	Destination port number
5	proto	Protocol type
6	state	The state
7	dur	Row total duration
8	sbytes	Source to destination bytes
9	dbytes	Destination to source bytes
10	sttl	Source to destination time to live
11	dttl	Destination to source time to live
12	sloss	Source packets retransmitted or dropped
13	dloss	Dest. packets retransmitted or dropped
14	service	Such as http, ftp etc.
15	sload	Source bits per second
16	dload	Destination bits per second
17	spkts	Source to dest. packet count
18	dpkts	Dest. to source packet count
19	swin	Source TCP window adv. value
20	dwin	Source TCP window adv. value
21	stcpb	Source TCP base seq. num.
22	dtcpb	Dest. TCP base seq. num.
23	smeansz	Mean of the packet size transmitted by the srcip
24	dmeansz	Mean of the packet size transmitted by the dstip
25	trans_depth	The connection of http req./resp. transaction
26	res_bdy_len	The content size of the data transferred from http
27	sjit	Source jitter
28	djit	Destination jitter
29	stime	Row start time
30	ltime	Row last time
31	sintpkt	Source inter-packet arrival time
32	dintpkt	Dest. inter-packet arrival time
33	tcprtt	Setup round trip time
34	synack	Time between SYN and SYN_ACK packets
35	ackdat	Time between SYN_ACK and ACK packets
36	is_srn_ips_ports	If srcip(1)=dstip(3) and sport(2)=dsport(4), assign 1else 0
37	ct_state_ttl	No. for each state (6) according to values of sttl(10) and dttl(11)
38	ct_flw_http_mthd	No. of fows that has methods like Get and Post in http service.
39	is_ftp_login	If the ftp session is accessed by user and password then 1 else 0.
40	ct_ftp_cmd	No of fows that has a command in ftp session.
41	ct_srv_src	No. of rows of the same service(14) and srcip(1) in 100 rows
42	ct_srv_dst	No. of rows of the same service(14) and dstip(3) in 100 rows
43	ct_dst_ltm	No. of rows of the same dstip(3) in 100 rows
44	ct_src_ltm	No. of rows of the same srcip(1) in 100 rows
45	ct_src_dport_ltm	No. of rows of the same srcip(1) and the dsport(4) in 100 rows
46	ct_src_sport_ltm	No. of rows of the same dstip(3) and the sport(2) in 100 rows
47	ct_dst_src_ltm	No. of rows of the same srcip(1) and the dstip(3) in 100 rows
48	attack_cat	Type of attack
49	label	0 for normal and 1 for attack

Python is used to process the UNSW-NB15 dataset in conjunction with the Jupyter Notebook tool that is an open-source web application used for generating and sharing documents which contain live code, equations, visualizations, and text. In particular, the following Python libraries [15] have been used in the analysis, processing and creation of the classification models: Pandas, NumPy, matplotlib.pyplot, Seaborn and sklearn (Scikit-learn).

The UNSW-NB15 dataset is defined by two files, a training set and a testing set (UNSW_NB15_training-set.csv and UNSW_NB15_testing-set.csv respectively). The training set includes 175,341 records, while the testing set includes 82,332 records. Accordingly, 31.95% of the records belong to the testing set, and 68.05% of the records belong to the training set. Each record can represent some of the nine types of attacks or normal traffic.

5-Fold Cross-Validation is used to build an ML model, so the UNSW-NB15 dataset is divided into 5 parts. In the first iteration, the first section is used to validate the model, and the rest (the other 4 sections) are used to train the model. In the second iteration, the second division is used as a validation set, while the others serve as training sets. This process is repeated until each of the five divisions is used as a validation set. This method is used for building an ML model for each of the analyzed algorithms, including: Naive Bayes, Logistic Regression, Decision Tree and Random Forest.

4 ANALYSIS OF RESULTS

In this paper, a research is done that develops a system for detecting attacks by differentiating anomalies from normal data flow based on network behavior. One advantage of this approach is that when an attack occurs, the network behavior will deviate from the normal pattern of behavior and the anomaly will be detected. In order to avoid the effect of data sampling when assessing the IDS, 5-fold cross-validation (CV) method is used. Four different machine learning algorithms (NB, LR, DT and RF) are applied on the dataset.

The analyzed metrics of the experiment are: CV fit time, CV accuracy mean, CV precision mean. CV recall mean, CV F1 mean, CV AUC mean, Accuracy test, Precision test, Recall test, F1 test and AUC test. The CV fit time refers to the required time for fitting the estimator on the train set for each of the five CV splits. In fact, the performance CV metrics reported by 5-fold cross-validation are calculated as an average of the values computed in 5 steps. In each of the steps, the model is trained using 4 of the folds as training data, and validated with the remaining part of the data. After that, final evaluation is done on the testing set, by measuring the Accuracy test, Precision test, Recall test, F1 test and AUC test values. Accuracy identifies how many observations, both positive and negative, were properly classified. Precision represents the ratio of properly predicted positive observations to the total predicted positive observations. Recall is the ratio of properly predicted positive observations to the all observations in an actual class. F1 Score combines precision and recall in one metric by calculating the harmonic mean between them. AUC is the area under the ROC (Receiver Operator Characteristic) curve, which is used to show the diagnostic ability of binary classifiers. The results can be seen on Table 2.

Table 2: Analysis of anomaly detection with NB, LR, DT, and RF classification algorithms over UNSW-NB15 dataset.

Metric	NB	LR	DT	RF
CV fit time [s]	0.37489	2.12689	2.85655	59.11744
CV accuracy mean	0.79568	0.85093	0.94884	0.95991
CV precision mean	0.84303	0.83782	0.96293	0.96320
CV recall mean	0.85993	0.96845	0.96186	0.97848
CV F1 mean	0.85136	0.89841	0.96239	0.97078
CV AUC mean	0.86780	0.86922	0.94299	0.99354
Accuracy test	0.70620	0.70598	0.86123	0.87093
Precision test	0.68783	0.65981	0.82205	0.81771
Recall test	0.85396	0.96199	0.95460	0.98522
F1test	0.76195	0.78275	0.88338	0.89368
AUC test	0.79999	0.81454	0.85322	0.97730

The results given in Table 1 show that the Random Forest algorithm gives better results for each of the analyzed metrics, during the model validation and testing, compared to the results obtained for the other three algorithms (exception is Test Precision).

The following text provides a more thorough evaluation of the results for the CV Fit Time, F1 test and Recall test parameters obtained for the observed algorithms (NB, LR, DT and RF).

4.1 Analysis of CV Fit Time Metric

Figure 2 shows the CV Fit Time for each of the four algorithms (NB, LR, DT and RF). According to that, it can be seen that the learning time of Random Forest is 20-30 times longer than for the other algorithms (LR and DT). Also it can be seen that Naive Bayes classifier has very small CV Fit Time, that is 158 times faster than the one attained for the Random Forest algorithm.

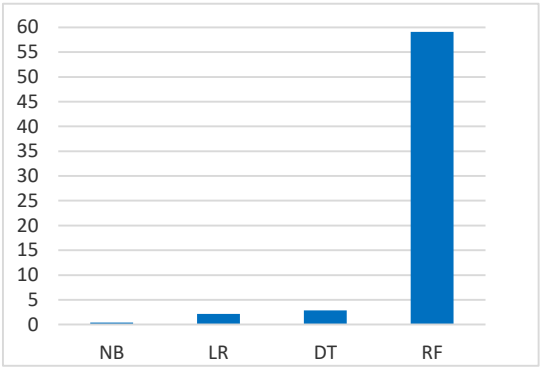


Figure 2: Comparison of CV Fit Time [sec] for NB, LR, DT and RF algorithms, implemented over UNSW-NB15 dataset.

4.2 Analysis of F1 Test Metric

The created anomaly detection model should have a relatively high coverage capability and high accuracy. Accordingly, the F1 Test is selected as the assessment metric. The F1 result can be interpreted as the average of precision and recall, where the F1 result reaches its best value when it is 1 and its worst result when it is 0. Figure 3 shows the F1 Test results for each of the four algorithms. From there it can be seen that all the algorithms (NB, LR, DT and RF) are more close to the 1, so all of them are valid and acceptable models. However, for the analyzed UNSW-NB15 dataset the Random Forrest algorithm is the best basic model for classification.

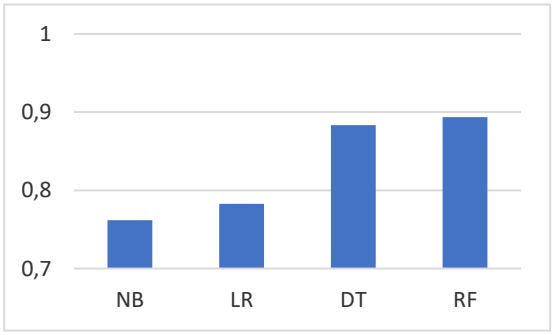


Figure 3: Comparison of F1 Test metric for NB, LR, DT and RF algorithms, implemented over UNSW-NB15 dataset.

4.3 Analysis of Recall Test Metric

When an unbalanced classification problem for anomaly detection is being analyzed, the recall metric should be observed as well. This metric is used to determine how many of the classified attacks were a real attack. Figure 4 shows the values of Recall Test metrics for each of the four algorithms. Accordingly, it can be noticed that Random Forest provides the best Recall Test result of 0.985 (i.e. 98,5%), but also Logistic Regression algorithm is very close to achieve the maximal score, given that its Recall Test result is 0,961 (i.e. 96,1%).

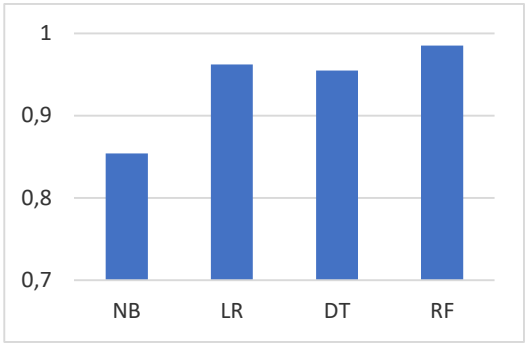


Figure 4: Comparison of Recall Test metric for NB, LR, DT and RF algorithms, implemented over UNSW-NB15 dataset.

5 CONCLUSIONS

This paper presents an implementation of IDS based on the UNSW-NB15 dataset. The dataset is trained and tested for nine class attack categories. With the application of four machine learning algorithms, Naive Bayes, Logistic Regression, Decision Tree, and Random Forrest, the UNSW-NB15 dataset has been successfully classified into network traffic of normal records and attack logs. From the analysis of the ML models for each of the methods, it was shown that the classification with Random Forrest is more successful than with Naive Bayes, Logistic Regression, and Decision Tree. According to the obtained results, the Random Forest classifier provides F1 and Recall values of 89.3% and 98.5%. The good results of Random Forrest training indicate that this algorithm requires far less need to find hyper-parameters, which are left as default. On the other hand, the Naive Bayes classifier shows the least effectiveness when applied in the UNSW-NB15 data set. In order to provide more extensive analysis, other ML classification algorithms and feature selectors could be applied to the UNSW-NB15 data set in the future.

REFERENCES

[1] L. H. Yeo, X. Che, and S. Lakkaraju, “Understanding modern intrusion detection systems: a survey,” in Cryptography and Security Journal, 2017.

[2] P. Amudha, S. Karthik, and S. Sivakumari, “Classification techniques for intrusion detection-an overview,” in International Journal of Computer Applications, vol. 76, no. 16, 2013.

[3] M. Almseidin, M. Alzubi, S. Kovacs, and M. Alkasassbeh, "Evaluation of machine learning algorithms for intrusion detection system," in Proc. of IEEE International Symposium on Intelligent Systems and Informatics, 2017.

[4] V. Golman, "An efficient hybrid intrusion detection system based on C5.0 and SVM," in International Journal of Database Theory and Application, vol. 7, no. 2, 2014, pp. 59-70.

[5] S. S. Tanpure, G. D. Patel, Z. Raja, J. Jagtap, and A. Pathan, "Intrusion detection system in data mining using hybrid approach," in International Journal of Computer Applications, 2016, pp. 0975-8887.

[6] S. A. Hajare, "Detection of network attacks using big data analysis," in International Journal on Recent and Innovation Trends in Computing and Communication, vol. 4 (5), 2016, pp. 86-88.

[7] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study," in Journal of Information Security and Applications, vol. 50, 2020.

[8] D. D. Protić, "Review of KDD CUP '99, NSL-KDD and KYOTO 2006+ datasets," in Military Technical Courier, vol. 66 (3), 2018.

[9] M. Nour, J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in Proc. of IEEE Military Communications and Information Systems Conference, 2015.

[10] S. M. Othman, N. T. Alsohybe, F. M. Ba-Alwi, and A. T. Zahary, "Survey on intrusion detection system types," in International Journal of Cyber-Security and Digital Forensics, vol. 7, no. 4, 2018, pp. 444-462.

[11] B. Caswell, J. Beale, and A. Baker: Snort Intrusion Detection and Prevention Toolkit. MA, Burlingthon: Syngress, 2007.

[12] S. Danish, A. Nasir, H. K. Qureshi, A. B. Ashfaq, S. Mumtaz, and J. Rodriguez, "Network intrusion detection system for jamming attack in LoRaWAN join procedure," in Proc. of IEEE International Conference on Communications, 2018.

[13] K. Hutchison, "Wireless intrusion detection systems," SANS Institute, White Paper, 2005.

[14] W. Stallings, "Network security essentials: applications and standards", 6th ed. USA: Pearson, 2017.

[15] S. Madhavan, "Mastering python for data science", UK: Packt Publishing, 2015.

[16] R. Ioshi, "Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures," 2016, [Online]. Available: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/.

IoT System for Monitoring Quality of Water

Marija Gjosheva, Zlate Bogoevski, Zdravko Todorov and Danijela Efnusheva
Computer Science and Engineering Department, Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University, 18 Rugjer Boshkovik Str., Skopje, R. N. Macedonia
marijagjosheva@yahoo.com, zlate1999@gmail.com, {todorovz, danijela}@feit.ukim.edu.mk

Keywords: Water Quality, Monitoring System, Internet of Things, Arduino Uno, Sensors.

Abstract: Water as a natural resource has so much impact on human lives. The diversity of water use makes this resource in constant demand, so its quality is crucial for humans. Pollution of water and many other critical natural resources is an increasingly common problem people face. That is why they have been trying to find various ways to prevent pollution. When it comes to water quality, it can be said that its traditional monitoring is very time-consuming and is also subject to irregularities. Thus, the advances in the Internet of Things (IoT) technology have provided innovative and valuable approaches for designing special purposed embedded systems. This paper proposes an IoT system design intended to monitor water quality in real-time. The prototype is based on an Arduino Uno microcontroller and several sensors to detect water parameters, such as pH-value, temperature, and turbidity. The results obtained from the proposed system are forwarded through the ThingSpeak web IoT platform and then are displayed on an Android application, which is specially developed to provide real-time monitoring. This implementation has been shown as a helpful solution for observing the water quality that people consume daily.

1 INTRODUCTION

The Internet of Things, known as the IoT, refers to millions of physical devices connected to the Internet to collect and share data. In a broader sense, it is a dynamic and global network infrastructure in which intelligent objects and entities are used in conjunction with electronics, sensors and software to improve connectivity, data collection and exchange [1]. This type of network generally has many nodes that communicate with the environment and exchange data while reacting to events, activating control operations, or changing the physical world. Because of the cheap computer chips and the ubiquity of wireless networks, it is possible for any object, regardless of its size, to be part of the IoT system if it can be connected to the Internet.

The IoT is also expanding its capacity for environmental issues [2]. Globally, a large number of deaths are caused by polluted water. The garbage that water collects during its flow has a harmful and destructive impact on vegetation and ecosystems. Therefore, ways to prevent this are being devised daily. First, checks are made to determine if the water consumed is clean and of good quality. To check this, people have designed water quality testing systems. All this is done to enable the water quality to be brought to the desired level.

The conventional method of water quality control consists of sampling the water that should be tested. These samples are then taken to a testing and analysis laboratory [3]. However, this approach is not expensive for everyday use and, because it uses time and power significantly. As a result, this paper proposes implementing a water quality monitoring system as a low-cost system, which provides water quality checks in real-time and storing the data on the cloud. This proof-of-concept implementation is based on quality control through several sensors using IoT technology. Any object that can get its IP address can be part of the proposed IoT system. Connecting objects themselves and adding sensors gives them a certain level of digital intelligence. The IoT makes the world around us more intelligent, simpler, and more responsible [4].

This paper is organized as follows: Section 2 gives an overview of different state of the art solutions. Section 3 describes the proposed IoT system for monitoring the quality of water. Section 4 presents the implemented system design and discusses the real-time monitoring results. Section 5 concludes the paper.

2 STATE OF THE ART

Although 71 per cent of the Earth is water-covered, it is slowly but surely losing its quality [5] Unfortunately, people are not aware that they are losing the most important resource on Earth. In order to eliminate such problems related to river water quality, several technologies have been created for their protection. For example, the authors of [6] researched a wireless sensor network (WSN) to collect real-time water quality parameters. Additionally, the authors of [7] proposed an online water quality monitoring system, where the information was transmitted using GPRS, and thus the water quality parameters were considered. Furthermore, the authors of [8] designed a web-based network to monitor pollution using ZigBee. Data was collected from multiple sensors, which were then routed to a web server over the WiMAX network to monitor water quality on long distances. This system was able to monitor pollution in real-time. Another approach of water quality monitoring was proposed in [9]. This system was powered by solar energy and used a remote sensor network. The base station received information from those sensors, and it was powered by solar energy (energy harvesting).

All of these systems mentioned above measure the quality of water in rivers. River water quality can be determined visually, but drinking water is not the case. So far, not many solutions have been found to measure the quality of drinking water.

IoT is very suitable for implementing a water quality monitoring system [10]. In such a scenario, a device with built-in sensors can connect to a platform that integrates data received from different sensors and then perform analysis. Such a robust IoT platform can determine which information is helpful and ignored. The information can identify patterns, make recommendations, and identify potential problems before they occur. Several proposals of embedded and IoT based solutions of water-quality monitoring systems are discussed in [10-12]. Even neural networks can be used in this kind of project. For example, the authors of [13] proposed a new water quality prediction method that was based on long short-term memory (LSTM) neural network, an artificial recurrent neural network (RNN) architecture used in deep learning. Here, a prediction model was established, and data set of water quality indicators in Taihu Lake was used as training data. A series of simulations and parameters selection was carried out. Finally, the proposed method was compared with two methods: one was based on a backpropagation neural network, and the other was based on an online sequential extreme learning machine. Another interesting approach that used a multiple linear regression model to determine the correlation among water quality parameters is given in [14].

3 PROPOSED SYSTEM DESIGN

In order to determine the quality of drinking water in real-time, a simple system has been constructed that would calculate the water quality, depending on the value of the parameters that the system itself would detect. In our proposal, not only a hardware model is developed, but also it is integrated with a web and mobile application for water quality monitoring. The proposed hardware device is connected to the cloud using a web service (ThingSpeak.com). The end-user will see the values obtained through the sensors through a developed Android application in a format understandable to him. The architecture of the proposed IoT-based system, including: hardware local unit, ThingSpeak platform and Android application, is presented in Figure 1.

The Core of the IoT-based system is the web service ThingSpeak.com. The hardware local unit detects the data received from the sensors and sends them to the Internet, to a specific channel on ThingSpeak.com. The Android application reads the same data from the status channel of ThingSpeak.com and displays them to the user in an appropriate way suitable for him. The types of data that the sensors detect are pH value, temperature and turbidity of the water. Each of theese sensors measures the value separately through a code, then sends that value through the corresponding channel on ThingSpeak. With a unique ID key, this information reaches the Android application of the end user.

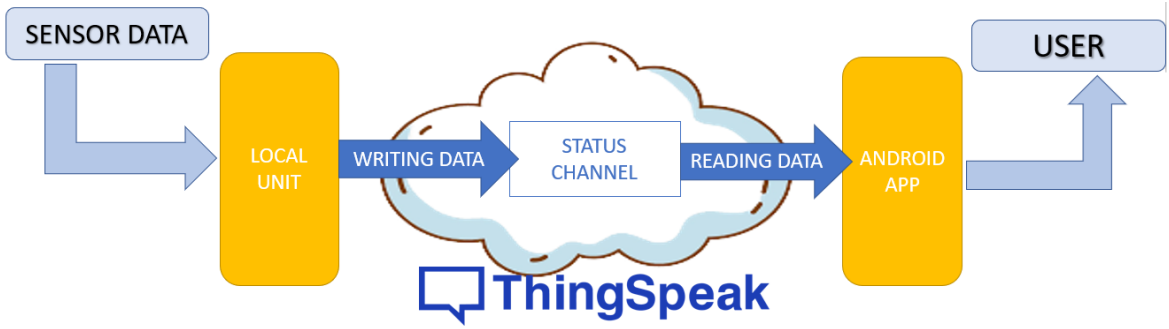


Figure 1: Proposed IoT system for monitoring quality of water.

3.1 Hardware Design

The hardware design of the local unit in the proposed system is based on a low cost Arduino Uno microcontroller board that is easily programmable and flexible [15]. It is based on the ATmega328P microcontroller. The Arduino Uno board is programmed using the Arduino IDE software which is the official software introduced by Arduino.cc. For the purpose of designing a water quality monitoring system, the Arduino Uno microcontroller board is connected to three sensors purposed to detect: pH value, temperature and water turbidity.

The pH value or hydrogen indicator is a measure of the activity of hydrogen ions in solution [16]. The pH sensor gives a result in the range from 0 to 14 pH value. Acidic solutions have a lower pH value, while alkaline ones have a value greater than 7. The natural pH value of water is about 7 (neutral point). The pH sensor can determine if the substance being tested is acidic, basic or neutral. The pH sensor is powered by 5V and easily connected to the Arduino board. Measuring the pH value can provide indications for corrosion of pipes, accumulation of solid materials etc. In the environment, a variable pH value can be an indicator of pollution. The normal range for pH value is from 6 to 8.5. If the pH value reaches above 8.5, the water is considered hard, which would probably cause damage to the pipes.

Water turbidity is a quantitative measure of the presence of particles in a liquid. It is an optical characteristic of water and measures the amount of light scattered by a material in the water when light is passed through a sample of water [17]. The turbidity sensor is quite simple, and the output is also quickly produced. The value required to calculate the turbidity is NTU (units for nephelometric turbidity). Low NTU values indicate high water clarity, while high values indicate low water clarity. There is a special connection between the voltage and the water turbidity. The graph representing this connection is given in Figure 2. The given equation of turbidity and voltage dependence is only applicable if the sensor gives a value of around 4.2V for pure water and if the output for voltage is in the range of 2.5 to 4.2V. If the correct value is not obtained during testing, a calibration is required.

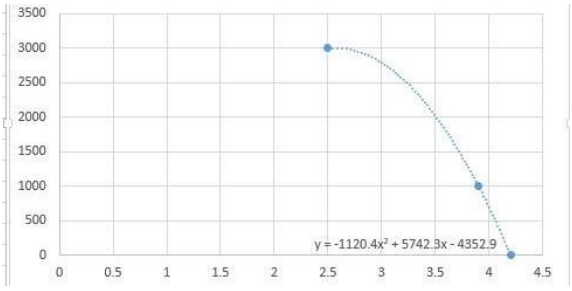


Figure 2: Voltage dependency of turbidity.

The DS18B20 temperature sensor is a convenient choice of water temperature sensor [18]. It gives results with high precision and has a fast response. Increased temperature also means increased voltage (ex. 250 mV means 25 °C.). The temperature that can be measured with the mentioned sensor is in the range of -55°C (-67°F) to 125°C (257°F).

The Wi-Fi module is a crucial part of the project development, as it is a necessary component of the proposed system that makes it an IoT product. The proposed system would not be IoT-based if the entire system were not connected to the Internet and if the obtained data was not displayed on the Internet. ESP8266 ESP-01 [19] is a low-cost Wi-Fi module that allows microcontrollers access to the Wi-Fi network. This module is placed on a miniature development board (24.8 x 14.3 mm) that contains a Wi-Fi chip with an integrated TCP/IP protocol stack. This module is a standalone SOC (system on chip)

that does not need a microcontroller to manipulate inputs and outputs as we would typically do with an Arduino since ESP-01 acts as a small computer. The ESP8266 ESP-01 module is pre-programmed with AT command set firmware, which allows it to connect directly to an Arduino device and get as much Wi-Fi connectivity as a Wi-Fi shield. Once the connection is established, the Wi-Fi is tested using AT commands. Those commands are used for controlling MODEMs.

3.2 ThingSpeak Platform

The most important part of the whole research is the web service ThingSpeak.com [20]. It is an open IoT platform that enables collection of sensor data to cloud, analysis and visualization of collected data and triggering actions according to the received data. The software is written in the Ruby programming language and allows users to communicate with devices that are connected to the Internet. This web platform facilitates data access by providing APIs to devices. HTTP and MQTT protocols are used for data transmission over the Internet.

In this research, the ThingSpeak web service is used to visually display the data received from the sensors of the implemented system for monitoring quality of water. Initially, in order to use this web service, a user profile must be created. When creating the profile, a channel is created through which the data are going to be displayed on the website. The created channel contains three fields for the three sensors (pH value, temperature and water turbidity) and additionally one field which is empty, but is created for the purpose of future expansion of the project and its debugging.

When creating a channel, a special key (API key) is generated, which is the channel identification code. This API code can be used to access the channel in order to display the values obtained during the measurement. Initially, the sensors in the project record the values locally. The next step would be to transfer these values to ThinkSpeak.com. In order for the results to reach the appropriate ThingSpeak.com channel, a TCP connection must be created (via its IP address). The string of values, which will be sent through a successfully established TCP connection, will be graphically displayed.

3.3 Android Mobile Application

The water quality monitoring system displays the specific parameters important for measuring water quality. In the previous part of the research, the obtained values were displayed on the monitor in the Arduino IDE and on the website ThingSpeak. However, these values are not correctly presented so that everyone would understand. That is why an Android application has been designed.

The Android application consists of several cards. Each of the cards represents a different parameter measurement. When a measurement is made, and its values are displayed on the web service, a card is also created to display them on the Android application. Figure 3 shows one of the cards representing values of a measurement that do not meet the standards for clean water.

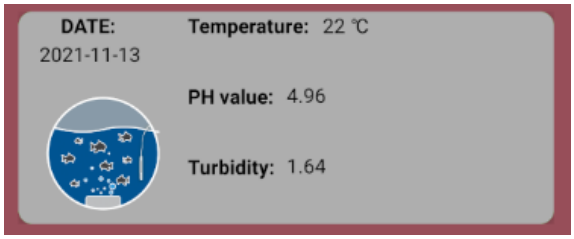


Figure 3: Card layout for displaying result values.

Figure 3 also shows the date when the measurement was made. The layout is quite simple but still meets the required functionality. The values of each measurement are displayed in separate fields. The turbidity value depends on the sensor data obtained in volts. The background of the card itself depends on the measured results. A blue background will appear if the pH value is greater than 6.5 and the results obtained from the turbidity sensor are greater than 2 NTU. For all other possible scenarios, the data will be red. Does the question arise in what way the values are obtained? As already mentioned, each channel in ThingSpeak has its key to access that channel. The application uses the same key to access ThingSpeak.com. The channel is accessed via a URL string where a key is specified[1].

The result of this would be a JSON string that contains the parameters of all the measurements along with the time and date when they were made. The string looks like this:

{"channel":{"id":1529598,"name":"Water Quality Monitoring System","latitude":"0.0","longitude":"0.0","field1":" Actuator 1","field2":"Actuator 2","field3":"Temperature","field4":"Turbidity","field5

[1]https://api.thingspeak.com/channels/1529598/feeds.json?api_key=IWRDIXMCOFUGBU1S

":"PH","field6":"Spare","created_at":"2021-10-07T18:00:01Z","updated_at":"2021-10-07T18:00:01Z","last_entry_id":8},"feeds":[{"created_at":"2021-10-28T17:41:33Z","entry_id":1"field3":"25","field4":"0","field5":"6.67","field6":"0"},{"created_at":"2021-10-28T17:44:14Z","entry_id":2"field3":"25","field4":"1.59","field5":"6.59","field6":"0"}]}

This is actually the string that the application gets, but it filters it in a special way in order to present the values more clearly to the end user.

4 PROPOSED SYSTEM IMPLEMENTATION

All components of the monitoring system that were previously discussed are complete. The application is designed, the code is written and tested, and the hardware is connected. The next step is to connect all the parts correctly and check if the system operates as expected and if the obtained results are acceptable. The test was performed on drinking water, with samples taken from various sources.

The hardware components are correctly connected so that they can get a value for the appropriate parameter, hopefully in an acceptable range. When designing the device, the focus is on the design of the end-user application.

The microcontroller is powered via a USB cable connected to a computer, but of course, there is the possibility of developing and expanding this device for further power supply from a battery or solar energy through a solar collector. The critical element for Internet access, the Wi-Fi module, is powered in the same way. Figure 4 shows the fully connected system that measures the parameters.

Once the first step in the final test is completed, i.e. the components are connected, it is necessary to execute the appropriate code in order to be able to read and display the values obtained during the testing of the water sample. Concrete testing is performed on relatively clean water, as expected that the values obtained will be minimal below the normal clean/water limits. The corresponding test results displayed on the Arduino IDE serial monitor are shown in Figure 5.

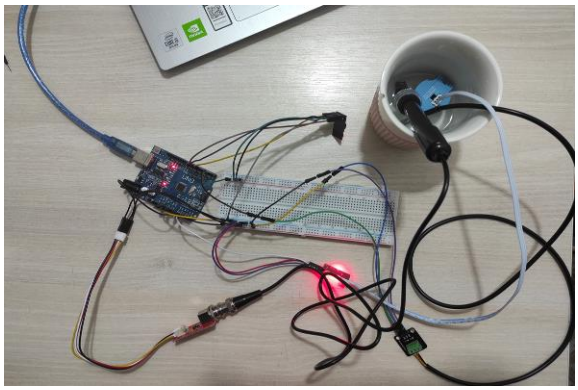


Figure 4: Implementation of the proposed IoT system for monitoring quality of water.

From the data shown in Figure 5, it can be seen that the sample on which the test was performed is relatively but not completely pure water. Values are expected - turbidity is 3.74 NTU, the temperature is 19 degrees, and the pH value is 5.36. When sending the sensor data, a TCP connection with ThingSpeak.com has opened automatically, and the values are displayed on the website in the appropriate field in the created channel. The sensor data values are represented by their size and the time the test was performed. These results are shown in Figure 6.

Finally, it is expected that these same data will be appropriately displayed on the card in the designed Android application. According to the JSON_URL that the application receives, the appropriate data of the measurement are filtered. Figure 7 shows that the results in the Android application are displayed exactly as expected. The parameters are written in a way that is understandable to everyone, and the measurement date is also given.

The results obtained when testing pure water show that the tested water is suitable for consumption, but still, more tests are needed to determine further the validity of the data and the system's operation before it can be deployed elsewhere. However, the prototype does not have many capabilities and therefore requires upgrades and extensions; to be able to transfer data over a wireless network to a remote laptop or mobile phone at any given time and location, and a stronger memory drive or the ability to store data in databases.

```
Reseting.......
RESET
Turbidity : 3.74
Temperature : 19 C
PH value :  5.36
Send ==> Start cmd: AT+CIPSTART="TCP","184.106.153.149",80
Send ==> lenght cmd: AT+CIPSEND=83
Send ==> getStr: GET /update?api_key=XJ4UOKUXXX3MAW6E&field3=19&field4=3.74&field5=5.36&field6=0
```

Figure 5: Results from testing, shown in Arduino IDE serial editor.

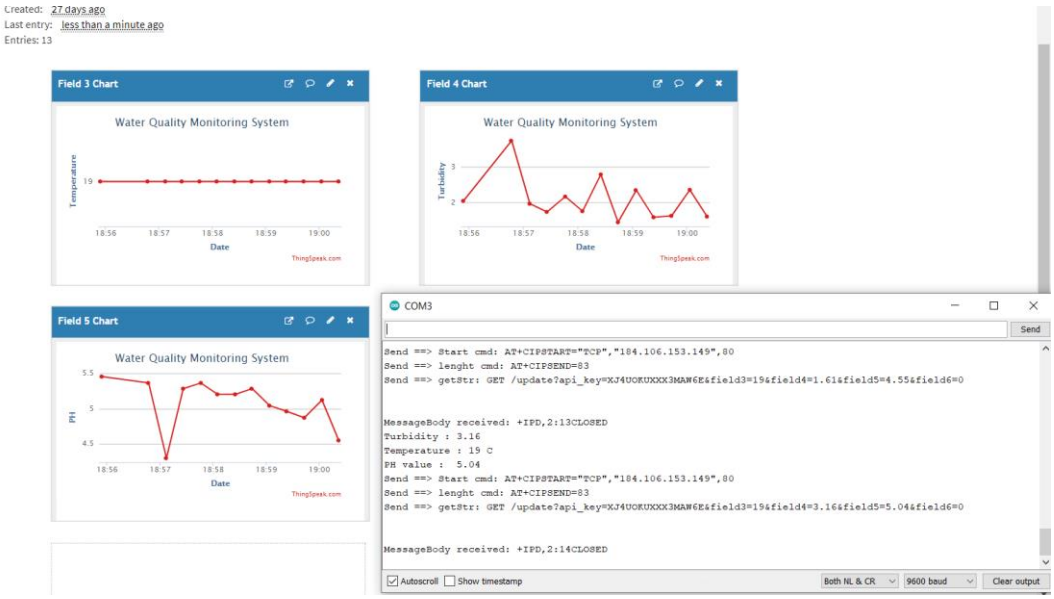


Figure 6: Results from testing, shown in ThingSpeak.com.

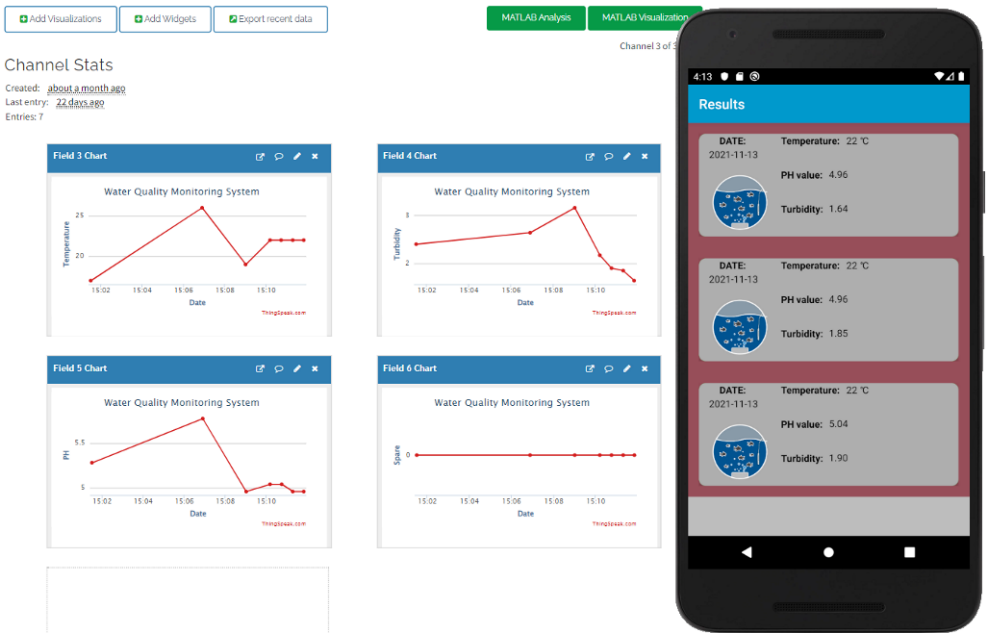


Figure 7: Results from testing, shown in Android mobile application.

According to that, an SD card could be used to allow the storage of quantities of Arduino data. The combined central system must be lined to prevent potential harmful defects from weather exposure. Regarding the sensors, the blur sensor must be upgraded or completely replaced. The opening at the top of the sensor for turbidity allows water to enter if the sensor is lowered too deep into the stream, and the inflow causes direct distortion in the readings. Its short cable means it is also strictly limited in use. A suitable replacement may be the Relihones digital blur sensor which is waterproof and offers a longer cable. Implementation of additional sensors in the project, thus its expansion will allow monitoring of additional parameters, especially the concentration of different ions.

Finally, changes may be made to the environment and the test schedule. Conducting the test of different liquids and more frequent reading of the data should provide much needed additional data and greater accuracy in determining the purity of the water and the validation of the monitoring system.

5 CONCLUSION

A fast, efficient, low-cost monitoring system is implemented and tested. The proposed system measures water quality in real-time and does not require people on duty. With this system, water testing is more economical, convenient, and faster. Although there is no need for external intervention, officials can still use it. They can consider the level of pollution that occurs in the water and transmit the warnings directly to the public. This can help prevent diseases caused by polluted waters and the presence of metals in them.

The proposed monitoring system for water quality is applicable when quick and effective actions are required to prevent extreme pollution levels. Indeed, the proposed system is easy to install and easily placed close to the target area. Monitoring can be performed even by less trained people. Approximately 20 euros were spent on the overall design of this device. It is also flexible and extensible. Given the low cost of this device, any household can afford it. By replacing the appropriate sensors and changing software programs, this system can monitor other water parameters. The monitoring time interval can also be changed as needed.

The current implementation of the proposed system has a wide application for meager costs. It provides real-time monitoring of water quality, which is way more efficient than traditional. What is of most importance, in this research, the ecological environment of the water resources is protected.

REFERENCES

[1] F. DaCosta, Rethinking the Internet of Things, Apress, 2013, pp.12-122.

[2] R. Liu, P. Gailhofer, C. Gensch, A. Köhler, and F. Wolff “Impacts of the digital transformation on the environment and sustainability,” Technical Paper, Berlin, 2019.

[3] UN/ECE Task Force on Laboratory Quality Management & Accreditation, “Guidance to operation of water quality laboratories,” Technical Report, 2002.

[4] J. Salazar, and S. Silvestre, Internet of Things, 1st Edition, Czech Republic: Techpedia, 2017.

[5] J. Mateo-Sagasta, S. Marjani Zadeh, and H. Turral, “Water pollution from agriculture: a global review,” Technical Paper, 2017.

[6] T. Le Dinh, W. Hu, P. Sikka, P. Corke, L. Overs, and S. Brosman, “Design and deployment of a remote robust sensor network: experiences from outdoor water,” Proc. of 32nd IEEE Conf. on Local Computers, pp 799-806, 2007.

[7] Q. Tie-Zhn, and S. Le, “The design of multiparameter on line monitoring system of water quality based on GPRS,” Proc. of IEEE International Conference on Multimedia Technology, China, 2010.

[8] S. Silva, H. N. Nguyen, V. Tiporlini, and K. Alameh, “Web based water quality monitoring with sensor network: employing ZigBee and WiMAX technology,” Proc. of 36th IEEE Conf. on Local Computer Networks, 2011.

[9] M. K. Amruta, and M. T. Satish, “Solar powered water quality monitoring system using wireless sensor network,” Proc. of IEEE Conf. on Automation, Computing, communication, control, and compressed sensing, pp. 281-285, 2013.

[10] S. S. Babu, "Water Quality Monitoring and Filter System to Preserve Water Resource Using IOT", Bangalore, India , July 2020.

[11] M. O. Faruq, I. H. Emu, M. N. Haque, M. Dey, N. K. Das, and M. Dey, “Design and implementation of cost-effective water quality evaluation system,” in Proc. of IEEE Region 10 Humanitarian Technology Conference, pp. 860-863, 2017.

[12] Y. K. Taru, and A. Karwankar, “Water monitoring system using arduino with Labview,” in Proc. of IEEE International Conference on Computing Methodologies and Communication, 2018.

[13] Y. Wang, J. Zhou, K. Chen, Y. Wang, and L. Liu, “Water quality prediction method based on LSTM neural network,” in Proc. of IEEE International Conference on Intelligent Systems and Knowledge Engineering, pp. 1-5, 2017.

[14] K. S. D. Krishnan, and P. T. V. Bhuvaneswari, “Multiple linear regression based water quality parameter modeling to detect hexavalent chromium in drinking water,” in Proc. of IEEE International Conference on Wireless

Communications, Signal Processing and Networking, pp. 2434-2439, 2017.

[15] Arduino, "Arduino Uno," DataSheet, 2021.

[16] Sensores, "Types of pH sensors: what you need to know," 2019. [Online]. Available: https://sensorex.com/blog/2019/09/09/ph-sensors-need-to-know/, last accessed 2021/12/06.

[17] "Turbidity Sensors," 2021. [Online]. Available: https:// www.campbellsci.eu/turbidity, last accessed 2021/12/06.

[18] DS18B20, DataSheet, 2019.

[19] "Getting started with the ESP8266 ESP-01," 2021. [Online]. Available: https://www.instructables.com/ Getting-Started-With-the-ESP8266-ESP-01/, last accessed 2021/12/06.

[20] "ThingSpeak for IoT projects," 2021. [Online]. Available: https://thingspeak.com/, last accessed 2021/12/06.

I 3D3P: an Intelligent 3D Protein Prediction Platform

Mohamed Hachem Kermani[1] and Zizette Boufaida[2]
[1]*LIRE Laboratory, National Polytechnic School - Malek Bennabi, Constantine, Algeria*
[2]*LIRE Laboratory, University of Constantine 2 - Abdelhamid Mehri, Constantine, Algeria*
hachem.kermani@enp-constantine.dz, {hachem.kermani, zizette.boufaida}@univ-constantine2.dz

Keywords: Computational Biology, 3D Protein Structure, Protein Structure Prediction, Multiple Sequence Alignment, Machine Learning, Intelligent Platform.

Abstract: Proteins are macromolecules consisting of a chain of smaller molecules (i.e. amino acids) known as monomers. Three levels of protein structure are distinguished: primary, secondary and tertiary. Determining the three-dimensional (3D) structure of a protein when only a sequence of amino acids is given, is one of the most important and frequently studied issues in bioinformatics and computational biology. Therefore, in this paper, we propose an Intelligent 3D Protein Prediction Platform, which aims to completely determine the tertiary protein structure of a given protein primary structure (i.e. the amino acid sequence). The proposed intelligent platform is based on multiple sequence alignment and machine learning techniques to predict automatically 3D protein structures. We also present a software application and an experiment of the proposed platform, which will be used by experts for a better understanding of protein functions and activities in order to develop effective mechanisms for disease prevention, personalized medicine and treatments and other healthcare aspects.

1 INTRODUCTION

Proteins are vital molecules that play many important roles in the human body; they contribute to the tissue growth and maintenance, the catalysis of organic reactions, the communication between cells, tissues and organs and help improve immune health. Each protein is a macro-molecule consisting of a chain of amino acids, which are assembled through peptide bonds (i.e. an amino acid group of carboxylic acid with a neighboring amino acid group and thus form the primary structure [1]. Then comes secondary protein structure which is the three-dimensional form of local protein segments. Alpha helices and beta sheets are the two most common secondary structural elements, which form spontaneously as an intermediate before the protein folds into the tertiary three-dimensional structure where the α-helixes and β-pleated-sheets are folded into a compact globular structure. Some proteins, known as oligurics (i.e. made up of several polypeptide chains, each chain has a primary, secondary and tertiary structure), such as hemoglobin, reach a quaternary structure by adopting a symmetrical structure [2]. Many computational methodologies and algorithms have been proposed as a solution to the 3D Protein Structure Prediction problem, including comparative modeling methods and sequence alignment strategies, deterministic computational techniques, optimization techniques, data mining and machine learning approaches [3]. In our case we combine both sequence alignment and machine learning techniques to automatically predict 3D protein structure of a given amino acid sequence. Furthermore, the proposed intelligent platform provides experts with all information needed for a deeper understanding of proteins functions and activities. The rest of this paper is organized according to the following. Section 2 provides an overview of research that is related to our approach. Section 3 presents our proposal which is the Intelligent 3D Protein Prediction Platform. Section 4 presents a software application and experimentation. Section 5 presents a discussion. Finally, Section 6 concludes the paper and suggests some directions for future research.

2 RELATED WORK

X-ray crystallography, which is a time-consuming and relatively expensive method, has determined most of the protein structures available in the Protein Data Bank [4]. Hence computational methods have been

developed to compute and predict protein structures based on their sequences of amino acids.

2.1 X-ray Crystallography Method

X-ray crystallography is a technique for determining the structure of molecules in three dimensions, including complex biological macromolecules such as proteins and nucleic acids. It is a powerful method at atomic resolution in elucidating the three-dimensional structure of a molecule. The X-ray crystallography technique uses diffraction patterns that are generated by irradiating a crystalline sample of the molecule of interest with X-rays, making diffraction quality crystals mandatory for this process [5].

Although this method provides a powerful tool in elucidating the three-dimensional structure, the major drawback is time. Thousands of experiments on crystallization can be performed daily in a single laboratory, each experiment is observed over time, with the normal time span being weeks to months [6]. It was for this reason that computational methods were developed to reduce time and costs.

2.2 Computational Methods

Proteins fold into one or more specific conformations to exercise their biological functions [7]. The Determination of a protein's structure can be achieved through computational techniques that automatically predict protein structures based on their amino acid sequences. The three common bioinformatics methods used to predict the protein structure are: comparative modeling, fold recognition and ab initio prediction.

2.2.1 Comparative Modeling

Also known as homology modeling, it is a technique which uses known information from one or more homologous partners to predict the structure of an unknown protein. Comparative modeling usually involves three steps: a) identifying template structures for modeling the query protein, b) aligning the template with the query sequence, and c) modeling the query structure [8]. This family of methods enables greater number of potential templates to be produced and better templates to be identified [9]. To predict the three-dimensional protein, both the template and the query can be submitted to a comparative modeling program once the better template has been identified.

2.2.2 Fold Recognition

We model the proteins in fold recognition by threading which have the same fold as the proteins of known structures. Protein threading is used for protein that is not stored in the Protein Data Bank (PDB) with its homologous protein structures [10]. Many algorithms for determining the correct threading of a sequence into a structure have been proposed. They employ some form of dynamic programming. The problem of identifying the best alignment for the complete 3D threading is very difficult (it is an NP-hard issue for some threading models) [11]. Researchers have therefore proposed many methods of optimization, such as Conditional random fields, simulated annealing, branch and bound and linear programming, in order to achieve heuristic solutions.

2.2.3 Ab-initio Prediction

The ab-initio method is a technique that attempts to predict protein structures based solely on information about sequences and without using templates. Ab-initio modelling is often referred to as de-novo modeling [12]. The fundamental procedure followed by the protein structure prediction ab-initio method begins with the primary amino acid sequence, which is searched for the various conformations which lead to the prediction of native folds [13]. After recognition and prediction of the folds, the model assessment is carried out to verify the quality of the predicted structure.

Numerous methods of predicting protein structures, including X-ray crystallography, and computational methods are currently being used [14]. Each method has advantages and disadvantages, but they all have the same goal of building a consistent 3D protein model that can be useful for a detailed understanding of protein and enzyme function.

3 I 3D3P

Proteins, consisting of long or short amino acid sequences, respectively called polypeptides and peptides, are assembled from amino acids based on the information contained in the genes[15]. Protein synthesis is the process in which cells produce proteins by determining a protein's various structures: primary, secondary, and tertiary. The proposed Intelligent 3D Protein Prediction Platform (I 3D3P) aims to determine the three-dimensional protein structure from a given amino acid sequence, based on a multi-

ple sequence alignment technique or a machine learning method.

First the platform compare the amino acid sequence introduced by the user with all amino acid sequences of already known proteins existing on the available protein sources, then based on the results of this comparison, the platform predict the 3D protein structure as illustrated in Figure 1.

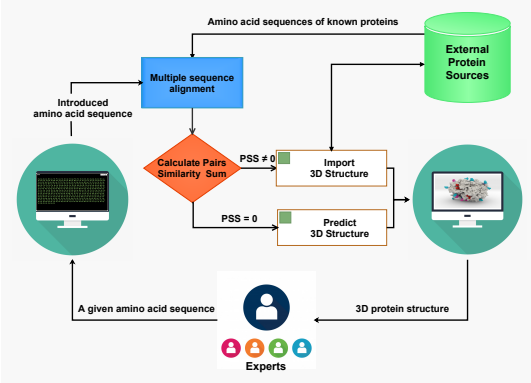


Figure 1: The 3D protein prediction process.

The proposed platform enables to predict the 3D protein structure from a given amino acid sequence based on two techniques, 3D structures importing from the available protein sources or the automatic prediction of 3D structures.

3.1 Importing 3D Structures

This technique consists of comparing the given amino acid sequence with all sequences of the already known proteins available in different protein sources in order to import the corresponding 3D protein structure. Therefore, we developed a multiple sequence alignment technique to compare the given sequence with all known proteins available in the existing protein sources.

The proposed sequence alignment technique consists to recuperate all protein sequences from the available sources and align them with the given sequence. This alignment creates a similarity score matrix between the given amino acid sequence, denoted below as X and all known protein sequences, denoted as Seq1, Seq2,...Seq$_n$

	X	Seq1	Seq2	Seq3	Seq$_n$
X	Ss = 1	Ss (X, Seq1)	Ss (X, Seq2)	Ss (X, Seq3)	Ss (X, Seq$_n$)
Seq1	Ss (Seq1, X)	Ss = 1	Null	Null	Null
Seq2	Ss (Seq2, X)	Null	Ss = 1	Null	Null
Seq3	Ss (Seq3, X)	Null	Null	Ss = 1	Null
Seq$_n$	Ss (Seq$_n$, X)	Null	Null	Null	Ss = 1

The pairwise similarity score between X and Seq1, Seq2, Seq3, ...Seq$_n$ depends on the similarities and dissimilarities between the amino acids in each sequence position. A correspondence between the amino acids is counted as 1, C = 1, and a dissimilarity or a gap in the case of local alignment is counted as 0, D = 0. The pairwise similarity score is calculated as follows:

$$Ss(X, Seq_n) = \frac{\sum C, D}{NAA}. \quad (1)$$

Where C and D represent the similarities and dissimilarities between the amino acids and NAA represents the number of amino acids constituting the sequence, as illustrated in the following examples:

Example 1:

X:	Lys	-	Glu	-	Thr	-	Lys
Seq1:	Lys	-	Glu	-	Thr	-	Lys
	1		1		1		1

$$Ss(X, Seq1) = \frac{\sum C, D}{NAA} = \frac{4}{4} = 1 \quad 100\,\% \quad (2)$$

Example 2:

X:	Lys	-	Glu	-	Thr	-	Lys
Seq2:	Thr	-	Glu	-	Thr	-	_
	0		1		1		0

$$Ss(X, Seq2) = \frac{\sum C, D}{NAA} = \frac{2}{4} = 0.5 \quad 50\,\% \quad (3)$$

Example 3:

X:	Lys	-	Glu	-	Thr
Seq3:	Thr	-	Lys	-	Glu
	0		0		0

$$Ss(X, Seq3) = \frac{\sum C, D}{NAA} = \frac{0}{3} = 0 \quad 0\,\% \quad (4)$$

The calculation of all the pairwise similarity scores will enable to get the following similarity score matrix.

Sequences	Seq1	Seq2	Seq3	Seq$_n$
X	Ss (X, Seq1) = 1	Ss (X, Seq2) = 0.5	Ss (X, Seq3) = 0	Ss (X, Seq$_n$) = 0.2

Based on this similarity score matrix we can calculate the Pairs Similarity Sum as below:

$$PSS(X, Seq_n) = \sum Ss(X, Seq_n) \quad (5)$$

The multiple alignment results will be one of the following cases:

1) PSS AX, Seq$_n$) $\neq$0 : The given sequence matches perfectly a sequence of a known protein and/or matches partially some known proteins. In this case, we will have two different situations based on the similarity score of each pair:
   - Ss (X, Seq$_n$) = 1 : The given sequence matches perfectly a sequence of a known protein. In this case, the platform will import the 3D protein structures in order to provide it to experts.
   - $0 < Ss < 1$: The given sequence partially matches some known proteins. The platform will import all the partially similar 3D protein structures and display them for experts.
2) PSS (X, Seq$_n$) = 0 : The given sequence does not match any known protein. In that case, the Intelligent Platform will automatically predict the 3D protein structure based on a machine learning technique.

3.2 The Automatic Prediction of 3D Structures

The inference of a protein's three-dimensional structure from its amino acid sequence remains an extremely difficult and unsolved task. A prediction for proteins consists of assigning regions of the amino acid sequence to be probable alpha helices, beta strands. Different methods for predicting 3D structures have been developed. One of the first algorithms was the Chou-Fasman method [16, 17], which relies mainly on the probability parameters determined from the relative frequencies of the appearance of each amino acid in each type of secondary structure [18].

In addition, overtime computational prediction methods were developed which are based on techniques of sequence alignment and methods of machine / deep learning. In our case, we propose a machine learning technique in order to automatically predict 3D protein structures. This technique is a work in progress and will be presented in our future work.

4 SOFTWARE APPLICATION AND EXPERIMENT

In this section, we present a software application and an experiment of the proposed platform. To illustrate our proposed I3D3P we developed the software application with Java programming language (see Figure 2).

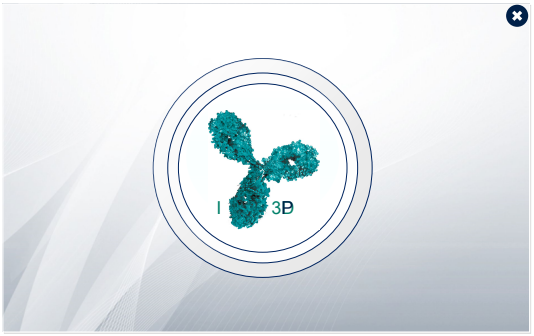


Figure 2: The Intelligent 3D Protein Prediction Platform.

I3D3P is an intelligent platform enabling experts to introduce a given amino acid sequence in order to get its 3D structure. Users can browse the file of an amino acid sequence as presented in Figure 3.

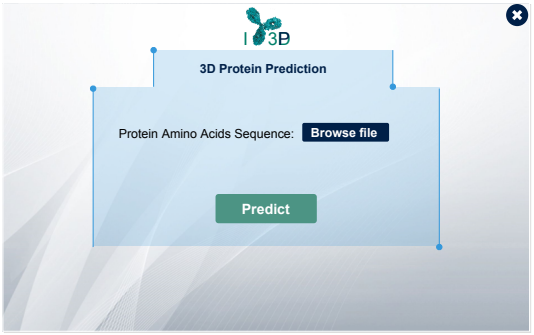


Figure 3: Browsing a given amino acid sequence.

According to our proposed intelligent platform, the prediction of the 3D protein structure depends on the results of the multiple sequence alignment. Meanwhile, the software application illustrates the case where the given amino acid sequence perfectly matches an amino acid sequence of a known protein.

Figure 4 shows that the given amino acid sequence is similar to the "Glycated Hemoglobin" amino acid sequence. In this case the 3D file will be imported from the external protein source, then displayed to allow the 3D protein visualizing, which will enable experts getting all information needed for a better understanding of protein functions and activities (Figure 5).

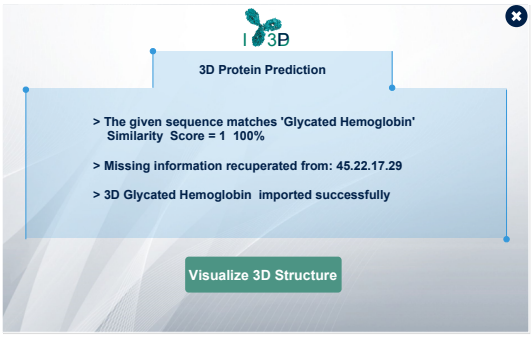


Figure 4: Importing 3D protein structure.

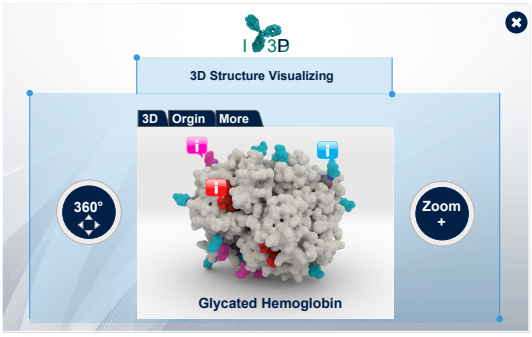


Figure 5: 3D protein visualizing.

5 DISCUSSION

Each method used to determine protein structures, including X-ray crystallography, and computational methods has advantages and disadvantages. The major drawback of X-ray crystallography is that method only aims to determine the 3D structure. In addition, laboratory experimentation with this method is time-consuming and requires days or weeks before the dynamic behavior or the expected results can be observed [19]. Instead, computational methods were developed with the aim of reducing time-consuming and getting faster results[20]. Therefore, we propose an Intelligent 3D Protein Prediction Platform, which is a computational tool that aims to determine 3D protein structures in a faster way. The proposed intelligent platform is based on computational methods by combining sequence alignment and machine learning techniques. The I3D3P will enable information on protein structures to be obtained altogether, which will allow a better understanding of protein functions and activities.

6 CONCLUSION

Protein structure prediction is the inference of a protein's three-dimensional structure from its amino acid sequence, – i.e., the prediction of its folding and tertiary structure from its primary structure. Determining a protein's 3D structure gives us a lot of information about how it operates, which we can use to control or modify its function, predict what molecules attach to it and understand diverse biological interactions. Therefore, protein structure prediction has been an important open research problem for more than 50 years. As a result, several approaches and techniques, including X-ray crystallography and computational methods, have been proposed as 3D protein structure prediction solutions. Our proposed platform is based on computational methods that combine multiple sequence alignment and machine learning techniques to predict 3D protein structures from a given amino acid sequence. I3D3P enabled the identification of 3D protein structures, allowing all protein information to be available at the three different structures. These information will be used to gain a better understanding of the proteins' functions and activities in order to develop effective mechanisms for disease prevention, personalised medicine and treatments, and other aspects of healthcare. However, our proposal has some drawbacks. We intend to address these issues in the future by presenting a machine learning method for 3D structure prediction, as our platform currently relies solely on multiple sequence alignment to predict 3D protein structures.

ACKNOWLEDGEMENTS

The authors acknowledge support from the General Directorate of Scientific Research and Technological Development (DGRSDT), Ministry of Higher Education and Scientific Research, Algeria.

REFERENCES

[1] D. T. Haynie and B. Xue, "Superdomains in the pro-tein structure hierarchy: The case of ptp-c2," Protein Science, vol. 24, no. 5, pp. 874–882, 2015.

[2] I. Kumari, P. Sandhu, M. Ahmed, and Y. Akhter, "Molecular dynamics simulations, challenges and op-portunities: a biologist's prospective," Current Pro-tein and Peptide Science, vol. 18, no. 11, pp. 1163–1179, 2017.

[3] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A.Žıdek, A. Potapenko et al., "Highly accurate protein structure prediction with alphafold," Nature, vol. 596, no. 7873, pp. 583–589, 2021.

[4] D. S. Goodsell, S. Dutta, C. Zardecki, M. Voigt, H. M. Berman, and S. K. Burley, "The rcsb pdb "molecule of

the month": inspiring a molecular view of biology," PLoS biology, vol. 13, no. 5, 2015.

[5] J. A. Brito and M. Archer, "X-ray crystallography," in Practical Approaches to Biological Inorganic Chemistry. Elsevier, 2013, pp. 217-255.

[6] M. H. Kermani, Z. Guessoum, and Z. Boufaida, "A two-step methodology for dynamic construction of a protein ontology." IAENG International Journal of Computer Science, vol. 46, no. 1, 2019.

[7] Y. Zhang, "I-tasser server for protein 3d structure prediction," BMC bioinformatics, vol. 9, no. 1, pp. 1-8, 2008.

[8] S. D. Lam, S. Das, I. Sillitoe, and C. Orengo, "An overview of comparative modelling and resources dedicated to large-scale modelling of genome se-quences," Acta Crystallographica Section D: Struc-tural Biology, vol. 73, no. 8, pp. 628-640, 2017.

[9] B. John and A. Sali, "Comparative protein structure modeling by iterative alignment, model building and model assessment," Nucleic acids research, vol. 31, no. 14, pp. 3982-3992, 2003.

[10] L. A. Kelley, "Fold recognition," in From Protein Structure to Function with Bioinformatics. Springer, 2009, pp. 27-55.

[11] T. Jo, J. Hou, J. Eickholt, and J. Cheng, "Improving protein fold recognition by deep learning networks," Scientific reports, vol. 5, p. 17573, 2015.

[12] D. Xu, L. Jaroszewski, Z. Li, and A. Godzik, "Aida: ab initio domain assembly for automated multi-domain protein structure prediction and domain–domain interaction prediction," Bioinformatics, vol. 31, no. 13, pp. 2098-2105, 2015.

[13] R. Townshend, R. Bedi, P. Suriana, and R. Dror, "End-to-end learning on 3d protein structure for interface prediction," Advances in Neural Information Process-ing Systems, vol. 32, pp. 15 642-15 651, 2019.

[14] M. H. Kermani and Z. Boufaida, "A2pf: An automatic protein production framework," in International Conference on Intelligent Systems Design and Appli-cations. Springer, 2020, pp. 80-91.

[15] M. H. Kermani, Z. Boufaida, S. Benredjem, and A. N. Saker, "An mvc-inspired approach for an intelligent annotation of a protein ontology : Iapronto," Inter-national Journal of Computer Information Systems and Industrial Management Applications, vol. 13, no. 2021, pp. 308-318, 2021.

[16] P. Y. Chou and G. D. Fasman, "Prediction of protein conformation," Biochemistry, vol. 13, no. 2, pp. 222–245, 1974.

[17] P. Y. Chou and G. D. Fasman, "Empirical predictions of protein conforma-tion," Annual review of biochemistry, vol. 47, no. 1, pp. 251-276, 1978.

[18] P. Y. Chou, "Prediction of the secondary structure of proteins from their amino acid sequence," Advances in enzymology and related areas of molecular biology, vol. 47, pp. 45-148, 1978.

[19] J. Jiménez, M. Skalic, G. Martinez-Rosell, and G. De Fabritiis, "K deep: protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks," Journal of chemical information and mod-eling, vol. 58, no. 2, pp. 287-296, 2018.

[20] M. H. Kermani and Z. Boufaida, "A state of art on bi-ological systems modeling," in 2016 IEEE Intl Con-ference on Computational Science and Engineering (CSE) and IEEE Intl Conference on Embedded and Ubiquitous Computing (EUC) and 15th Intl Sympo-sium on Distributed Computing and Applications for Business Engineering (DCABES). IEEE, 2016, pp. 712-715.

Linguistic Difference of Human-Human and Human-Chatbot Dialogues about COVID-19 in the Russian Language

Aleksandr Perevalov[1], Aleksandr Vysokov[2] and Andreas Both[1]
[1]*Anhalt University of Applied Sciences, 55 Bernburger Str., Köthen, Germany*
[2]*Perm National Research Polytechnic University, 29 Komsomolsky avenue, Perm, Russia*
{*aleksandr.perevalov, andreas.both*}*@hs-anhalt.de, sshvsk@mail.ru*

Keywords: Quantitative Text Analysis, Human-Chatbot Interaction, Human-Computer Interaction, Chatbot Interaction Design, Language Features, Wizard of Oz, COVID-19.

Abstract: This work describes the quantitative analysis of the linguistic difference in human-human and human-chatbot dialogues. The research is based on conducting a set of experiments where respondents communicate with a human or a chatbot in the domain of COVID-19 questions. In the case of the human-human dialogues, the approach of the inverted "Wizard of Oz" experimental setting is used. During the experiments, 35 human-human and 68 human-chatbot dialogues in Russian language were performed. The dialogues were collected during 4 months and thereafter analyzed with a set of quantitative text measures such as descriptive statistics of a text, syntactic complexity, lexical density, and readability. As a result, a set of measures demonstrated a statistically significant linguistic difference between the language structure of questions that were asked to the human and to the chatbot. Specifically, respondents were using shorter sentences and words, simpler syntax while communicating with the chatbot. Moreover, lexical richness of the human-chatbot dialogue data is lower while the readability is higher – these markers indicate that humans use simpler language constructions while speaking with a chatbot.

1 INTRODUCTION

The use of virtual assistants such as chatbots or question answering systems is a fairly relevant topic and develops every year more and more[1]. In today's highly competitive realities, chatbots bring real benefits to society [1, 2, 3]. A chatbot responds instantly in comparison to a human, making it less likely that the user will leave without getting a response and also simplifies real-world problems (e.g., customer service, corporate information search) by doing a lot of routine work.

Researchers and developers strive to bring their chatbot products to such an extent that communication with them could not be distinguished from communication with a real human in the context of fulfilling information needs. However, chatbots may not always sufficiently cover all the data and knowledge which is necessary for successful communication. The process of interaction between a human and a chatbot also has social and psychological aspects, as a result of which a person can somehow change his behavior and adjust his way of communication to a chatbot [4].

This work analyzes the linguistic difference between human-human and human-chatbot dialogues by using the method of quantitative text analysis [5]. To model a real-world scenario, the emergent knowledge domain of COVID-19 pandemic was selected[2]. Thereafter, a chatbot capable of answering frequently asked questions (FAQ) based on public data provided by government was developed. The working language of the chatbot and therefore all the collected data is the Russian language. During the experiment, the respondents were randomly distributed to one of the two groups: Group_1 – respondents communicate with a human-expert, Group_2 – respondents communicate with the developed FAQ chatbot. After that, the corresponding dialogues were collected in the textual form and analyzed using different measures, such as syntactic complexity, readability, lexical diversity and other descriptive statistics (see Section

[1]https://www.gartner.com/smarterwithgartner/chatbots-will-appeal-to-modern-workers

[2]It is worth underlining, that the subject of this study is not to analyze what questions were asked, but how they were asked.

2.2 for details). The experimental results highlight significant difference between the human-human and human-chatbot dialogues w.r.t. the considered measures. Based on the results, the practical recommendations on the chatbots' development and design process were formulated.

In this work, the following research questions are answered: RQ_1 – Are human-human and human-chatbot dialogues different from each other in terms of quantitative textual measures, RQ_2 – If the first is true, what measures distinguish human-human and human-chatbot conversations, and RQ_3 – What recommendations to the chatbot developers can be created in this regard? The scientific novelty of the work consists of the following: a dialogue dataset based on the topic of consulting people on COVID-19 FAQ questions was collected[3]. The dataset consists of human-human and human-chatbot dialogues, (2) the collected data was analyzed using several quantitative text measures, and (3) the corresponding practical recommendations for the developers were formulated. The practical value of the work is that it can help researchers and developers understand how to make the human-chatbot interaction process more natural and useful for both the user and the developer.

This paper is organized as follows. In Section 2 the related work connected with this research is presented. In Section 3 the analysis approach is described. Section 4 presents experimental setup. In Section 5 the analysis and discussion of the results is described. Section 6 summarizes the research and explains the limitations and future plans of the work.

2 RELATED WORK

2.1 Human-Chatbot Interaction from Linguistic Perspective

It is well known that chatbots tend to limit free text from the user in order to ensure conversational structure [6]. Developers have to deal with this limitation by hiding it from a user such that one doesn't see explicitly the borders of dialogue. Thus, a chatbot triggers the fact that the language used by a human while interacting with a chatbot differs from a conversation with a human. There are studies showing that a human tends to imitate the vocabulary of a chatbot [7] and to match its language style [8]. The role of used language constructions is not limited to the utterances used by chatbot. If such a system uses machine learning or corpus based methods, its performance is also biased to the available datasets [9]. Hence, it may also negatively influence the quality of machine learning models used in a chatbot (e.g., intent classification) as human-human and human-chatbot language constructions are different. To the best knowledge of the authors, there are only few recent studies that compared human-human and human-chatbot from linguistic perspective [8, 10, 11]. The main research question of these studies is similar to this work: "Do humans communicate differently when they know their conversational partner is a computer as opposed to another human being?". However, this work is different from the mentioned study by a knowledge domain of a chatbot and the language of experiments. *Authors of this is work are not aware of any publications studying differences in human-human and human-chatbot dialogues from the linguistic perspective in Russian language.*

2.2 Measures of Quantitative Text Analysis

The syntactical complexity represents how complex are the sentence structures used in the text. This characteristic is represented by the *Mean Dependency Distance* (MDD) [12]. The MDD is calculated as shown in (1).

$$MDD = \frac{1}{n-s}\sum_{i=1}^{n}|DD_i| \tag{1}$$

Where n is the total number of words in a document, s is the number of sentences in a document, DD_i is the dependency distance of the i-th syntactic link of the document[4] [13].

The lexical diversity represents how complex and dense is the vocabulary used in the text. This characteristic is represented by the following measures. *Lexical Density* (LD) is calculated according to the Ure's method [14] (see (2)).

$$LD = \frac{N_{lexicalitems}}{N_{words}} * 100 \tag{2}$$

Where N_x corresponds to number of a corresponding variable in the formula.

Another well-known measure is *Type-Token Ratio* (TTR) [15]. The term "type" corresponds to the number of unique words in a text corpus. The TTR is calculated according to (3).

$$TTR = \frac{V}{N} \tag{3}$$

[3]Authors will publish the dataset online after the paper acceptance decision.

[4]The connection between words or group of words in a string.

Table 1: Language-specific coefficient values for FRE and FKG metrics.

Language	k_1	k_2
Flesch Reading Ease (FRE)		
English [18]	1.015	84.6
Russian [19]	1.3	60.1
Flesch-Kincaid Grade (FKG)		
English [20]	0.39	11.8
Russian [19]	0.5	8.4

Where V is the number of types and N – number of tokens.

There are also several TTR-related measures such as *Herdan's C* or *LogTTR* [15] (see (4)), *Summer's Index* [16] (see (5)), and *RootTTR* [17] (see (6)).

$$LogTTR = \frac{log(V)}{log(N)} \tag{4}$$

$$S = \frac{log(log(V))}{log(log(N))} \tag{5}$$

$$RootTTR = \frac{V}{\sqrt{N}} \tag{6}$$

The readability measures demonstrate how hard is to read the text. There are several such measures supported in the presented software. The well-known *Flesch Reading Ease* [18] (FRE) depends on the syllables per word ($\frac{n_{sy}}{n_w}$) and words per sentence (*ASL*), see (7).

$$FRE_{lang} = 206.835 - k_1^{lang} * ASL - k_2^{lang} * \frac{n_{sy}}{n_w} \tag{7}$$

The coefficients (k_1^{lang}, k_2^{lang}) mentioned in the Formula are language-specific. The corresponding values are demonstrated in Table 1.

The value of Flesch-Kincaid Grade (FKG) correspond to a U.S. grade level that is required to read a given text [20]. The (8) is dependent on the language-specific coefficients as FRE (see Table 1).

$$FKG_{lang} = k_1^{lang} * ASL + k_2^{lang} * \frac{n_{sy}}{n_w} - 15.59 \tag{8}$$

The *Automated Readability Index* (ARI) [21] and its simplified version – sARI are also supported by the software (see (9) and (10) respectively).

$$ARI = 0.5 * ASL + 4.71 * AWL - 21.34 \tag{9}$$

$$sARI = ASL + 9 * AWL \tag{10}$$

Where *AWL* corresponds to the average word length. The *Coleman's Readability* [22] that includes number of one-syllable words is shown in (11) (where $n_{wsy=x}$ corresponds to the number of words with x syllables).

$$Coleman's = 1.29 * \frac{100 * n_{wsy=1}}{n_w} - 38.45 \tag{11}$$

The *Easy Listening Score* (ELS) [23] is the ratio between the number of words with 2 syllables or more and number of sentences (see (12)).

$$ELS = \frac{n_{wsy>=2}}{n_{st}} \tag{12}$$

3 APPROACH

In this section, the approach for collecting the dialogue data and therefore its comparison is described. Firstly, the chatbot that is able to answer frequently asked questions has to be developed. The underlying knowledge base D for the chatbot must be taken from the trustworthy and validated sources, and structured as pair $d_i = (Q_i, a_i)$, $d_i \in D$, where Q_i – is the list of possible questions targeting on a similar information need (e.g., "Will a mask protect me from the virus?" and "How helpful are the masks for COVID?"), a_i – is the answer text from a validated data source that is fulfilling the information need of Q_i. The algorithm of the FAQ chatbot works over the data D and selects the most relevant a_i for a given Q_i. Such an algorithm is just represented by a simple multi-class classifier as the number of unique a_i is much lower than a number of unique Q_i. This algorithm for FAQ answering of the chatbot was selected due to the lower quality requirements and implementation simplicity in comparison to the data-driven chatbots (e.g., [24]). The process of the FAQ chatbot development is presented in Section 4 in detail.

Secondly, the respondents has to be collected and assigned randomly to the two different groups: Group_1 – respondents communicate with a human-expert, Group_2 – respondents communicate with the developed FAQ chatbot. In case of Group_1, *a respondent knows that a human-expert is on the other side of the dialogue*. In its turn, the human-expert forwards a question to the chatbot and returns an answer to a user produced by the FAQ chatbot, s.t., it is hidden to a respondent (i.e., this is an inverted "Wizard of Oz" experiment following [25, 26]). The *human-expert does not change neither question nor answer* and is required in the experiment to create a trustworthy UI outlook such that a respondent is sure that a dialogue is conducted with a human. In case of Group_2, a user is provided with a link to the FAQ chatbot and prompted to have conversation with the chatbot. These two groups of respondents are independent and are in the same conditions, both utilize the same user interface (UI) while performing the dialogues with either a human-expert or a FAQ chatbot. The only difference is that they know who they are speaking to.

Finally, the collected dialogue data has to be anonymized and therefore analyzed using quantitative text analysis measures. For the analysis all the measures introduced in Section 2.2 were used, in addition such descriptive statistics of text as average- words per sentence, sentence length, word length and syllable per word were computed. The general schema of the approach is demonstrated in Figure 1.

4 EXPERIMENT

The knowledge base was collected from the FAQ section of the official portal of the Russian government "Stop Coronavirus"[5]. Thereafter, it was structured using a parser script as follows $d_i = (Q_i, a_i)$, $d_i \in D$ (see Section 3).

To develop the chatbot, the Google Dialogflow[6] framework was used. Each Q_i was considered as "intent" in the framework. Therefore, the knowledge base D was loaded into the Google Dialogflow platform to train the intent recognition model. After a particular intent i is recognized, the chatbot returns a_i as a response.

The chatbot was deployed as a Bot in the Telegram Messenger[7]. Consequently, it was accessible from any kind of device (e.g., mobile, desktop, tablet etc.). This option enables to use the same UI for both respondent groups and exclude the UI from the possible threats of validity of this study.

The respondents were attracted to the study via social media announcements. There were 103 respondents involved in our experiment in total. 35 respondents were part of the Group_1. The other 68 respondents were part of the Group_2. The first dialogue was conducted on March 16, 2021 and the last dialogue was conducted on July 29, 2021. Hence, the dialogue data collection process continued for more than 4 months. All the dialogues were carried out in Russian language. The following features were collected for each message: chat id, user id, data, first name and last name (anonymized), question, answer. The example of a collected dialogue translated to English is demonstrated in Table 2. Thereafter, a set of quantitative text measures was calculated on both datasets from Group_1 and Group_2 using the LinguaF Python package[8].

[5]https://stopkoronavirus.rf/faq

[6]https://cloud.google.com/dialogflow

[7]https://core.telegram.org/

[8]https://github.com/Perevalov/LinguaF

5 RESULTS AND DISCUSSION

The measures on which the calculation and comparison will be made are presented in Table 3. The columns "p-value" and "Is significant" correspond to the result of two-sample unequal variance two-tailed T-test with significance level (α) 0.01. If the result of this statistical test is "Yes", then the difference is significant.

5.1 Descriptive Statistics and Syntactical Complexity

It is evident from the Table 3 that in case of human-human dialogues, respondents use more words and therefore sentences are longer (see metrics 1, 2). For instance, in a human-human, the respondent uses 1.8 times more words than in a human-chatbot dialogue. There were also significant differences seen on measures 3 and 4. Hence, *respondents use longer or more complex words*. Since the average distance between dependent words in human-human dialogues (2.210) is greater than in human-chatbot dialogues (1.651), the Syntactic complexity will also be higher (see Metric 5).

5.2 Lexical Diversity

Lexical density and Type-Token Ratio (cf., Section 2.2) values demonstrate that the differences between human-human and human-chatbot dialogues are insignificant. However, considering such measures as Log Type Token Ratio, Root Type Token Ratio and Summer Index, it is obvious that Lexical diversity in human-human dialogues is much higher than in human-chatbot ones. Thus, in the dialogue with the chatbot, the respondents choose simpler phrases. Consequently, *when developing such a chatbot at the stage of preparing the training data, it is worth considering the fact that dialogues with such systems are not always similar to human ones*, and, if necessary, make adjustments to the data (e.g., simplification of the questions).

5.3 Readability

The metric Flesch Reading Ease also signals the use of more complex language constructions in "human" dialogues. According to the obtained values of the Flesch Reading Ease metric, in the case of a dialogue with a human, the respondents operate close to the language level of a university student, while *in a dialogue with a chatbot, the level of language constructions is two steps lower, which corresponds to the 7th*

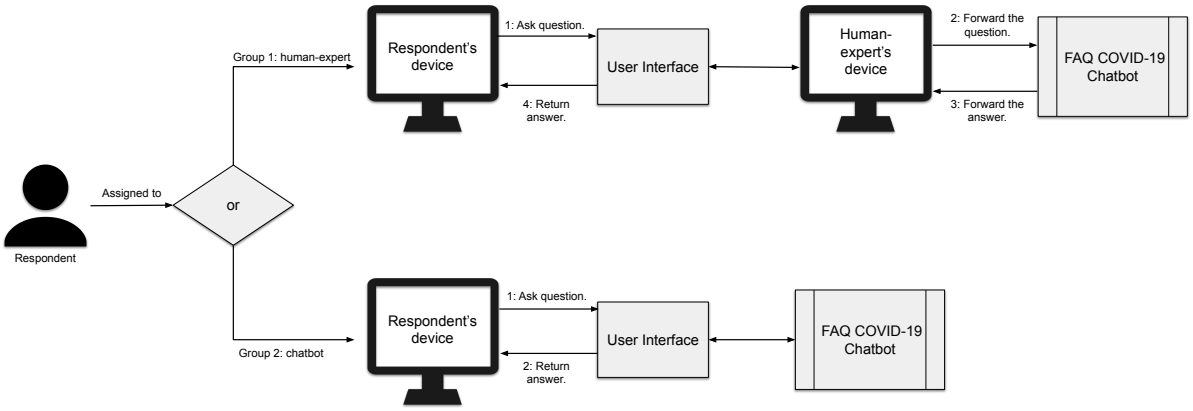


Figure 1: The general approach schema. The human-expert is just "playing" the role of an expert, while actually forwarding all the questions and answers to/from the FAQ chatbot (inverted "Wizard of Oz"). In this setting, the respondent thinks that the communication partner is a real human-expert, however, the human-expert is fully controlled by the chatbot.

Table 2: The examples of collected questions and their intentions. For each intention, an answer in the knowledge base is defined.

№	Question	Intention	Group
1	What vaccines there are?	Information on vaccines	2
2	Can I put Pfizer in a private clinic, for example?	Information on vaccines	2
3	What should I do if I see signs of Coronovirus?	Symptoms	1
4	What antibodies can be detected and what does this tell me?	Antibodies	1

Table 3: Results of the measures' calculation. The column "Avg. Human Data" contains the values of the human-human dialogues (i.e., $Group_1$). The column "Avg. Chatbot Data" contains the values of the human-chatbot dialogues (i.e., $Group_2$). The values of the "Is significant" column are calculated according to the "p-value" results and the significance level $\alpha = 0.01$.

№	Measure Name	Avg. Human Data	Avg. Chatbot Data	p-value	Is significant
	Descriptive Statistics				
1	Avg Words Per Sentence	6.287	3.370	0.000	Yes
2	Avg Sentence Length	37.692	18.894	0.000	Yes
3	Avg Word Length	6.144	5.812	0.007	Yes
4	Avg Syllable Per Word	2.348	2.160	0.000	Yes
	Syntactical Complexity				
5	Mean Dependency Distance	2.210	1.651	0.000	Yes
	Lexical Diversity				
6	Lexical Density	69.902	67.946	0.237	No
7	Log Type Token Ratio	0.986	0.806	0.000	Yes
8	Root Type Token Ratio	2.509	1.774	0.000	Yes
9	Type Token Ratio	0.997	0.998	0.189	No
10	Summer's Index	0.986	0.806	0.000	Yes
	Readability				
11	Automated Readability Index	10.650	7.627	0.000	Yes
12	Automated Readability Index Simple	61.580	55.674	0.000	Yes
13	Coleman Readability	41.439	51.837	0.000	Yes
14	Easy Listening	1.998	1.252	0.000	Yes
15	Flesch-Kincaid Grade	7.279	4.239	0.000	Yes
16	Flesch Reading Ease	57.532	72.639	0.000	Yes

grade student. The Flesch-Kincaid Grade indicates the level of readability as 7th grade in human-human dialogues and 4th grade in human-chatbot dialogues. While *the absolute values may be debatable, the relation between them shows that the language used in the human-chatbot dialogues is simpler*. Also, the Automated Readability Index and Simplified Automated Readability Index metrics show that in human-human dialogues, the perception of text is more complex than in human-chatbot dialogues. Based on the calculation of all the measures in this section, the complexity of readability in human-human dialogues appears to be several levels higher than in human-chatbot dialogues. It is worth noting that the "complexity" may be given by the subject of the dialogues, as in this case numerous medical terms are used.

5.4 General Statistics and Qualitative Analysis on Dialogues

The average number of messages in human-human dialogues is 9.4. The average number of messages in human-chatbot dialogues is 5.3. The total average number of messages in all dialogues is 7.35. The duration of the dialogues also varied. In the case of human-human, the dialogues ranged up to 30 minutes. In the case of human-chatbot dialogues, they are no more than 15 minutes long. The human-human dialogues also appear to be longer w.r.t. the number of questions that were asked on average (9.4 vs 5.3). This difference may be due to the social and psychological aspects of human-chatbot interaction.

The qualitative analysis of the dialogue data showed that in human-human dialogues, in most cases, respondents used more complex sentence structures and gave more clarifying information to get an answer to similar questions than in human-chatbot dialogues. *In the case of human-chatbot dialogues, respondents often ask personal questions*, such as: "How are you?", "What do you know?", "Who is your creator?", whereas *when talking to a human, only questions on the subject were used*. Moreover, in human-chatbot dialogues, the respondents frequently use obscene language.

5.5 Summary

In this section, two dialogues types – human-human and human-chatbot– were analyzed. The content of human-human dialogues is predominantly more complex and rich than in the case of human-chatbot dialogues (RQ_1). It is confirmed by the set of measures related to the Descriptive Statistics of text, Syntactic Complexity, Lexical Diversity, and Readability (RQ_2). The respondents, when communicating with chatbots, construct their speech intentionally or subconsciously in a simpler and clearer way. Therefore, chatbots and other dialogue systems should be designed so that they are prepared to work effectively with the simplified language input. To achieve this, it is proposed to use simplification techniques on the training data that is used to build such systems (RQ_3).

6 CONCLUSION

In this work, the differences between human-human ($Group_1$) and human-chatbot ($Group_2$) dialogues were analyzed. For this purpose, the respondents of the experiment were randomly assigned to one of the two aforementioned groups. The domain of the dialogues was focused on frequently asked questions about COVID-19 and the language of the dialogues was set to Russian. To conduct the experiment, the FAQ chatbot was created based on publicly available data, provided by the Russian government.

In total, 103 respondents were involved in the experiment. The scope of the analysis contained different quantitative text measures related to syntactical complexity, lexical diversity, readability and others. Given the experimental results, it was possible to identify significant difference w.r.t. the considered measures between human-human and human-chatbot dialogues. Specifically, respondents in the human-chatbot dialogues used shorter and simpler language constructions, which is reflected in the experimental results. Based on this, the authors of this work recommend researchers and developers of chatbots and dialogue systems to consider simplifying the training data used for the systems, as the majority of users are not asking well-formed questions. To summarize the discussion, the research questions of the study were fully answered.

However, several limitations may be identified in this work. Firstly, the selection process of the respondents might be biased due to the used social media connections of the authors of this work. In addition, only one UI was used in the experiments, hence, it may have added some bias into the dialogues as well. Finally, the work is limited by the used language of dialogues and knowledge domain.

For the future work, it is worth considering extending the analysis towards more languages, especially, low resource ones. In order to reduce the bias of the respondents, the selection process should be aligned it the way it ensures maximal diversity of the respondents and statistical stability of the results. Moreover, different chatbot user interfaces should be

used to reduce a possible corresponding bias. Finally, after each dialogue a (short) feedback from a respondent should be collected (e.g., "Do you think that you were talking to a chatbot?" , "Were the answers helpful?", etc.) to determine the current impressions of the users. In the same context, It would be very interesting at what frequency of such questions the users measurably change their behavior.

REFERENCES

[1] U. Gnewuch, S. Morana, and A. Maedche, "Towards designing cooperative and social conversational agents for customer service." in Proceedings of the 38th International Conference on Information Systems (ICIS), 2017.

[2] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Deliv-ering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully au-tomated conversational agent (woebot): a randomized controlled trial," JMIR mental health, vol. 4, no. 2, p. e7785, 2017.

[3] A. Androutsopoulou, N. Karacapilidis, E. Loukis, and Y. Charalabidis, "Transforming the communication between citizens and government through ai-guided chatbots," Government Information Quarterly, vol. 36, no. 2, pp. 358-367, 2019.

[4] A. P. Chaves and M. A. Gerosa, "How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design," International Journal of Human–Computer Interaction, vol. 37, no. 8, pp. 729-758, 2021.

[5] M. R. Mehl, "Quantitative text analysis." 2006.

[6] D. Duijst, "Can we improve the user experience of chatbots with personalisation," Master's thesis. University of Amsterdam, 2017.

[7] M.-C. Jenkins, R. Churchill, S. Cox, and D. Smith, "Analysis of user interaction with service oriented chatbot systems," in International Conference on Human-Computer Interaction. Springer, 2007, pp. 76-83.

[8] J. Hill, W. R. Ford, and I. G. Farreras, "Real con-versations with artificial intelligence: A compari-son between human–human online conversations and human–chatbot conversations," Computers in human behavior, vol. 49, pp. 245-250, 2015.

[9] A. Schlesinger, K. P. O'Hara, and A. S. Taylor, "Let's talk about race: Identity, chatbots, and ai," in Proceed-ings of the 2018 chi conference on human factors in computing systems, 2018, pp. 1-14.

[10] Y. Mou and K. Xu, "The media inequality: Comparing the initial human-human and human-ai social interactions," Computers in Human Behavior, vol. 72, pp. 432-440, 2017.

[11] E. Silkej, "Linguistic differences in real conversations: Human to human vs human to chatbot," 2020.

[12] M. Oya, "Syntactic dependency distance as sentence complexity measure," Proceedings of the 16th International Conference of Pan-Pacific Association of Applied Linguistics, 01 2011.

[13] H. Liu, "Dependency distance as a metric of language comprehension difficulty," Journal of Cognitive Science, vol. 9, pp. 159-191, 09 2008.

[14] J. Ure, "Lexical density and register differentiation," Applications of Linguistics, pp. 443–452, 1971.

[15] G. Herdan, Quantitative Linguistics. London: Butterworth, 1960.

[16] H. H. Sommers, "Statistical methods in literary analysis," The computer and literary style, pp. 128-140, 1966.

[17] P. Guiraud, Problèmes et Méthodes de la Statistique Linguistique. Paris: Presses universitaires de France, 1960.

[18] R. Flesch, "A new readability yardstick," Journal of Applied Psychology, pp. 221–233, 1948.

[19] V. Solovyev, V. Ivanov, and M. Solnyshkina, "Assessment of reading difficulty levels in russian academic texts: Approaches and metrics," Journal of Intelligent and Fuzzy Systems, vol. 34, pp. 1-10, 04 2018.

[20] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom, "Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel," Naval Technical Training Command Millington TN Research Branch, Tech. Rep., 1975.

[21] R. Senter and E. A. Smith, "Automated readability index," CINCINNATI UNIV OH, Tech. Rep., 1967.

[22] E. B. Coleman, "Developing a technology of written instruction: Some determiners of the complexity of prose," Verbal learning research and the technology of written instruction, pp. 155-204, 1971.

[23] I. E. Fang, "The "easy listening formula"," Journal of Broadcasting & Electronic Media, vol. 11, no. 1, pp. 63-68, 1966.

[24] A. Both, A. Perevalov, J. R. Bartsch, P. Heinze, R. Iudin, J. R. Herkner, T. Schrader, J. Wunsch, A. K. Falkenhain, and R. Gürth, "A question answering system for retrieving german covid-19 data driven and quality-controlled by semantic technology," in Joint Proceedings of the Semantics co-located events: Poster&Demo track and Workshop on Ontology-Driven Conceptual Modelling of Digital Twins co-located with SEMANTiCS 2021, ser. CEUR Workshop Proc., vol. 2774, CEUR-WS.org, 2021.

[25] L. D. Riek, "Wizard of oz studies in hri: a systematic review and new reporting guidelines," Journal of Human-Robot Interaction, vol. 1, no. 1, pp. 119-136, 2012.

[26] J. Newn, R. Singh, F. Allison, P. Madumal, E. Velloso, and F. Vetere, "Designing interactions with intention-aware gaze-enabled artificial agents," in IFIP Conference on Human-Computer Interaction. Springer, 2019, pp. 255-281.

Robotic System Operation Specification on the Example of Object Manipulation

Leonid Mylnikov, Pavel Slivnitsin and Anna Mylnikova
Perm National Research Polytechnic University, 29 Komsomolsky avenue, Perm, Russia
leonid.mylnikov@pstu.ru, slivnitsin.pavel@gmail.com, novikova@yandex.ru

Keywords: Function Modeling, Diagram, Metalanguage, Recognition, Identification, Positioning, SLAM, Computer Vision, Robotic System.

Abstract: Currently, robotic system tasks are formalized with help of procedural programming languages that do not take into account the specificity of robots and are not generic in their application. The goal of the paper is to develop a method of semantic description of the sequence of operations performed by a robotic system on the example of object manipulation around them. To achieve the goal, a method of a graphical representation of a robotic system operation specification and its semantic description (metalanguage) are proposed. The paper considers the approaches to the objects' representation, determines the way object characteristics are stored, and provides the list of possible operations with objects. The obtained methods of graphical and semantic robotic system operation specification allow to assign the task without being bound to a specific technical solution. In addition, the paper provides the examples of operation assignments for the robotic arm.

1 INTRODUCTION

The current development trends of automation technologies and robotic systems are their autonomation and more enhanced complexity of the operations performed. In industry, this is caused by the need to increase production performance and quality, the need to provide high-precision machining in the context of enhanced complexity, reduced size, and shorter life cycles of products. The automation of routine operations such as welding, painting, assembly, and sorting [1] is becoming insufficient.

These operations cannot be performed in the context of a dynamic environment without solving the task of object recognition and considering the variability of motion paths. More complex operations such as manipulation of various objects, assembly and disassembly, repair, adjustment are not possible at all without taking into account the peculiarities of the environment.

Object manipulation tasks are becoming more and more widespread. They are often used for robotic systems positioning, such as warehouse maintenance robots [2], urban infrastructure facilities' maintenance [3], etc.

In order to improve the efficiency of such systems, there are various competitions. For example, the competition on the automatic object picking and sorting [4], [5].

2 THE DESCRIPTION OF OBJECT MANIPULATION OPERATIONS

The task of objects' manipulation performed by a robotic system is implemented as a sequence of operations with the manipulator and objects, which can be represented by the Figure 1, where Subj is a subject that performs the operation (manipulator), Obj is an object of the operation, V is the operation to be performed, NS are features that distinguish the subject, NV is the goal (where to move/shift/place/etc. an object), NO are the features that distinguish the object.

To specify operations and their describing structures, we will use the following symbols [6]: «=» – clarifying the concept; «{}» – merge; «<>» – mandatory part; «[]» – optional part; «|» – or; «&» – and; «\» – clarifying a new variable, «""» – rigidly given element.

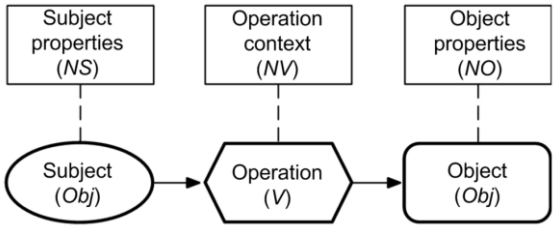


Figure 1: The functional diagram of the robotic system task assignment.

Then, syntactically a task can be described as the following structure:

<Subj>[NS]<V>[NV]<Obj>[NO].

For example, the task "Use the manipulator A to move the object B from X to Y" can be described as follows:

\Subj="A"
\V="move"
\Obj="B"
\NV=Y
\NO=X
Subj V NV Obj NO.

Let's consider the task of outdoor luminaire replacement described in [3], which is represented in natural language by the following expression: **«Use the manipulator A, replace the luminaire B with the luminaire C on the lighting column D».**

In this example, we are dealing with a complex opeartion that can be decomposed into a number of operations (e.g., remove and install) and a complex object (consisting of the objects B and C). The Figure 2 shows the structure of this task.

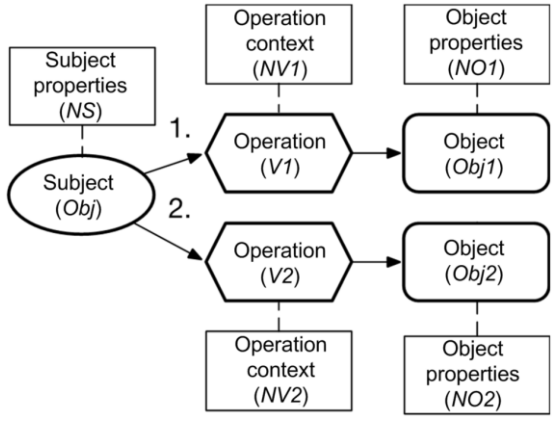


Figure 2: The functional diagram of the robotic system task assignment with a complex operation and a complex object.

Semantically, this operation can be represented as follows:

\Subj="A"
\V1="remove"
\Obj1="B"
\NV1=D
Subj V1 NV1 Obj1
\V1="install"
\Obj2="C"
\NV2=D
Subj V2 NV2 Obj2.

From the above examples we can make a conclusion that the robotic system operation can be defined by a set of operations **V={produce, repair, assemble, disassemble, replace, take, lift, put, place, insert, throw, separate the gripper, bring the gripper together, move the gripper to some location, etc.}** (some of which can be complex **V={V1,V2, etc.}**), as well as by the ability to identify objects and their specified features **NO={location, shape, color, material, etc.}** based on which a number of certain operations can be performed.

Complex operations can be decomposed in different ways into elementary ones, provided that the principle of equivalence is observed. This provides an opportunity to consider the optimization of the robotic system operation.

The optimization process of the robotic system operation is defined as follows: 1) the algorithms that implement elementary operations, 2) object recognition algorithms, 3) the knowledge about the environment (the underlying circumstances of the operation).

For example, the movement of the manipulator from the position X to the position Y can be implemented by the application of the shortest path algorithm or the ant colony optimization algorithm. In the first case, if obstacles occur in the environment, the operation cannot be performed. Moreover, different implementations of this operation will have different power consumption, different execution time, as well as they can have different execution risk assessments.

The efficiency of operation execution algorithms is not relevant if it is not possible to identify the object or there is a lack of information about the operation context.

3 APPROACHES AND METHODS FOR OBJECT IDENTIFICATION AND LOCATION DESCRIPTION IN MANIPULATION TASKS

The recognition of objects in the environment is a complex task, that includes the localization of objects, their identification, search for interaction tools with objects, the mapping of the environment and building the knowledge bases that describe the environment.

Existing object recognition methods are based on the selection of certain features, a set of elements or templates that are used to identify objects. Along with the feature detection, it is also important to consider the relationships between features that have an impact on recognition (for example, see the Thatcher effect or Thatcher illusion [7] shown in the Figure3).

Figure 3: The Thatcher Effect presented by Peter Thompson.

The recognition process and its application can be represented by the scheme shown in the Figure 4. In the paper we will consider several steps, which indicated in the figure by the numbers 2-4. Depending on the implementation algorithm, some blocks can be combined or can be executed simultaneously.

3.1 Approaches to Object Representation

The ideas of object representation (block 3 in Figure 4) used in computer vision systems are based on the theories of human object recognition. At the moment, there is a number of theories that describe the approaches to the recognition and classification of objects by humans.

The template matching theory (exemplar theory) [8] assumes that for each object there is a template in the memory. By the processing of new information, object identification requires an exact match between the object and the template from the memory. The disadvantage of the template matching theory is the need to store a large number of templates.

Another theory is the **prototype theory** [9], [10]. It involves the comparison of new information not with templates, but with some abstract object prototypes (see the Figure 5). A prototype is based on a set of examples of the object and describes their common features. There are two models for prototype formation: the central tendency model and the attribute-frequency model. [10].

In the context of computer vision, the algorithms SIFT [11] and SURF [12] are well-known, which have elements of both theories. They are based on the attribute extraction of template objects, which is used to detect objects in the image.

Later theories have developed the ideas of attribute extraction to form a prototype. For example, according to the **feature analysis theory** (see [5] and [6]), the human visual system includes feature detectors and object recognition is based on the extraction of the simplest features of objects (see Figure 6). The theory assumes a layered recognition. The simplest feature detectors detect simple features. The next feature detectors are capable of detecting more complex features. In the case of the occurrence of same features or the same combination of features for certain objects, it becomes possible to determine that these objects belong to the same class.

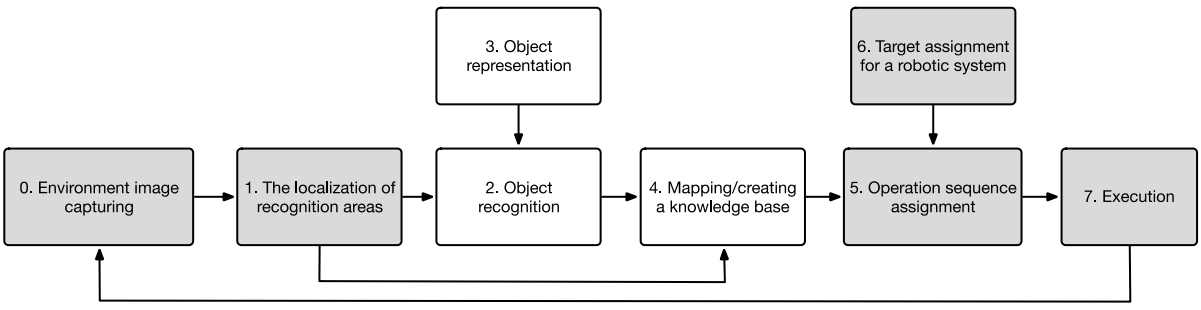
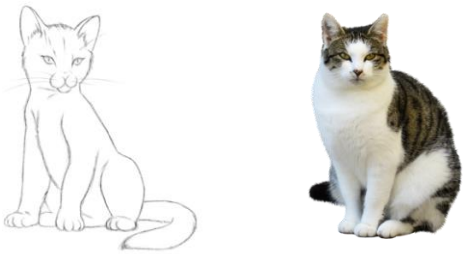


Figure 4: The sequence of steps performed by a robotic system to collect and use information about the environment.



a) *b)*

Figure 5: Prototype theory: *a)* abstract prototype, *b)* class instance.

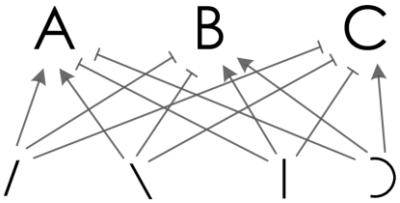
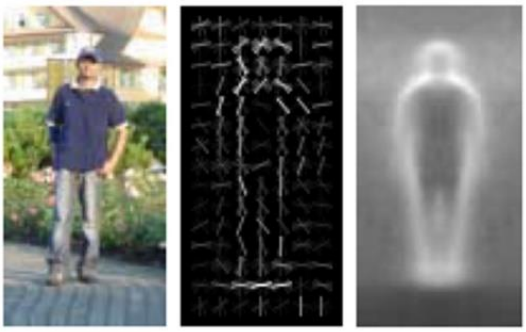


Figure 6: An example of identifying English alphabet characters using features.

The elements of this approach can be traced in the Dalal-Triggs method [15] and the Viola-Jones method [16].

The Dalala-Triggs method is based on the calculation of histograms of oriented gradients (HOG) (see the Figure 7).

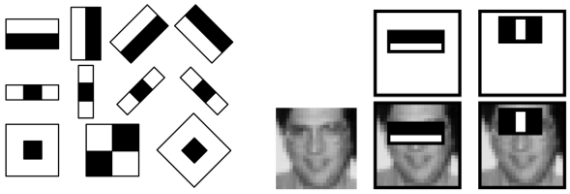


a) *b)* *c)*

Figure 7: The Dalal-Triggs Method: *a)* an image of a person, *b)* oriented gradients (features), *c)* trained structure (prototype) [15].

The Viola-Jones method is based on an integral image representation and describes objects using a combination of typical features from a limited set (Figure 8).

The ideas of feature analysis are used in convolutional neural networks (CNN) (Figure 9) [17].



a) *b)*

Figure 8: The Viola-Jones Method: *a)* a set of features, *b)* matching the features to the object (face) [16].

In this approach, a convolution operation is used, which is based on image processing with help of a convolution filter, which allows to extract the object features. The result of the convolution operation in the image is a new image. The intermediate (hidden) convolutional layers represent a matrix (a feature map, which can also be represented as an image) or a set of matrices, depending on the number of filters applied to the previous layer. Each next hidden layer of the neural network detects more complex object features compared to the previous one.

The approach described by A. Kononyuk in [19] can also be associated with the feature analysis theory. It is based on the **feature extraction of benchmark objects** from the training set **using predicate logic**.

In this approach features can capture color, spatial arrangement, etc. Each object is represented by a specific set of features. Each feature is described by a predicate. Predicate arguments are elements of the benchmark images that indicate the occurrence of the feature in it. The representation of the benchmark image (and object in it as well) is the conjunction of all the object features. An object class is represented by a training set as a disjunction of all benchmark images that contain an object of the class $\bigcup_{\omega_k \in \Omega} \bigcap_{i=1}^{h} P_i(c_{1i}, \ldots, c_{ni})$, where ω_k is the k-th benchmark image of the class Ω, h is the number of features, c_i are parts of the image that describe the occurrence of the i-th feature in the benchmark image.

Based on the classes described in the way as specified above, the object recognition task is performed as a logical inference task, which allows to identify object classes on the new image. Thus, the recognition task is to prove equivalence of objects.

Within this approach, the same set of primary features can occur for different objects, which corresponds to the feature analysis theory. For example, the primary features are lines (vertical, horizontal, parallel), and the secondary features represent more complex combinations of lines (rectangles, etc.).

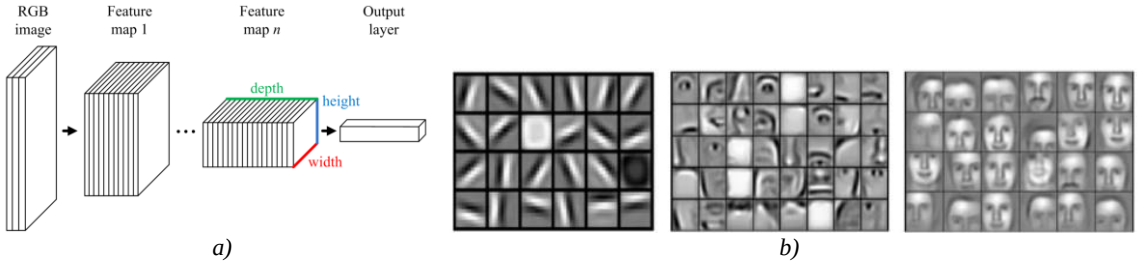


Figure 9: A convolutional neural network: a) structure, b) feature maps [18].

The implementation of feature analysis methods generated a group of approaches to object recognition that are based on the filter and mathematical functions' application for image processing and assigning objects to certain classes [20], [21].

A further development of recognition theories is D. Marr's **computational theory of human stereo vision** [22]–[24], which assumes that recognition is multistage and involves more enhanced degree of object details.

At the first stage the information about contours, edges and spots is processed. At the second stage information about the depth and object surfaces position is processed. After that, at the third stage a three-dimensional model of the detected object is formed. According to this theory, a three-dimensional representation is based on the canonical forms (e.g., cylinders). Thus, objects can be represented as a set of cylinders of different sizes with an axis, depending on the degree of detail (see [15] and [16], the Figure 10).

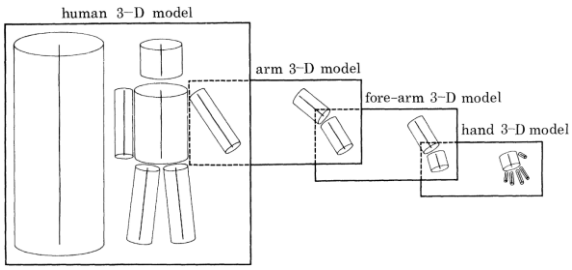


Figure 10: A visual representation of a three-dimensional model of a human using the computational theory [23].

I. Biederman developed the **recognition by component theory** [25]. According to this approach, a human being perceives objects of the real world through a certain set of geometric primitives called geons and relations among them. Thus, every object and every scene in the image can be represented by a set of primitives (Figure 11).

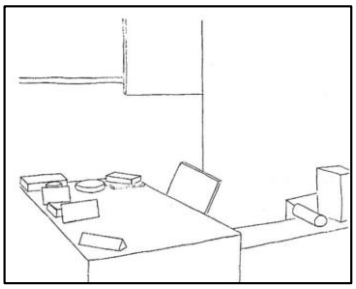


Figure 11: Scene presentation using geons [26].

Each primitive can also be described through nonaccidental properties of shapes, i.e. properties that do not change when the angle of view changes (e.g. collinearity, curvilinearity, symmetry, etc.). Thus, each component of an object can be represented by the relationship of a number of primitives, which are described by a number of nonaccidental properties. Biederman distinguishes the following relations between geons: verticality, relative size, centering, surface size join [26].

In position recognition systems, such as human pose detection [27] (see the Figure 12) or hand gesture detection [28], there are methods based on the extraction of key points or key structures of objects (e.g., skeleton structures).

Figure 12: Human pose detection using a skeleton structure [27].

The main differences of the analyzed approaches are shown in the Table 1.

Table 1: The comparison of object representation approaches.

	Features	Relations
1. The template matching theory (exemplar theory)	A set of templates	—
1.1. SIFT	The extraction of the features of the template objects to detect them in the image	—
1.2. SURF		
2. The prototype theory	A set of prototypes	—
3. The feature analysis theory	The sets of geometric features	—
3.1. Dalala-Triggs Method	A unique set of features forming the prototype	—
3.2. Viola-Jones Method	A set of features is formed for each required class	A class is defined by a combination of features
3.3. Neural networks	Each layer of the neural network represents a map of features	—
4. The computational theory	Features are relations between a number of primitive objects of the same type	Relations between the primitives form the class
5. Recognition by components theory	Features are a set of geometric primitives	The relation type is considered as a feature
6. Extraction of key points or key structures	Features are key points or key structures that describe the object's position	Relations between key points
6.1. Neural networks	The feature is a set of key points that form the object prototype structure (object's configuration)	

3.2 Object Recognition Using Features

Object recognition on an image includes the object detection [29] and object classification, and in some cases, image segmentation (semantic or instance [29]) or detection of object parts (key points) [27].

In the object recognition task, it is necessary to define classes of objects, each of which is assigned a specific set of features (based on information that is extracted from the image, including the object location on the image). For example, in the case of using object parts as features, the difference between classes can be demonstrated as shown in the Figure 13.

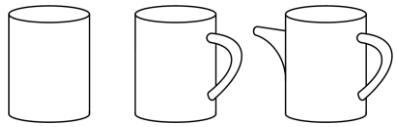


Figure 13: Objects with the same main shape feature (a – glass (cylinder), b – mug (cylinder + handle), c – kettle (cylinder + handle + cone)).

In multiple classification tasks, we often have to deal with false-positive and false-negative object identification (for example, objects can have features which can be associated with two or more classes or important object elements are not visible on the image). In this regard, specific algorithms for object recognition are being developed, which include:

- multi-step recognition process (the number of classes is reduced, the information about the object is gradually clarified) (e.g., changing a view angle [4]);
- object part recognition (objects are classified by the presence or absence of some predefined parts and their relative position; information about the position of object parts allows, among other things, to make assumptions about the position of the object in the image).

At the moment, among the methods which allow to work with objects as complex elements of the environment there are classification models and methods [30] (classifiers such as Naïve Bayes, SVM, kNN, decision trees), associative rule learning methods (Apriori, Eclat, FP-growth, OPUS, SlopeOne [31]), expert systems, predicate logic, neural networks (see [17], [18], [27], [28]) (see the Table 2).

Table 2: The peculiarities of different groups of methods used for object recognition with features.

A method	Tasks	Recognition approach
1. Classifiers	Object classification	Detection of object features and relations between them in an image
2. Association rule learning	The analysis of relations between the components of the object	Pairwise comparison of relations between features
3. Expert systems	Hypothesis testing based on information about the object	The detection of object features
4. Predicate logic	The description of objects as a set of elements (parts) and relations between them	The detection of object features and relations between them in an image
5. Neural networks	The detection of objects and their configuration (position, pose)	Feature extraction from an image using a trained neural network

Holistic objects (consisting of parts) can also be considered as elements. Thus, information about the objects position in the image allows us to estimate the relative positions of objects in the environment.

3.3 Building a Base of Knowledge of the Environment

The task is related to the tools used to obtain initial information about the environment (photo/video images, images from stereo cameras, 3-D scans, etc.), to form maps of the environment, and to store information about it (classes of objects, relations between objects, etc.).

Environment mapping is used in robotic systems' positioning and navigation (obstacle detection and shortest path search).

The SLAM methods [32] focus on collecting information about the environment in order to build a map of an area within which it is possible to move the manipulator. This approach does not provide information about the content of the environment, which includes a number of objects. The information about obstacles is enough for the robotic system navigation, but in the tasks related to the object manipulation it becomes insufficient, because it is necessary to identify objects in order to interact with them. Information just about the classes of objects and their position in the image (or in point cloud) can also be insufficient. A more detailed analysis of the object, its parts and its position can be required for object manipulations as it allows to determine the list of operations that can be performed with an object. Thus, in addition to accumulating information about the environment, it is necessary to accumulate information about its content.

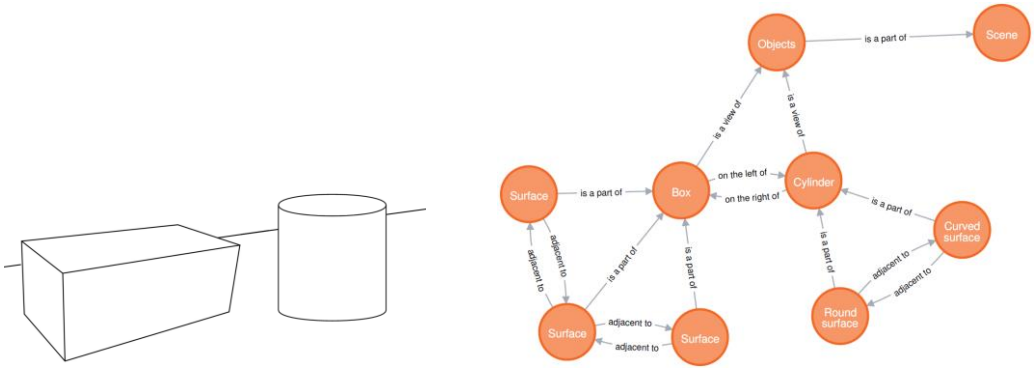
Therefore, for object manipulation, the recognition task includes the recognition of object features, the recognition of object parts, the identification of elements with which it is possible to interact and the relations between them, as well as the recognition of the scene or map of the environment. For this purpose, a knowledge base about the environment is built.

To describe the environment it is necessary to describe what objects it contains, as well as the relations between the objects. Objects can be represented as nodes with a set of some properties (color, material, etc.), and object relations can be represented as edges between nodes (left, above, etc.). Semantic networks, frame-type expert systems, network data models [33] have been used to describe such information (see an example in the Figure 14).

These approaches can also be applied to describe the object structure (for example, the representation of objects as spatial combinations geometric primitives). Primitives can be represented by nodes, and spatial relations between these primitives can be described by edges between nodes. A set of primitives, in this case, defines a list of operations that can be performed with an object.

4 CONCLUSIONS

The paper considers a method of semantic task description (metalanguage) for robotic systems and its graphical representation. The paper investigates approaches to object identification using features as well as approaches to describe the environment. It allows to formulate tasks, taking into account the context of the operation and the object features, that allows the following:



a) *б)*

Figure 14: An example of applying a semantic network for scene description: a) the scene, b) a semantic network.

- to bring the task assignment for robotic systems closer to the operations description in natural language;
- to bring invariance to the task execution, depending on the algorithm efficiency;
- to not impose algorithmic constraints to the practical implementation of basic operations.

ACKNOWLEDGMENTS

The reported study was partially supported by the Government of Perm Krai.

REFERENCES

[1] Fanuc Europe, "Robot industrial applications." [Online]. Available: https://www.fanuc.eu/de/en/industrial-applications. [Accessed: 01-Jun-2021].

[2] A. Delfanti and B. Frey, "Humanly Extended Automation or the Future of Work Seen through Amazon Patents," Sci. Technol. Hum. Values, vol. 46, no. 3, pp. 655–682, 2021.

[3] P. Slivnitsin, A. Bachurin, and L. Mylnikov, "Robotic system position control algorithm based on target object recognition," in Proceedings of International Conference on Applied Innovation in IT, 2020, vol. 8, no. 1, pp. 87–94.

[4] A. Zeng et al., "Multi-view self-supervised deep learning for 6D pose estimation in the Amazon Picking Challenge," Proc. - IEEE Int. Conf. Robot. Autom., pp. 1386–1393, 2017.

[5] A. Zeng, S. Song, J. Lee, A. Rodriguez, and T. Funkhouser, "TossingBot: Learning to Throw Arbitrary Objects with Residual Physics," IEEE Trans. Robot., vol. 36, no. 4, pp. 1307–1319, 2020.

[6] A. Novikova, "Direct Machine Translation and Formalization Issues of Language Structures and Their Matches by Automated Machine Translation for the Russian-English Language Pair," in Proceedings of International Conference on Applied Innovation in IT, 2018, pp. 85–92.

[7] P. Thompson, "Margaret Thatcher: A New Illusion," Perception, vol. 9, no. 4, pp. 483–484, Aug. 1980.

[8] R. M. Nosofsky, "The generalized context model: an exemplar model of classification," Form. Approaches Categ., pp. 18–39, 2012.

[9] E. Rosch, "Cognitive representations of semantic categories.," J. Exp. Psychol. Gen., vol. 104, no. 3, pp. 192–233, Sep. 1975.

[10] P. G. Neumann, "Visual prototype formation with discontinuous representation of dimensions of variability," Mem. Cognit., vol. 5, no. 2, pp. 187–197, Mar. 1977.

[11] D. G. Lowe, "Object Recognition from Local Scale-Invariant Features," in Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999, p. 8.

[12] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3951 LNCS, no. July 2006, pp. 404–417, 2006.

[13] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," Psychol. Rev., vol. 65, no. 6, pp. 386–408, 1958.

[14] O. G. Selfridge, "Pandemonium: a paradigm for learning," in Proceedings on the Symposium on Mechanisation of Thought Processe, 1959, pp. 511–529.

[15] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," Proc. - 2005 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition, CVPR 2005, vol. I, no. 16, pp. 886–893, 2005.

[16] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in Proceedings IEEE Conf. on Computer Vision and Pattern Recognition, 2001, pp. 511–518.

[17] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[18] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Unsupervised learning of hierarchical representations with convolutional deep belief networks," Commun. ACM, vol. 54, no. 10, pp. 95–103, 2011.

[19] A. E. Kononyuk, Obshchaya teoriya raspoznavaniya. Matematicheskiye sredstva opisaniya raspoznavayemykh obyektov i raspoznayushchikh protsessov. Kiyev. 2012.

[20] P. J. Diggle and J. Serra, “Image Analysis and Mathematical Morphology.,” Biometrics, vol. 39, no. 2, p. 536, Jun. 1983.
[21] Y. V. Vizilter, Y. P. Pyt’ev, A. I. Chulichkov, and L. M. Mestetskiy, “Morphological Image Analysis for Computer Vision Applications,” 2015, pp. 9–58.
[22] D. Marr and T. Poggio, “A computational theory of human stereo vision,” Proc. R. Soc. London - Biol. Sci., vol. 204, no. 1156, pp. 301–328, 1979.
[23] D. Marr and H. K. Nishihara, “Representation and recognition of the spatial organization of three-dimensional shapes,” Proc. R. Soc. London. Ser. B. Biol. Sci., vol. 200, no. 1140, pp. 269–294, Feb. 1978.
[24] D. Marr and L. Vaina, “Representation and recognition of the movements of shapes,” Proc. R. Soc. London. Ser. B. Biol. Sci., vol. 214, no. 1197, pp. 501–524, Mar. 1982.
[25] I. Biederman, “Recognition-by-Components: A Theory of Human Image Understanding,” Psychol. Rev., vol. 94, no. 2, pp. 115–147, 1987.
[26] I. Biederman, “Matching Image Edges To Object Memory.” pp. 384–392, 1987.
[27] S. Jin et al., “Whole-Body Human Pose Estimation in the Wild,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12354 LNCS, pp. 196–214, Jul. 2020.
[28] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, “Hand keypoint detection in single images using multiview bootstrapping,” Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 4645–4653, 2017.
[29] L. Jiao et al., “A survey of deep learning-based object detection,” IEEE Access, vol. 7, no. 3, pp. 128837–128868, 2019.
[30] L. A. Mylnikov, Statisticheskiye metody intellekttsualnogo analiza dannykh. SPb.: BKhV-Peterburg. 2021.
[31] D. Lemire and A. Maclachlan, “Slope {One} {Predictors} for {Online} {Rating}-{Based} {Collaborative} {Filtering},” SIAM Data Min. (SDM’05), Newport Beach, California, April 21-23, 2005.
[32] D. Vershinin and L. Mylnikov, “A review and comparison of mapping and trajectory selection algorithms,” Proc. Int. Conf. Appl. Innov. IT, vol. 9, no. 1, pp. 85–92, 2021.
[33] G. F. Luger, Artificial Intelligence: Structures and Strategies for Complex Problem Solving, vol. 5th. 2005.

An Adaptive Technique of Digital Maturity Integral Estimation for an Organisation

Yuliya Shevtsova[1], Tatiana Monastyrskaya[2], Aleksei Poletaykin[3] and Gleb Toropchin[4]
[1]*Department of Mathematical Modelling and Digital Development, Institute of Problem Solving, Siberian State University of Telecommunications and Information Science, 86 Kirova Str., Novosibirsk, Russia*
[2]*Department of Social and Communicative Technologies, Institute of Problem Solving, Siberian State University of Telecommunications and Information Science, 86 Kirova Str., Novosibirsk, Russia*
[3]*Department of Information Technologies, Kuban State University, 149 Stavropolskaya Str., Krasnodar, Russia*
[4]*Faculty of Humanities, Novosibirsk State Technical University, 20 Karl Marx* avenue*, Novosibirsk, Russia*
shevcova_yuliya@mail.ru, t.monastyrskaya@mail.ru, alex.poletaykin@gmail.com, glebtoropchin@mail.ru

Keywords: Digital Maturity, Digital Transformation, Adaptive Technique.

Abstract: The paper reviews the digital transformation (DT) management technology for an organisation at the stage of evaluating its digital maturity. A characteristic feature of DT consists in profound changes in approaches to management, corporate culture, external communications, as well as an abrupt increase in efficiency. The relevant goal is to estimate the progress of DT and record the indicators of this progress in order to improve the efficiency of the further DT, as well as secure the transparency of an organisation in terms of its digital maturity. We present the developed system of estimating the digital maturity of an educational organisation. The objective of estimating digital maturity has been formalised thanks to analysing the existing approaches to validation of the digital maturity for an organisational object. The utilisation of the model developed is demonstrated on the example of higher education in the Russian Federation. The novelty of the given study is in introducing the mechanism of integral estimation calculation of digital maturity into the digital maturity evaluation methodology. The weights in this modified weighted average estimation are assigned to indicators of DM pronouncement level in such a way so that the integrated estimation could be transformed into a 100-point grading scale. The technique is adaptive and can be applied not only in educational organisations but also in organisations in other spheres of activity.

1 INTRODUCTION

The development of digital economy poses new challenges for education given the conditions of instability and uncertainty. The changes are so swift that the managerial decisions have to be made rapidly whereas innovations should be promptly introduced into the educational process. Transaction costs, linked to the implementation of the aforementioned transformations and innovations, also escalate. All this combined determines the need for DT as a process of permanent digitalisation of all the procedural components in the educational activities, which leads to minimising the costs referred to above thanks to creating and using digital services and platforms.

In general, maturity models as a software engineering method have already been applied in a plethora of spheres, from healthcare to education [1]. There is no single established definition of digital maturity in the scientific literature; nevertheless, it is possible to single out some basic features most authors agree upon. As such, I.V. Aslanova and A.I. Kulichkina [2], having analysed the vast accumulated experience, define digital maturity as a "gradual process of integration and implementation of [organisation] processes, human, and other resources into digital processes and vice versa". When applied to the realm of higher education, digital maturity most of all affects teaching and learning, curriculum, and management, among other things [3].

Manifold models of estimating digital maturity have been offered by the specialists worldwide. T. Thordsen, M. Murawski and M. Bick present a comprehensive review [4] of 17 related models, and come to a conclusion that most of those do not meet necessary standards. The researchers imply that it is

vital to define the semantics within digital maturity as a notion, also stating explicitly that the transparency of data collection is of great importance from an empirical point of view. M.V. Kupriyanova and co-authors admit that the available information on the existing models of digital maturity may not be very well-developed in terms of terminology, whereas sometimes it is habitual to concentrate on digital readiness or digital dividends instead. As an alternative, their work advances the concept of hierarchy analysis [5]. Croatian experts, V. Đurek, N. Begičević and R.N. Kadoić [6], employ the DSR (design science research) paradigm in their respective model specifically for universities. It incorporates seven layers, including leadership, planning and management, ICT resources and infrastructure, ICT culture and several others. M. Al-Ali and A. Marks, in turn [7], build their model on multiple instruments, including survey, interviews and direct observation. They pay special attention to the discrepancy between the DM requirements and practicalities of their implementation.With regard to case studies applying some of the models, R. Doneva, S. Gaftandzhieva, and G. Totkov [8] use their UniDigMaturity model for assessing the situation in Bulgarian higher educational institutions. They claim that even though their model makes allowance for national context, it can be easily adapted to any country and the desired level of educational system. K. Hummel and B. Schenk [9] provide their experience pertaining to a university of applied sciences, namely UAS in Baden-Württemberg. It was found that on a scale of 0 to 4, the UAS displayed an indicator of digital transformation at just 1.4 (the study had nine dimensions on the whole). H. Keshavarz and Ya. Norouzi [10] concentrate on digital maturity of university libraries, and they present their MMDIM (Maturity Model for Digital Information Management), including 5 levels, 10 dimensions, 20 components. In their study, the authors conclude that most of the organisations – in this case, libraries – are at Level 3 of DM.

The analysis of approaches to estimating digital maturity (DM) made it possible to single out two most interesting techniques allowing one to determine the DM of an organisation.

The first method is suggested by the Institute of Digital Development of Science and Education, FSAEI HE “MIPT”, and is described in [11]. The project of a digital passport for a HEB (higher educational body) under development includes 42 scalar indicators, distributed across five layers: Users and services; Information systems; Data control; Infrastructure; HR. All the indicators are formalised, characterising a basic – technical and engineering – level of digitalisation and basically rely on the requirements of governing and regulatory authorities. Along with that, the utilisation of the technique under discussion necessitates a well-organised system of data collection for calculations. This, in turn, requires a developed integrated information system of university management, which is not frequently found, for instance, in Russia, as it demands significant investments.

The second technique has been developed by the Consulting Analyst Company “Center for Advanced Governance” and has been tested in M.K. Ammosov North-Eastern Federal University [12]. This technique includes 20 indicators, distributed across seven layers: Infrastructure and instruments; Organisational culture; HR; Processes; Products; Models; and Data. Considering the methodological particularities of the calculations, the indicators have a pronounced cognitive directionality and express as a whole the digital potential (ability) and the desire to implement it (readiness) by the entities of the educational sphere at the present level of digital technology potential in the organisation. An advantage of the method is the simplicity of data collection for further calculation through surveying key specialists and processing the statistical data.

The conducted analysis of the existing models of estimating digital maturity in the educational organisations and companies in other spheres, as well as digital transformation of economy, allowed the authors to form their approach to estimating the level of digital maturity in an organisation.

2 METHODOLOGY AND TECHNIQUES

The suggested model is based upon the methods formed by the Consulting Analyst Company “Center for Advanced Governance” [12]. In their respective work, the authors hammered out original indicators of a certain level of digital maturity in an educational organisation for every layer of digital level indicators, reflecting the special features of educational, administrative and R&D processes in terms of a HEB [13].

“Level 0 – Beginner” of a digital maturity indicator reflects the beginner level of DM. The characteristics of this level of digital maturity in a HEB are:

- ineffective automation of basic business processes in a HEB;
- underdeveloped digital infrastructure;
- data handling is only limited to meeting the requirements as per regulatory legal acts;
- a low level of digital competencies among the students, academic staff, and administrative personnel.

A zero level of digital maturity in a HEB limits the potential of its development due to ineffective automation of basic business processes in a HEB and underdeveloped digital infrastructure that does not make it possible to implement the digital transformation projects. This results in a HEB's falling behind compared to other educational organisations with a higher level of DM, which, thanks to digital technologies, improve their efficiency. That means they also gain traction in terms of their attractiveness for parties in interest (companies, state, and students).

"Level 1 – Basic" determines the level of automating the processes, i.e. the implementation of IT solutions reproducing the existing processes. The characteristics of this level of digital maturity in a HEB are:

- non-systemic (discrete) optimisation of business processes in a HEB;
- a low level of digital infrastructure development;
- a low level of work culture in data handling;
- lack of systemic actions aimed at developing digital competencies in students, academic staff, and administrative personnel.

HEBs at "Level 1" of digital maturity are only entering the process of digital transformation and have not yet reached the primary effects of implementing their digital transformation strategies, which become more pronounced at the later stages of digital maturity, such as a better quality of rendering services, or decreasing labour costs etc.

"Level 2 – Advanced" corresponds to the stage of the process digitalisation in an organisation where the existing processes are ameliorated thanks to implementing IT solutions, their re-engineering and optimisation, whereas decisions are made based on data analysis. The characteristics of this level of digital maturity in a HEB are:

- preliminary optimisation of basic business processes thanks to orderly inoculation of services into the HEB's activities;
- modernisation of the existing infrastructure;
- introduction of data-driven management;
- digital capacity building in students, academic staff, and administrative personnel.

A HEB at "Level 2" of digital maturity can be recommended to use best practices aimed at digital maturity, develop the existing infrastructure for subsequent expansion of their basic business processes, take action with a view to further build digital capacity among students, academic staff, and administrative personnel, continue their transition to data-driven management.

Finally, "Level 3 – Perfect" models the state of actual digital transformation, where the activities of an organisation are permeated by novel processes, products and models with conceptually new properties. The characteristics of this level of digital maturity in a HEB are:

- a high level of basic business processes optimisation thanks to introducing services in most of the business processes in the HEB activities;
- well-developed digital infrastructure;
- a high level of work culture in data handling;
- a high level of digital competencies in students, academic staff, and administrative personnel.

HEBs that reached a top level of digital maturity are capable of providing effective management, improving the quality of educational and scientific activities thanks to creating a unified digital environment provided with services, implementing new forms of organising basic processes, based on data management. HEBs that reached the given level, are recommended to develop novel models of managing their basic business processes taking into account the capabilities acquired in the process of digital transformation, bring the existing services to the level of an ecosystem, implement consultancy and methodological support for other players in the area [13].

Besides, we upgraded the model of DM estimation developed by "Center for Advanced Governance" with two more original indicator layers: "Global digital environment" and "Personality factor".

The "Global digital environment" indicator layer reflects the degree of digital unity between HEB activities and its external relations, as well as the degree of HEB's belonging to the global digital educational and research environment. The introduction of such a layer is predetermined, among other factors, by the fact that it is "creation, development and exploitation of IT infrastructure and information systems in the sphere of science and

higher education in Russian Federation" that is one of the priority directions of digital transformation of science and higher education suggested by Russian Ministry of Science and Higher Education [14].

The "Personality factor" indicator layer in turn reflects the degree of intolerance to digital immaturity of certain processes and activities, degree of impact of digitalisation processes in a HEB on employees' personal development, degree of democratism in HEB's digitalisation processes, as well as the degree of adequate understanding of ethical and social aspects of digitalisation in education and science by HEB employees.

As such, the suggested technique of estimating the level of DM in an educational organisation is structured as the following model (Table 1).

This leaves open the question of how to calculate the integrated level of digital maturity in an organisation. As such, in the model developed by the "Center for Advanced Governance" (upon which our technique is based), defining an overall level of digital maturity is not envisioned at all, whereas visual representation of the estimates obtained is only implemented in the form of a radar chart [12].

We suggest determining an integrated indicator of organisations' digital maturity (*DM*) using the following (1):

$$DM = \sum_{i=1}^{n} \sum_{j=1}^{m} k_{ij}^{l} x_{ij}, \; l = \overline{0,3}, \tag{1}$$

where x_{ij} is a variable of the *j*th reply of a respondent to the *i*th question in the questionnaire: $x_{ij}=1$, if a respondent attributed their *j*th answer to the *i*th question, $x_{ij}=0$ if vice versa; k^{l}_{ij} is the weight of the *j*th answer to the *i*th question; *l* is the index of digital maturity level in an organisation.

Table 1: Structural model of estimating the digital maturity of an organisation.

Layer of DM indicators	DM indicators
Organisational culture: Support of constant advancement and innovation processes facilitating effective change control	Developed digital tools for task management
	Effectors' pro-activeness when managing tasks
	Inter-operational control and assessment of results
Competencies: Personnel possessing competencies necessary for successful work in the digital economy environment	Development level of digital competencies among the staff
	Proficiency in using digital and analytical tools
	Maturity of the approach to developing digital competencies
Processes: Process-based management practices: methods of optimising processes, lean management, design thinking; monitoring processes and constant updates	Process management maturity
	Opportunities to optimise processes
	Degree of process automation
Products: Analysis of existing digital projects, their requirements and related activities	Participation in the creation of digital projects
	Managing digital products requirements
	Applying digital technologies in product creation
Models: Using various types of analytical models, updating them constantly, ensuring their validity and applying the results in the processes	Degree of proficiency in analytical approaches
	Degree of learning trajectories digitalisation
Data: Access to data for real-time decision-making taking into account their integrity, quality and safety for work	Degree of data classification
	Data processing performance level
	Data quality
Infrastructure and instruments: Access to modern digital infrastructure and maintaining workability on all types of devices	Workplace engineering
	Existence of developed digital services for the personnel
	Safety and security arrangements
Global digital environment: Access to modern global digital educational and research environment	Degree of digital unity
	Clarity of understanding one's belonging to the global digital environment
Personality factor: Employees' ability to embrace positive and constructive digital transformation	Degree of intolerance to digital immaturity of processes
	Degree of democratism in digitalisation processes
	Degree of impact of digitalisation processes on personal development
	Degree of adequate understanding of ethical and social aspects of digitalisation in education and science

The (k_l) weights are expertly set to correspond the answers and serve as normalising factors, bringing the integral results to certain values of a DM level (2):

$$k^l = \frac{\mathrm{DM}_l}{n}, l = \overline{0,3}. \quad (2)$$

As such, we deem it natural to use the following model parameters: DM_0=0 points is "Level 0 – Beginner"; DM_1=33.3 points is "Level 1 – Basic"; DM_2=66.7 points is "Level 2 – Advanced"; and DM_3=100 points is "Level 3 – Perfect".

3 RESULTS

The developed technique of estimating the DM was approbated at the Faculty of Computer Technology and Applied Mathematics, Kuban State University (KubSU). A total of 25 faculty members took part in the survey. Among those surveyed are primarily members of Department of Data Analysis and Artificial Intelligence, Associate Professors aged from 41 to 50, having worked at the university for less than 10 years. Figure 1 presents the level of digital maturity at the KubSU Faculty of Computer Technology and Applied Mathematics.

The results obtained allow us to make a conclusion that the DM estimations at most of the layers (8 out of 9), as well as the integrated DM at the faculty on the whole, fit the interval of a basic maturity level, i.e. slightly above average (55 points). The interpretation is as follows:

1) The digital support of advancement and innovation processes is hardly implemented at the faculty (42 points). The penetration rate of digital tools (34 points) and follow-up action regarding the goals set (32 points) is barely at the described beginner level.

2) The academic staff skills mostly correspond with the basic level of digital competencies, according to the Plan [14]. HR tools outreach to the staff (61 points) and data tools proficiency (58 points), are still not sufficient, though.

3) The adoption of process-based practices is quite well implemented (58 points). The degree of workers' understanding of the corresponding processes (including their content) is sufficient (68 points). However, the level of process optimisation and automation (45 points) leaves a lot to be desired.

4) The analysis of digital products and involved activities existing at the faculty demonstrates an average level of DM (48 points). Of concern is an insufficient penetration rate of digital products development tools (40 points).

5) The application of analytical tools and mathematical models in the activity processes, given the specificity of the faculty, is in general almost sufficient (58 points), which is mainly reflected upon the sufficient degree of their understanding by the academic staff. However, their application for organising educational activity is clearly underdeveloped (52 points for analysing data in one's principal field and 45 points for models of forming students' individual educational trajectories).

6) Data handling is also at an average level of development. Data completeness (56 points) and quality (61 points) for decision-making also suffer from a certain insufficiency. Similar to the "Processes" layer, the level of optimisation and automation in data processing (36 points) leaves a lot to be desired.

7) The access to modern digital infrastructure is also at an average level (49 points) and requires additional development in terms of creating extra automated jobs and digital services for the academic staff. The factor of cybersecurity appears essentially underdeveloped (36 points).

8) The maturity of global digital environment at the faculty is clearly insufficient (47 points), especially at the actual external communications level (36 points), which should be developed preferentially.

9) Up to the mark is the "Personality factor" layer (70 points), which is declarative of the sufficient HR quality at the faculty and encourages optimism regarding further digital development of the organisation. Of concern is the underdevelopment of the teamwork at the university level in terms of the expediency of its digital transformation (60 points).

The obtained digital maturity level estimations of layers and their separate indicators make it possible to single out priority directions of digital transformation in an organisation, which is especially important bearing in mind permanent limitations on various kinds of resources (including financial, human, material resources etc.) As such, for example, in the department under consideration (KubSU Faculty of Computer Technology and Applied Mathematics) the "Organisational culture" layer has the lowest level of digital maturity. The head of the Faculty mainly allocates tasks and controls their execution through primitive instruments at the minimum level of digitalisation (in this case these are emails, messengers or phone calls).

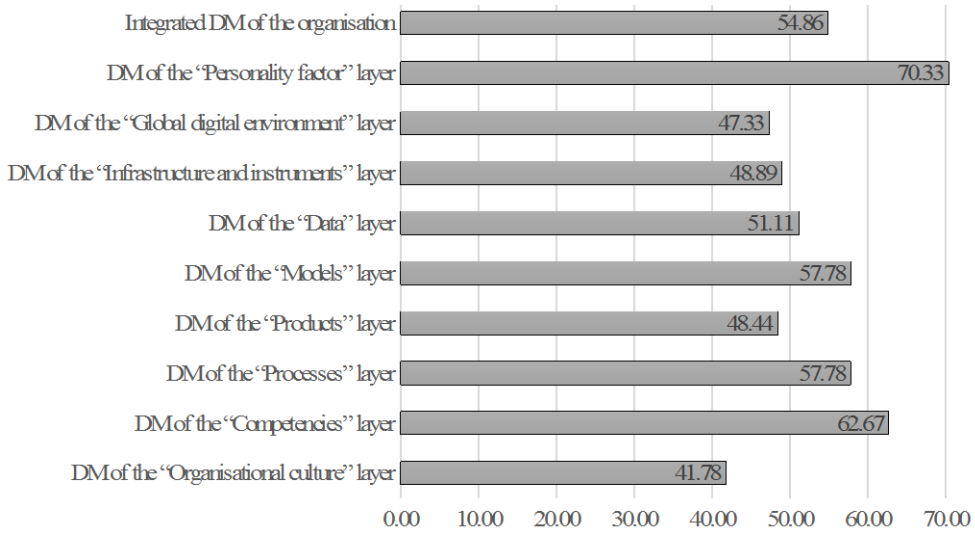


Figure 1: Digital maturity profile for the Faculty of Computer Technology and Applied Mathematics, KubSU, in accordance with the developed technique.

Consequently, the development of digital instruments and technologies of organisational interaction is seen as a priority direction of digital development at the KubSU Faculty of Computer Technology and Applied Mathematics when administering tasks of primary activities.

Therefore, estimating digital maturity in an organisation is, in fact, its internal check-up, which allows one to estimate its growth potential, single out priority development directions and elaborate an individual strategy for its digital transformation.

All in all, one can notice the existence of a definite basis for digital development at the faculty, a certain inefficiency of which should be compensated for thanks to constructive workload of highly-qualified human resources with sufficient supply of material resources.

4 CONCLUSION

The novelty of the given study is in introducing the mechanism of integral estimation calculation of digital maturity into the digital maturity evaluation methodology. This will allow one to draw a correct comparison between organisations or their separate departments, having different indicators in terms of scope of their activity. Apart from that, a change in an overall DM level of an organisational object in its dynamics can be regarded as an efficiency indicator in terms of digital transformation it is undergoing, which is a relevant applied research task of project management.

An application of the given description of calculating intergated result not just allows one to obtain a folding of estimations for individual indicators. In the description we suggest an integral indicator represents a modified weighted mean estimation of DM, where weights are assigned to replies but not questions, as it has been traditionally done in case of various techniques, including the mentioned method that we use as a basis for our work [12].

The suggested approach to estimating the digital maturity level of an organisation was endorsed at Faculty of Computer Technology and Applied Mathematics, Kuban State University. In general, its results demonstrate the presence of a certain basis for digital development of the faculty. The insufficiency of this basis should be compensated thanks to increasing constructive workload of high-quality HR coupled with adequate material support.

Even though we sought to estimate maturity of an educational organisation when developing our technique, the latter is adaptive and can be applied in organisations concerned with other areas of activities.

REFERENCES

[1] E. Tocto-Cano, S. Paz Collado, J.L. López-Gonzales, and J. E. Turpo-Chaparro, "A systematic review of the application of maturity models in universities,"

Information (Basel), 11(10), 2020, p. 466, [Online]. Available: https://doi.org/10.3390/info11100466.

[2] I.V. Aslanova, and A.I. Kulichkina, "Digital Maturity: Definition and Model." In Proceedings of the 2nd International Scientific and Practical Conference "Modern Management Trends and the Digital Economy: From Regional Development to Global Economic Growth" (MTDE 2020), Paris, France: Atlantis Press, 2020, [Online]. Available: https://doi.org/10.2991/aebmr.k.200502.073.

[3] M. Alenezi, "Deep Dive into Digital Transformation in Higher Education Institutions," Education Sciences 11 (12), 2021, p. 770, [Online]. Available: https://doi.org/10.3390/educsci11120770.

[4] T. Thordsen, M. Murawski, and M. Bick, "How to Measure Digitalization? A Critical Evaluation of Digital Maturity Models," In Lecture Notes in Computer Science, 2020, pp. 358–369. Cham: Springer International Publishing, [Online]. Available: https://doi.org/10.1007/978-3-030-44999-5_30.

[5] M.V. Kupriyanova, E.N. Evdokimova, I.P. Solovyova, and I.P. Simikova, "Methods of Developing Digital Maturity Models for Manufacturing Companies," E3S Web of Conferences, 224, p. 02034, [Online]. Available: https://doi.org/10.1051/e3sconf/202022402034.

[6] V. Đurek, N. Begičević, and R.N. Kadoić, "Methodology for Developing Digital Maturity Model of Higher Education Institutions," Journal of Computers, 14 (4), 2019, pp. 247-256, [Online]. Available: https://doi.org/10.17706/jcp.14.4.247-256.

[7] M. AL-Ali, Maytha, and A. Marks, "A Digital Maturity Model for the Education Enterprise," Perspectives Policy and Practice in Higher Education, 2021, pp. 1-12, [Online]. Available: https://doi.org/10.1080/13603108. 2021.1978578.

[8] R. Doneva, S. Gaftandzhieva, and G. Totkov, "Digital Maturity Model for Bulgarian Higher Education Institutions," In EDULEARN19 Proceedings, IATED, 2019, doi:10.21125/edulearn. 2019.1474.

[9] K. Hummel and B. Schenk, "Digital maturity in the administration of a university of applied sciences," In Proceedings of the Central and Eastern European E|Dem and E|Gov Days, May 2-3, 2019, Wien, Austria, pp. 307-318, doi: 10.24989/ocg.v335.25.

[10] H. Keshavarz and Y. Norouzi, "A Maturity Model for Digital Information Management in University Libraries: A Design Science Study," The International Information & Library Review, 2022, pp. 1-16. [Online]. Available: https://doi.org/10.1080/10572317. 2021.2022388.

[11] Kadry dlja cifrovoj jekonomiki: federal'nyj proekt nacional'noj programmy "Cifrovaja jekonomika RF". (In Russ.), [Online]. Available: https://digital.gov.ru/ ru/activity/ directions/866/ (accessed 01.10.2021).

[12] Metodika raschjota indeksa cifrovoj zrelosti obrazovatel'nyh organizacij vysshego obrazovanija. II Mezhdunarodnyj IT-forum s uchastiem stran BRIKS i ShOS (In Russ.), [Online]. Available: https://www.youtube.com/ watch?v=tQVEaGyhX3Y (accessed 01.10.2021).

[13] Metodicheskie rekomendacii po razrabotke strategii cifrovoj transformacii obrazovatel'noj organizacii vysshego obrazovanija, podvedomstvennoj Ministerstvu nauki i vysshego obrazovanija Rossijskoj Federacii. Moscow, Ministry of Science and Higher Education of the Russian Federation, 2021, p. 28.

[14] Plan dejatel'nosti Ministerstva nauki i vysshego obrazovanija Rossijskoj Federacii na period s 2019 po 2024 god. Oficial'nyj sajt Minobrnauki Rossii (In Russ), [Online]. Available: http://fgosvo.ru/ uploadfiles/prikaz _miobr/Plan_deyatelnosti_2019-2024.pdf (accessed 01.10.2021).

The Clustering and Fuzzy Logic Methods Complex for Big Data Processing

Larysa Globa, Rina Novogrudska and Andrii Liashenko
Igor Sikorsky Kyiv Polytechnic Institute, 37 Peremohy avenue, Kyiv, Ukraine
lgloba@its.kpi.ua, rinan@ukr.net

Keywords: Fuzzy Logic, Clustering Algorithms, Smart System, Statistical Numerical Data, Fuzzy Knowledge Bases, Fuzzy Logical Rules.

Abstract: Currently, telecom operators are facing a problem that is conditionally called "Big Data". The telecom industry is growing rapidly and dynamically, new technologies are emerging (IoT, M2M, D2D, P2P), new companies are using them, new information and communication services are being introduced to automate production processes, and so on. Methods of statistical analysis, A\B testing, data fusion and integration, Data Mining, machine learning, data visualization are used in the Big Data processing and analysis, but due to the fact that large amounts of Big Data are not structured, come in real-time with various delays related to bandwidth and network congestion, in each case the processes of processing and analysis of Big Data are extremely costly in terms of time and resources. As a result, telecom operators need not only to process large amounts of data but also to extract knowledge from them. However, the analytical processing of large data is characterized by blurred boundaries, which determine certain logical relationships between data. This study proposes the flexible complex of clustering and fuzzy logic methods for big data processing, which increases the speed and reliability of their processing in network nodes, as well as an architectural solution for analysis and processing Big Data realization using micro-services, which increases system scalability and reduces the load on the servers that process them. Experimental studies have confirmed the effectiveness of the proposed modifications. Studies of the K-means algorithm when processing 1500 rows in 3 columns showed decreasing in execution time by 2 seconds. Studies of the Fuzzy C-means algorithm have shown a reduction in execution time by almost 2 times. The validity of the developed fuzzy knowledge base for the K-means and fuzzy C-means algorithms increased by 9% and 4%, respectively.

1 INTRODUCTION

The rate estimation of in data volume increasing in communication networks is determined by the following trends: population growth [1], increasing of the number of mobile users, the number of Internet users and the number of social network users. These trends are driving an ever-increasing amount of content and data in the digital space. According to the source [2], the amount of data generated every second is more than 30,000 gigabytes. At the same time, improving the network infrastructure of information platforms for the provision of modern digital services is quite a complex and time-consuming and costly task. The state of the current infrastructure of telecom services requires the knowledge extraction from statistical data sets to effectively support and process significant amounts of information from both users and from various services. Currently, analytics [3] are used to obtain some knowledge from statistical data sets, but due to the fact that large amounts of big data are not structured, they come in real time with various delays associated with bandwidth and network congestion, simple statistical analysis of data is not enough. It is necessary to apply a set of methods that would allow to process, analyze data and form knowledge from them, using different modifications of clustering algorithms, to form logically connected groups of data that define logical dependencies. Choosing the right clustering algorithm is an important step in this process. K-means clustering algorithm is one of the most popular and simplest clustering technologies used in practice [4,5,6]. But the usual K-means algorithm has problems with the accuracy of cluster center selection. This algorithm requires solving the

problem of initialization of cluster centers and finding the right number of clusters [7].

At the same time, Big Data is characterized by fuzzy and requires the participation of experts during their analysis. Based on this, it is proposed to analyze them using fuzzy logic, forming fuzzy logical statements (rules) such as *IF… AND… THAN*, which are best for human perception. However, it is not possible to use a single method or algorithm to develop fuzzy logic rules. This process requires the creation of reliable sets of clusters from which fuzzy logical rules are formed, and to achieve this it is necessary to use clustering algorithms, construction of membership functions, formation of fuzzy logical rules for the transition from numerical data to logical statements.

The stages of developing the system for the formation of sets of fuzzy logical rules (fuzzy knowledge base) are determined by the following stages: expert estimation - loading of pre-cleared and structured statistics, the transition from data clusters to membership and union functions, and hence the transformation of numerical values into terms of linguistic variables (formation of fuzzy rules), checking the correctness of the model. In addition, the system for forming a fuzzy knowledge base should be flexible, take into account the peculiarities of data flows and increase the efficiency (speed and reliability) of processing large amounts of data. In this study, this is achieved by using a set of methods for constructing fuzzy logical rules based on clustering algorithms that focus on the features of statistical data sets processed in telecommunication systems, as well as architectural solutions for development Big Data analysis and processing using microservices.

The paper is structured as follows: Section 2 contains state of art analysis of Big Data processing methods problem. Section 3 explains the approach to be solved by proposed flexible complex of clustering and fuzzy logic methods for Big Data processing. Section 4 introduces the approach for the fuzzy knowledge base development. Section 5 presents the approach for the architecture development of the system based on fuzzy knowledge base. Section 6 shows efficiency of the proposed flexible complex of clustering and fuzzy logic methods for Big Data processing usage. Section 7 includes the summary and outlook on future work.

2 STATE OF THE ART AND BACKGROUND

Features of Big Data such as the speed of data generation, the complexity of their analysis has necessitated the use of machine learning and artificial intelligence. Along with the evolution of data analysis computer methods, their analysis is also based on traditional statistical methods. Processing and analysis of Big Data is performed in the conditions of streaming data as they appear, and then apply various methods of data analysis as they are created, to find behavioral patterns and trends. As the amount of data increases, so are developed the methods used to process it.

Methods of Big Data processing and appropriate tools analytics are classified [8]:

1) A / B testing - these data analysis tools involve comparing the control group with different test groups to determine which ways to influence or change will improve a given objective variable.

2) Data merging and integration is a common tool used in Big Data analytics. Different data from different sources are integrated together and data analysis is performed by combining methods of statistics and machine learning in database management. An example of such an analysis in the telecom industry is the collection of customer data to determine which customers are most likely to a proposal respond.

3) Data Mining - a method of data analysis designed to search for previously unknown patterns in large arrays of information. These patterns make it possible to make effective management decisions and optimize business processes. Data Mining methods include teaching associative rules, classification, cluster analysis, regression analysis, detection and analysis of deviations, and more.

4) Machine learning - methods of artificial intelligence, which are a method of data analysis that allows the machine, robot or analytical system to conduct independent learning by performing a group of similar tasks.

5) Artificial neural networks are mathematical models, as well as their software or hardware implementation, built on the principle of organization and functioning of biological neural networks - networks of nerve cells of a living organism.

6) Visualization of analytical data - a means of presenting statistical information in a form that is better perceived by humans, in the form of diagrams, drawings using animation to determine the results of the expert or for further analysis.

All these methods of processing Big Data and extracting patterns from them are characterized by the fact that they depend on the structure and type of data, for example, if signals can be analyzed visually, then such methods are not effective for other data. Since searching for previously unknown patterns in large amounts of information provides some knowledge about the behavior of a system from sets of statistical data, these methods are promising for the telecommunications industry.

In recent years, cluster analysis is widely used in Data Mining as one of the main methods [9, 4]. The purpose of cluster analysis is to assign to objects to homogeneous data sets (clusters). Assigning objects is done so that the objects are similar to each other in one cluster and different in others. The method of summarizing the observed data in clusters is determined on the basis of statistical information that can be stored in database tables or files, because at the beginning of the study there is no prior knowledge.

There are many clustering algorithms, the paper [10] proposes the structure of algorithm categorization. Different clustering algorithms can be broadly classified as follows:

Separate algorithms: in such algorithms all clusters are defined quickly. The initial groups are specified and redistributed into associations. In other words, distribution algorithms divide data objects into several partitions, where each partition is a cluster. These clusters must meet the following requirements:

- each group must contain at least one object,
- each object must belong to exactly one group.

For example, in the K-means algorithm, the center is the mean of all points and coordinates, which is the arithmetic mean. There are many other partitioning algorithms such as K-modes, PAM, CLARA, CLARANS and FCM.

Hierarchical algorithms: data is organized in a hierarchical way depending on proximity. Proximity defines intermediate nodes. The initial cluster is gradually divided into several clusters as the hierarchy continues. The process continues until the stop criterion is reached (often determined by the number of k clusters). However, the hierarchical method has a serious drawback, which is that once a step (merger or division) is performed, it cannot be undone. BIRCH, CURE, ROCK and Chameleon are some of the well-known algorithms in this category.

Density Based: Here, data objects are divided based on their density, connectivity, and boundary areas. A cluster is an associated dense component that grows in any direction that determines density. Density-based algorithms are able to detect clusters of arbitrary shape. The total data point density is analyzed to determine the functions of the data sets that affect its assignment to a particular cluster. DBSCAN, OPTICS, DBCLASD and DENCLUE are algorithms that use this method to filter noise and detect arbitrary clusters.

Grid-based: The space of data objects is divided into grids. The main advantage of this approach is its fast-processing time, because the data set is passed once to calculate statistical values for grids. Collected grid data makes grid-based clustering methods independent of the data objects number that use a single grid to collect specific statistics and then perform clustering in the grid rather than directly in the database. The performance of a grid-based method depends on the size of the grid, which is usually much smaller than the size of the database. However, for very irregular data distributions, the use of a single homogeneous grid may not be sufficient to obtain the required clustering quality or to meet processing time requirements. Wave-Cluster and STING are typical examples of this category.

When defining clustering methods, it is necessary to use specific criteria to assess the relative strengths and weaknesses of each algorithm in relation to the multidimensional properties of Big Data, the most important of which are volume, speed and diversity.

Volume: refers to the ability of a clustering algorithm to work with large amounts of data. To control the selection of the appropriate clustering algorithm for the Volume property, the following criteria are considered:

- data set size;
- high dimensional processing;
- processing of emissions / noisy data.

Variety: refers to the ability of a clustering algorithm to process different types of data (numerical, categorical, and hierarchical). To control the selection of the appropriate clustering algorithm for the Variety property, a criterion such as data set type is considered.

Velocity: refers to the clustering algorithm speed during the big data processing. To control the selection of the appropriate clustering algorithm for the Velocity property, the following criteria are considered:

- complexity of the algorithm;
- performance at runtime.

Clustering algorithms perform effectively with either numerical data or categorical data; most of them do not process well mixed categorical and

numerical data types, with large size and data set, and in case of errors. As the number of measurements increases, the data becomes sparser, so measuring the distance between pairs of points becomes impractical, and the average density of points anywhere in the data is likely to be low. The process of clustering large amounts of data takes too much time, which can be impractical. The K-Means and FCM (Fuzzy C-Means) algorithms are among the most efficient algorithms that meet the requirements for processing large amounts of numerical data in telecom systems. FCM clustering algorithm is more flexible than K-Means algorithm, it allows to form a base of fuzzy logical rules for business processes of telecom operators.

This research focuses on creating a set of clustering methods and fuzzy logic for Big Data processing, which take into account the peculiarities of statistical data flows, increase the speed and reliability of their processing in network nodes, and on developing the architecture of Big Data analysis and processing using micro-services. This increases the scalability of the system and reduces the load on the servers that perform their processing.

3 THE APPROACH TO DEVELOP THE FLEXIBLE COMPLEX OF CLUSTERING AND FUZZY LOGIC METHODS FOR BIG DATA PROCESSING

The Fuzzy C-means algorithm (FCM) is a separate algorithm, like the K-means algorithm. The main difference is that a point can belong to all centers of clusters, but with its degree of affiliation, which can range from 0 to 1. The higher the degree, the more probably it is that the object belongs to that cluster. The FCM clustering algorithm is more flexible than the K-Means algorithm because it uses some ambiguity as the value of the membership function. This allows to determine whether a sample object belongs to clusters with different degrees of membership without losing existing logical connections on cluster boundaries. This gives the possibility to realize the transition to fuzzy logical statements (fuzzy knowledge) [4, 11].

FCM algorithm convergence criteria:

For each element of the measurements sample, the sum of the its belonging degrees to the clusters should be equal to:

$$\sum_{i=1}^{c} \mu_{ij} = 1, \qquad \forall j = 1, \ldots . N,$$

The value of the affiliation degree is limited by the interval [0,1]:

$$\mu_{ij} \in [0,1], \forall i = 1, \ldots ., c, \;\; \text{i} \;\; \forall j = 1, \ldots . N$$

FCM clustering is performed by minimizing the objective function (1):

$$J = \sum_{i=1}^{n} \sum_{k=1}^{c} \mu_{ik}^{q} \, |x_i - v_k|^2, \qquad (1)$$

where

J – the objective function,
n – the number of objects in the data sample,
c – number of clusters,
μ – fuzzy membership value from the table,
q – fuzzy coefficient (value > 1),
x_i– the value of the *i-th* object in the sample,
v_k– the cluster center,
$||x_i - v_k|$-Euclidean distance determined by (2):

$$||x_i - v_k| = \sqrt{\sum_{i=1}^{n} (x_i - v_k)^2}. \qquad (2)$$

The calculation of the cluster center is determined by (3):

$$v_k = \frac{\sum_{i=1}^{n} \mu_{ik}^{q} x_i}{\sum_{i=1}^{n} \mu_{ik}^{q}}. \qquad (3)$$

The fuzzy membership table is calculated using (4):

$$\mu_{ik} = \frac{1}{\sum_{l=1}^{c} \left(\frac{|x_i - v_k|}{|x_i - v_l|}\right)^{\frac{2}{q-1}}}. \qquad (4)$$

Implementation steps:

Step 1: Setting the number of clusters, fuzzy parameter (*constant value> 1*), and stop parameter.

Step 2: Initializing the matrix of affiliation degrees.

Step 3: Setting the cycle counter *k = 0.*

Step 4: Calculating the centroids of the cluster, calculating the value of the objective function *J.*

Step 5: For each object and for each cluster calculating the value of membership in the matrix.

Step 6: If the value of *J* between successive iterations is less than the stop condition, then stop; otherwise set *k = k + 1* and go to step 4.

Step 7: Obtaining the membership matrix and the end of the algorithm.

However, this algorithm has significant shortcomings, namely the initial initialization of cluster centers and the correct number of clusters can affect the accuracy of the fuzzy logic rules development and subsequently on the construction of fuzzy knowledge bases [10].

Since the initial centers of clusters have a strong influence on obtaining the final sets of clusters, the

process of their formation depends on the choice of starting points as the initial centers of clusters [12]. As a rule, the initial centers of clusters are selected randomly, which directly affects the accuracy of cluster construction. If a cluster center is initialized as a "remote" point, it may simply not have associated points, and more than one cluster may be associated with a single cluster center. Similarly, more than one centroid can be initialized in one cluster, which will lead to poor clustering. Correct calculation of the initial centers of clusters allows to obtain more accurate and efficient groups of clusters, as well as to reduce the complexity of the clustering process over time. Studies conducted in [13] have identified as an effective way to initially initialize clusters methods K-means ++ with the simultaneous use of the method of "elbow", which avoids the initial centroids, which are located close to each other.

K-means ++ is similar to initializing random points because it randomly selects data points to use as initial centroids. However, instead of selecting these points uniformly at random, K-means ++ selects them sequentially so as to induce the initial centroids to be distributed. In particular, the probability that a point will be chosen as the starting center is proportional to the square of its distance to the existing starting centroids [14].

The "elbow" method is based on determining the sum of squares in the middle of clusters (WCSS - Within Cluster Sum of Squares).

$$WCSS = K_n \sum_{P_i \text{ in Cluster } k}^{n} distance(P_i, C_i)^2$$

A cluster that has a small sum of squares is more compact than a cluster that has a large sum of squares. The "elbow" method considers the nature of the change in the WCSS scatters with the increasing number of groups *k*. Combining all n observations into one group, we obtain the largest intracluster dispersion, which will decrease to *0* for $k \rightarrow n$. At some point, the decrease in this dispersion slows down - this happens at a point called the "elbow".

The disadvantage of the elbow method is that it measures only the general characteristics of clustering, but the algorithm of the "elbow" method is effective if the time to find the number of clusters is important to obtain the result [13]. Some clustering methods allow building fuzzy logical rules after the clustering process.

In the theory of fuzzy logic, true values of statements can take any value of truth from the interval of real numbers [*0*; *1*]. This provision allows building a logical system in which you can make approvals with uncertainty and assess the degree of truth of statements. One of the concepts of fuzzy logic is the concept of elementary fuzzy expression. In the set theory, an element either belongs to the set or not. The theory of fuzzy sets is based on the concept of partial belonging to the set: each element belongs to the fuzzy set partially. [15]. The fuzzy set is defined by the "membership function", which corresponds to the concept of "characteristic function" in classical logic. The membership function is an important element in fuzzy logic. On the one hand, it provides a convenient tool for analytically presenting the degree of membership of a given term, on the ordinate axis - *f(x)* is always delayed range from *0* (clearly does not belong) to *1* (clearly belongs) - the degree of membership, and the axis abscissa - quantitative indicators of the corresponding term. On the other hand, the membership function makes it possible to perform various operations on fuzzy sets [4]. The fuzzification procedure is to determine the degree to which a variable (for an example, measurement) belongs to a fuzzy set. The defuzzification procedure is to determine the numerical value of a variable based on the degree of its belonging to a fuzzy set. Fuzzy logic rule bases are the most commonly used tools in analytical software systems with fuzzy logic. They apply rules in the form such as: *IF "condition" THEN "result"*.

The main difficulty of the "Fuzzy inference" block is that it is not possible to form a rules base in advance due to unstructured data and their large volume. To develop a fuzzy knowledge base, the fuzzy inference mechanism is often used, which is called the Mamdani mechanism [4, 15,16]. In order to specify the number of fuzzy rules, the number of linguistic terms into which the input variables of statistical data are divided without an expert, it is necessary to identify the structure of the fuzzy system. Such identification is performed using fuzzy cluster analysis.

4 THE FUZZY KNOWLEDGE BASE DEVELOPMENT

The proposed approach to designing a fuzzy knowledge base works only with numerical data, which are presented in tabular form. This situation is typical for the technical infrastructure of telecom operators and is acceptable for fuzzy logic procedures in the fuzzification phase, which operates

with numerical data to determine the fuzzy value of the linguistic variable term.

The input data comes in the form of a table in CSV format. In this data structure, each column is an object property. And each line is an instance of the object state. The efficiency study of a set of methods for building a fuzzy knowledge base was conducted on the example of data on the processor's power consumption at different loads. These data were obtained from the Technical University of Dresden [17], where:

FR - the frequency with which the data is processed;

TR - number of threads;

EN - the energy that will be consumed by the processor (Figure 1).

	FR	*TR*	*EN*
1	*1300*	*2*	*637,727*
2	*2400*	*4*	*3448,939*
…	….	…….	……

Figure 1: A fragment of the table with input data.

From the point of view of the fuzzy model, the columns in the table are arranged so that the first *N*-columns (*FR* and *TR*) are conditions, antecedents, which are the left part of the fuzzy logical rule, and the last column (*EN*) is a consequent or fuzzy logical the conclusion on the right side of the fuzzy inference. The rule has the form *IF… AND… .TO*. The result of the fuzzy rule is a combination of propositions combined by "*TA*" operators. The "*OR*" instruction is not used to generate the result.

If tabular data are visualized, we get an *N*-dimensional space in which each instance of the state of the object, so the string will be displayed as a point, and each property of the object will be an axis (Figure 2) [18].

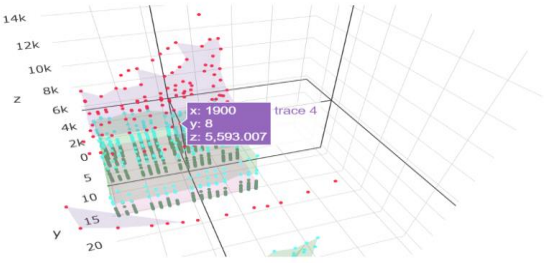


Figure 2: Representation of a point in space.

The next step is to determine that this structure is responsible for the linguistic variable rules and what is terms set of a certain linguistic variable. A rule contains a set of object characteristics, that is columns in a tabular structure. The set of these columns also determines the structure of the resulting linguistic rule. An object characteristic is a linguistic variable rule. A linguistic variable is an axis in *N*-dimensional space that represents term sets. *Term* - a description of the value of the column (for example, for data on the energy efficiency of computer servers - these will be the values for frequency, number of flows, and energy consumption in form as *low*, *high*, *medium*). The meaning of terms in the form of linguistic variables is determined by an expert.

Consider an example of the converting from numerical statistics, which have a tabular data structure, to a fuzzy set. In this case, the measuring space is a two-dimensional plane, in which the abscissa will be the value of the condition (antecedent), and the y-axis will be the final value (consequent) [18]. To move to ambiguity, it is necessary to find membership functions for each linguistic variable. The clustering procedure is performed that will divide all the data from the input space into cluster groups. After clustering, each cluster will be a term set (Figure 3).

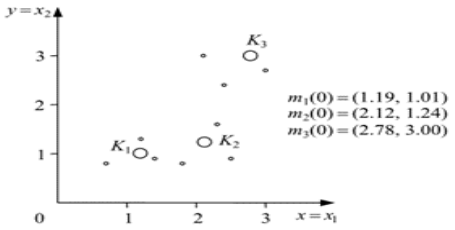


Figure 3: Clustering for two dimensional spaces.

The membership function is a function that has values from *0* to *1* on the ordinate axis, and on the abscissa axis the numerical values of this term. Membership functions can be of different types: triangular, trapezoidal or Gaussian. The proposed method uses the Gaussian function.

The Gaussian function has the form: $\mu(x) = e^{(-\frac{(m_{xi}-x)^2}{2\sigma^2})}$, where x – the center of a cluster, σ – standard deviation. The mathematical expectation for this function is the center of the cluster, and the standard deviation is the measure of the scatter of points near the center of the cluster. To find the width σ it's possible to use the $\frac{|m_{xi} - m_{x(i+1)}|}{N_x}$, where m_{xi} – the center of a cluster, $m_{x(i+1)}$ – the center of the next cluster, N_x – number of clusters.

After determining the membership functions, it is necessary to design them for each axis of the *N*-dimensional space of each cluster (terms) of the

Gaussian membership function of the form (Figure 4) [18,4].

Formation of fuzzy logical rules. After the step of building membership functions, each object in the input sample will create a separate new rule in the form *If.. AND… AND*. The mechanism of rule formation works so as to process all received statistical data [19, 4]. The system cannot have two identical rules, as this will use more computing resources or there will be conflicting inconsistencies, which is not correct. In Figure 5 shows the value of the Gaussian function (membership function) at a given point. Then the largest value of all values is determined, which indicates that the point refers to a fuzzy set (to the corresponding term).

Thus, for each numerical value of the string (object characteristics) there is a corresponding term set, which will be a fuzzy logical rule. The example is shown in Table 1.

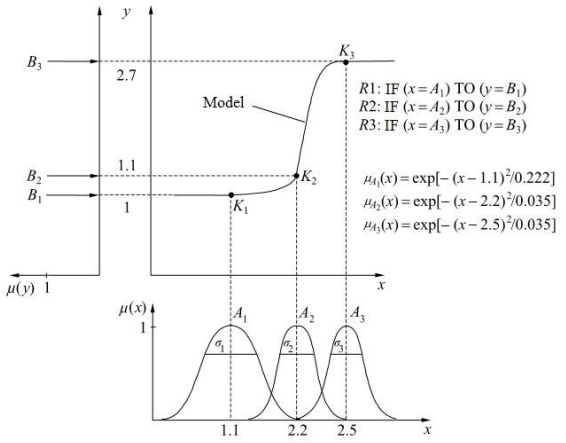


Figure 4: Designing of membership functions for each linguistic variable.

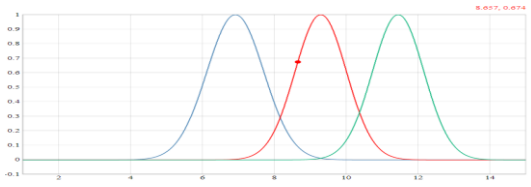


Figure 5: The value of the function at the appropriate point.

Table 1: The fragment of data converted into corresponding term sets.

FR	*TR*	*EN*
low	*low*	*low*
high	*low*	*high*
high	*low*	*low*
middle	*low*	*high*

Based on table1 the rules set in the form ***If… AND .. Then***:

IF ***FR –> low*** *AND* ***TR –> low*** *THEN* ***EN –> low***

IF ***FR –> high*** *AND* ***TR –> low*** *THEN* ***EN –> high***

IF ***FR –> high*** *AND* ***TR –> low*** *THEN* ***EN –> low***

IF ***FR –> middle*** *AND* ***TR –> low*** *THEN* ***EN –> high.***

At this stage, converting the data into terms of linguistic variables, conflicting and duplicate rules can form due to the different nature of the data or because the data during clustering is not properly distributed between clusters at the boundaries of the cluster distribution. In this case, you need to check the set of rules for correctness that can be done by analyzing the metagraphs.

Thus, the modified method for developing fuzzy logic rules for Big Data consists of the following steps:

1) Cleaning up the input data and presentation of them in the correct tabular structures;

2) Fuzzy clustering of the input data based on the FCM algorithm with the first primary initialization of the cluster's centers used the K-means++ algorithm and getting the correct number of the cluster's centers by the algorithm "elbow";

3) Formation of fuzzy logical rules for numerical data converting into a appropriate term-set.

4) The quality analysis of fuzzy logic rules development based on their visual analysis by the metagraphs theory.

Fuzzy logical rules should be as precise as possible, so clustering is an important element of methods complex. If at the final stage of clustering the centers of clusters are found incorrectly, the data will be incorrectly divided, because of it there will be an error in construction of membership functions. To prevent this, training procedures should be performed on different numerical data sets. As a result, the algorithmic and time complexity of the method will increase, but more important is the correctness and accuracy of developing fuzzy logical rules.

The proposed approach to Big Data processing has shown the possibility of designing fuzzy logical rules using numerical clustering methods from numerical statistical data sets to create a fuzzy knowledge base, which greatly simplifies the process of analyzing the effectiveness of telecom services infrastructure.

5 THE ARCHITECTURE OF THE SYSTEM BASED ON FUZZY KNOWLEDGE BASE

When designing real data analysis systems, there is a problem of high load of computing resources due to the large amount of data processed in the system. Microservice architecture is used in Big Data analysis systems to prevent congestion (Figure 6). Microservice architecture is an approach to creating a software package that involves the use of several small application services, each of which corresponds to a limited context. These software services run on different servers and interact with each other over the network, for example via HTTP [20]. The essence of microservice architecture is that each logical part of the system is allocated as a separate micro-service that can be easily connected and integrated into the system, regardless of what technology it uses in the implementation. In Figure 6 shows the architecture of the proposed system for processing Big Data based on the microservice approach.

Each part of the system is allocated as a separate project, which was hosted on a separate server [21].

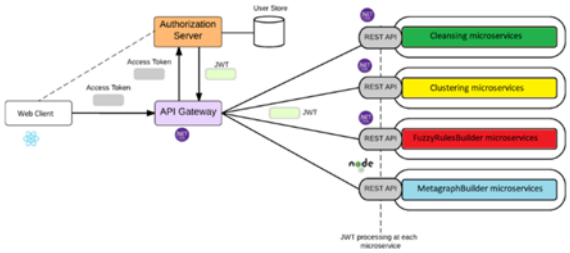


Figure 6: System architecture based on microservice approach.

The system is a website, the client part of which is written in Java-Script using the React library. The proposed architecture uses the OpenId Connect protocol. The Gateway API serves as a gateway between client services there, aggregates data from different services, and is responsible for the logic of executing service requests. The Gateway API performs load balancing, which distributes tasks across multiple network devices to optimize resource depletion, reduce query service time, scale cluster horizontally, and provide resiliency.

Cleansing microservice - a service for cleaning the data that the user uploads to the system. The service uses various algorithms to clean data from incorrect data. For example, the downloaded file may have different data emissions or incorrect format (there may be words instead of numbers). This service provides for data cleaning clustering algorithms. The service uses .Net core technology and is written in the C # programming language.

Clustering microservice - a service for clustering data received after data cleaning. This microservice uses K-means ++ and FCM clustering algorithms, as well as algorithms for initializing initial clusters and finding the number of clusters to analyze data and select similar objects into a homogeneous group. This step is required to construct membership functions for each feature of the object, and then to develop fuzzy logical rules using previously found membership functions. The service uses .Net core technology and is written in the C # programming language.

FuzzyRulesBuilder microservice - a service for designing fuzzy logical rules from statistics. It uses a method that finds the value of the Gaussian function at a point and relates this point to a fuzzy term of a linguistic variable, the service is realized in .Net Core technology.

MetagraphBuilder microservice is a service for displaying and visualizing a metagraph based on fuzzy logical rules. This service filters duplicate rules and builds a metagraph to verify the fuzzy rules of the knowledge base. The service uses Node.js technology.

All services use REST - a protocol for the interaction of components of a distributed system in the network. Steps of the business process of developing fuzzy logical rules from statistical data:

1) Uploading numerical statistics is a step in which the user uploads data to the system. This data must already have the appropriate tabular structure in the form of a CVS file.

2) Initial cluster initialization and finding the number of clusters is a system step in which the system finds the optimal input parameters for clustering before performing clustering: finding the correct number of clusters and performing initial initialization of cluster centers for input sampling.

3) Clustering of uploaded data is a system step in which the system clusters data with pre-selected FCM and K-means ++ clustering algorithms.

4) Initialization of terms for each linguistic variable is a step in which the user can initialize the terms of each linguistic variable or column from the input set. This stage runs after finding the correct number of clusters and performing the clustering procedure because the number of terms for each linguistic variable determines the number of clusters into which the input sample will be divided.

5) Designing of function graphs for each linguistic variable is a system step in which Gaussian membership functions are built. For each function, the mathematical expectation is the center of the cluster, and the standard deviation is the measure of the scatter of the data points around the cluster

6) Converting of statistical numerical data into appropriate term sets is a system step, at the stage of which the system determines for each number in the line the corresponding set of terms. This is done by finding the maximum value of the function at the point for this number.

7) Generating a JSON file with fuzzy logical rules is a system step, in which the system builds a fuzzy logical rule for each line of input statistics, and then passes it to the stage of removing conflicting and duplicate rules.

8) Checking the quality of development of fuzzy logical rules through their visual analysis in the form of metagraphs, retraining as needed.

6 THE EFFICIENCY OF PROPOSED CLUSTERING AND FUZZY LOGIC METHODS

To test the efficiency of constructing fuzzy logic rules in the proposed system, the efficiency is considered as the reliability of the construction of fuzzy logic rules and the time complexity of different clustering algorithms.

Reliability of results was performed by the method presented in [4] and the proposed modified method.

Algorithmic complexity depends on the amount of data received at the input of the clustering procedure. The time complexity of the algorithm K-means *O(ncdi)*, and FCM algorithm $O(ndc^2 i)$, where *n* is the number of points, input data, *d* is the dimension of space, *c* is the number of clusters, *i* - number of iterations for which clustering will be performed.

The experiment was performed with data in the two-dimensional plane. 1500 random points were pre-generated. The input data sequence is divided into 3 clusters by clustering algorithms K-means++ and FCM with different primary initializations of the starting points of the cluster centers. The number of clusters was chosen by the algorithm for finding the number of clusters, namely the "elbow" algorithm.

The number of iterations and the time spent running each of the algorithms for the same data in the same environment (on the same computer) were measured.

The results of the evaluation of efficiency are given in Table 2.

Table 2: Analysis of algorithmic complexity of K-Means and FCM algorithms.

	Primary initialization	Algorithmic complexity	Time spent (seconds)	Number of iterations
K-Means	Random	$O(ncdi)$	1.540	37
FCM	Random	$O(ndc^2 i)$	9.440	115
K-Means	kmeans++	$O(ndc^2 i + + ncdi + nd)$	0.033	8
FCM	kmeans++	$O(2ndc^2 i + nd)$	5.680	74

From the results we can conclude that the time complexity of the K-means algorithm is better than FCM, but the initial initialization greatly affects the final result of clustering, which reduces the time spent and the number of iterations. With the correct initialization, the K-means and FCM algorithms converge in much less time.

To verify the correctness of fuzzy logic rules development, it's needed to conduct an experiment in which there will be two samples of data: training and test. The test sample of data will contain ready-made and already formed fuzzy logical rules, and the training sample will contain only ordinary statistical data, from which fuzzy logical rules will be built by a modified method.

The test sample used data on energy efficiency of servers [17]. The expert used 3 terms for each linguistic variable. The terms had the following meanings: *low, middle, high*. Two experiments were performed using K-means and FCM clustering algorithms with initial K-means ++ initialization.

The training sample had 1,500 rows of columns, which contained the values of data processing frequency, number of streams and energy consumed by the server. The algorithm for finding the number of clusters showed that the number of clusters will be 3. The results of the experiment are given in Table 3.

Table 3: Results of the tests for reliability.

Algorithm	Number of clusters	Number of samples in the input sample	The number of correctly formed samples	Spent time, ms (milliseconds)	Rreliability, C (%)
K-Means (kmeans++)	3	1500	1257	5790	86, 4
FCM (kmeans++)	3	1500	1473	14540	98, 2
K-Means (random)	3	1500	1159	7950	77, 2
FCM (random)	3	1500	1421	35770	94, 7

Thus, it's possible to conclude that the modified method of constructing fuzzy logic rules has reduced the execution time for the two algorithms K-means and Fuzzy C-means, which are used in the proposed method. For K-means the time decreased by about 2 seconds, and for FCM the time decreased by about 2 times.

The reliability of built fuzzy logic rules increased for K-means and FCM algorithms by 10% and 4%, respectively. The reliability of developing fuzzy logic rules using fuzzy FCM is quite high for both methods (normal and modified), but the time complexity of this algorithm is greater than K-means. The choice of algorithm is determined by the characteristics of the data sets, based on the needs of analysis and subject area.

7 CONCLUSIONS

The analysis of characteristics, features and methods of Big Data processing allows to define:

1) Big Data are characterized by different sizes, are structured or unstructured, have different speed of their receipt, a significant amount is simultaneously obtained from different sources, belong to such information that is difficult to process using traditional processes and tools.

2) Some Big Data processing methods are not suitable due to poor structure (numerical data can be in multidimensional space) when managing computing processes in telecommunications nodes, so such data is recommended to analyze in the form of fuzzy logical rules that are close to human understanding.

3) The analysis of clustering algorithms allowed to determine the feasibility of using algorithms FCM, K-Means++ and the algorithm "elbow" to process numerical statistics, extract knowledge from them, so converting to a fuzzy knowledge base, which will be used at the stage of fuzzy inference.

4) A modified method of development fuzzy logic rules for Big Data processing is proposed, the feature of which is the use of algorithms for initialization of cluster centers and finding the number of clusters, using criteria such as reliability of fuzzy logic rules and computational complexity of algorithms.

5) The fuzzy knowledge base was trained on statistical numerical data, which allowed to increase the reliability of the designing of fuzzy logical rules and reduce the operating time of the proposed set of methods.

6) The software of the system and architectural solution for its development with the use of microservices is created, such solutions allow to increase the flexibility of Big Data processing processes and their productivity during data clustering, designing of fuzzy logical rules based on the proposed modified method.

7) The efficiency of the modified method is experimentally proved, which is confirmed by the fact that for K-means algorithm when processing 1500 rows in 3 columns the execution time decreased by 2 seconds, and for FCM execution time was reduced by almost 2 times. The reliability of the designed fuzzy knowledge base for K-means and FCM algorithms increased by 9% and 4%, respectively.

Future researches will focus on further study of the Big Data characteristics and approaches for their processing in information and communication systems, especially such as 5/6 G, and peculiarities of their processing based on data streams specifications.

REFERENCES

[1] Digital 2019: Global Internet use accelerates, [Online]. Available: https://wearesocial.com/blog/2019/01/digital-2019-global-internet-use-accelerates.

[2] M. Nathaz, W. James, Big data. Principles and best practices of scalable real-time data systems 1st Edition, 2015, pp. 185-192, [Online]. Available: https://www.amazon.com/Big-Data-Principles-practices-scalable/dp/1617290.

[3] E. Nada, Advances in Data Mining. Applications and Theoretical Aspects / E. Nada, E. Ahmed. // Big Data Analytics: A Literature Review Paper. – 14th Industrial Conference, ICDM 2014, St. Petersburg, Russia, July 16-20, 2014. Proceedings, Lecture Notes in Computer Science, Springer, pp. 214-227.

[4] Y. Buhaienko, L. S. Globa, A. Liashenko, and M. Grebinechenko, “Analysis of clustering algorithms for use in the universal data processing system”, in Proc. International scientific and technical conf. Open Semantic Technologies for Intelligent Systems (OSTIS-2020), Minsk, 2020, pp. 101-104.

[5] K. Hribernik, Z. Ghrairi, C. Hans, and D. Thoben, “Co-creating the Internet of Things - First experiences in the participatory design of Intelligent Products with Arduino”, in Proc. 17th International Conference on Concurrent Enterprising, Aachen, Germany, 2011, pp. 1-9.

[6] X. Lei, Z. Yan, and Y. ChunLi, “The application and implementation research of smart city in China”, in Proc. 2012 International Conference on System Science and Engineering(ICSSE), China, 2012, pp. 288-292.

[7] P. Fränti and S. Sieranoja, “How much can k-means be improved by using better initialization and

repeats?", Pattern Recognition, vol. 93, no. 2, pp. 95-112, 2019. doi:10.1016/j.patcog.2019.04.014.

[8] Concepts and Characteristics of Big Data Analytics, [Online]. Available: https://www.iunera.com/kraken/fabric/big-data/.

[9] M. Zgurovsky and Y. Zaychenko, "Big Data: Conceptual Analysis and Applications", Springer Nature Switzerlend, 2020, pp. 1-42.

[10] A. Fahad, N. Alshatri, Z. Tari et al., "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis", IEEE Transactions on Emerging Topics in Computing, vol. 2, no. 3, pp. 267-279, Sept. 2014, doi: 10.1109 / TETC.2014.2330519.

[11] S. Ghosh and S. Kumar Dubey, "Comparative Analysis of K-Means and Fuzzy CMeans Algorithms", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 4, no. 4, pp. 35-39, 2013.

[12] K. A. Abdul Nazeer and M.P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", in Proc. of the World Congress on Engineering 2009, London, 2009, ISBN: 978-988-17012-5-1.

[13] L. S. Globa, Y. M. Buhaienko, I. O. Ishchenko, and A. V. Liashenko, "Approach to determining the number of clusters in a data set", in Proc. International scientific and technical conf. Open Semantic Technologies for Intelligent Systems (OSTIS-2019), Minsk, 2019, pp. 151-154.

[14] Initialization: Where do you start?, [Online]. Available: http://www.salientiastuff.com/k-means-clustering-part-2.html.

[15] A. Pegat, "Sushchnost teorii nechetkikh mnozhestv", Nechetkoye modelirovaniye i upravleniye, 2015, (3th ed.), pp. 13-19, ISBN 3-7908-1385-0.

[16] F. Shevri, "Fuzzy logic", Schneider Electric, 2009, vol. 31, pp. 1-30, [Online]. Available: https://profsector.com/media/catalogs/566dd6af08f6c.pdf

[17] A. V. Lyashenko, "The approach to building fuzzy logical rules for big data", International scientific and technical conference: Modern challenges in telecommunications, 2020, [Online]. Available: http://conferenc.its.kpi.ua/proc/article/view/201702.

[18] A. Pegat, "Samoorganizuyushchiyesya i samonastraivayushchiyesya nechetkiye modeli", Nechetkoye modelirovaniye i upravleniye, 2015, (3th ed.), pp. 506-520, ISBN 3-7908-1385-0.

[19] V. V. Kurdecha, I. O. Ishchenko, and A. H. Zakharchuk, "Data processing method in distributed network Internet of Things", Young Scientist, 2017, vol. 10, no. 50, pp. 75-81.

[20] Microservice architecture, [Online]. Available: https://itnan.ru/post.php?c=1&p=320962.

[21] .NET Microservices: Architecture for Containerized .NET Applications, [Online]. Available: https://docs.microsoft.com/ru-ru/dotnet/architecture/microservices/.

Cross-Spectrum of Signals of Vibrations and their Application for Determination of the Technical Condition of Dynamic Equipment

Leonid Mylnikov and Nikita Efimov
Perm National Research Polytechnic University, 29 Komsomolsky avenue, Perm, Russia
leonid.mylnikov@pstu.ru, nfmv@mail.ru

Keywords: Equipment Condition, Vibration, Dynamic Equipment, Model, Ranking, Signal Analysis.

Abstract: The aim of the paper is to develop ranking techniques for dynamic equipment based on its technical conditions, the estimation of recovered resource value and the determination of critical points of time after which equipment operation has to be terminated. Accelerometer data, cross-spectrum for wave analysis and a TOPSIS-based method have been used to achieve the goal. The most significant result of the work is a method of estimating the technical condition of the equipment, which allows: 1) to perform the transition to condition-based equipment maintenance by predicting non-normative work time; 2) to plan preventive repairs; 3) to select performers for repairs and maintenance of equipment based on objective estimates of work quality. The importance of the results is as follows: 1) the application of multi-criteria ranking method allowed to make ranking according to the technical condition of the equipment units for which condition monitoring groups of sensors are used; 2) it is shown that equipment condition changing is non-linear and there are areas of accelerated degradation; when the latter ones are reached, an accelerated condition deterioration is encountered; 3) the application of the technique on the data simultaneously taken from four sensors has shown its ability to conduct a comprehensive estimation without reference to a specific type of failure in conditions when the data from individual accelerometers give different information about the failure due to the different distance from the problem area. The verification of the proposed theoretical results is carried out on the basis of operating time data before a bearing failure, as well as monitoring data on the operation of wind turbine gearboxes.

1 INTRODUCTION

The transition to predictive equipment maintenance and repair is one of the ways to improve the efficiency of production systems due to the fact that it makes it possible to plan equipment maintenance based on the upcoming load (deferred maintenance at high load and conducting predictive maintenance and repair at low load of the production system), as well as reduce the ecological load [1]. There are situations when monitoring and predicting the condition of equipment is a prerequisite for the production systems to function. For example, in China, for oil distillation stations located in the Gobi Desert [2], each maintenance event is very expensive, and because of the remoteness of the object, it is difficult to carry out timely repairs in case of an unexpected accident. The widespread use of wind turbines and their construction on the shelves of the seas and other remote locations from service centers make the task of reducing maintenance costs due to logistical features more acute than before [3].

To predict equipment failures and make an optimal maintenance schedule, statistical approaches were initially considered [4]. However, the use of statistics on equipment failures yields low accuracy, as shown by research in which we proposed to identify the equipment operating condition based on the analysis of the amplitude characteristics of the vibration signal using machine learning methods.

Furthermore, to process statistical data, methods of regression and data mining, machine learning, and neural networks are used, which do not give equally good results on all data and all types of equipment [5]. It is assumed that improving the accuracy is possible if one knows the distribution function of the occurrence of accidents [6] and the cause of failures, which requires large amounts of statistical information or data on the operating time of the investigated dynamic equipment units [7]. In practice, the use of excessive amounts of data leads to the

phenomenon of overfitting [8] and erroneous prediction of abnormal equipment operation. In addition, the collected datasets will be unbalanced due to the rarity of some phenomena, which makes the use of machine learning methods inefficient [9]. Currently, regression methods are used to predict the values of critical parameters (equipment characteristics) and machine learning methods to solve the classification problem (determining the current condition of the research object) [10]. In such cases, the parameters to be monitored are, as a rule, technological parameters secondary to the state of the equipment (current consumption values, resistance, friction, etc.).

A large group of methods that have received wide applications are wave methods, the implementation of which is possible after refining the equipment with vibration sensors, installed on the device body, the most important blocks or axes of rotating parts. The basis of all methods using such data is the suggestion that certain changes and/or configurations of recorded signals (wave characteristics) will tell about a particular condition or the process of approaching some desired or not condition [11]. Successful applications of wave process analysis can be found in various fields of science:

- To effectively predict strong earthquakes, seismic wave parameters are analyzed using methods such as spectral analysis based on Fast Fourier Transform and continuous wavelet transform. In seismic exploration, the elastic vibration field data are processed and further analyzed using amplitude control, migration, deconvolution, velocity analysis, and various types of filtering.
- Another example of wave-based diagnostics is the analysis of human sleep, particularly the detection of snoring activity, where sound pressure level and MFCC (Mel-frequency cepstral coefficients) are used to analyze the sound signal, based on which the classification models [12] are trained using the support vector method (SVM), deep learning and multi-core learning [13], also to detect apnea and asthma diagnosis, for which the spectrum analysis of breath noises is used.
- There are widely known cases of application of wave analysis generated by aircraft equipment for noise analysis, technical conditions, analysis of operating modes, and search for solutions to reduce acoustic cluttering of the space [14].
- Wave quality control of static products and monitoring of their condition during operation by wave reflections (it allows to find cracks, material irregularities, cavities) and analysis of vibrations of dynamic equipment are carried out.

Methods of this group can be divided into methods that evaluate changes in indirect parameters (e.g., the frequency parameters of the alternating current device), methods that use information from specially installed sensors (e.g., accelerometers) and methods that require action on the object to evaluate its condition (e.g., hammering and evaluation of wave propagation parameters) [15].

Two types of tasks are considered [16]: 1) predicting the equipment lifetime, 2) classification of the condition, and identification of faults.

The use of the wave description allows us to formulate a mathematical apparatus that will not impose requirements on the mechanisms and technical methods of obtaining a wave, i.e., to use any sensors capable of obtaining the necessary representation of the signal: sound, vibration, electromagnetic radiation, current, light, special characteristics of control systems.

In practice, the presence of a single sensor, as a rule, turns out to be insufficient, which leads to the statement of problems of multicriteria choice, among which only methods of ranking of decisions are applicable for the decision of the received problems as allow to receive stable decisions [17] and, do not apply convolution of criteria [18] (such approach brings certain assumptions in behavior and importance of estimations which lead to errors in decisions) and expert estimations [19] (the system should work in real-time).

Currently, a comprehensive equipment condition analysis taking into account a group of sensors is not carried out. As a result, only a certain type of fault is determined on complex devices [20] or the results obtained are not reproducible in conditions different from the initial conditions [21]. Thus, the purpose of this research is to develop a methodology for ranking equipment by its technical condition (based on information collected from a group of sensors), which will allow prioritizing the order of maintenance, making predictions about the operating time to failure, and estimating the value of restored resource after repairs and maintenance.

2 METHODOLOGY FOR INVESTIGATING THE CONDITION OF DYNAMIC EQUIPMENT

To analyze the vibration signal, it is necessary to identify the features that characterize the degradation

of the investigated equipment. For this purpose, the spectral analysis of time series was considered.

Each wave signal can be described by the corresponding power spectral density (PSD) of initial signals, characterizing the energy, which is carried by the considered frequency.

In general case PSD is defined as the Fourier transform of the covariance function[22]:

$$\varphi(\omega) = \sum_{k=-\infty}^{\infty} r(k) e^{-i\omega t},$$

where the covariance function $r(k) = E\{y(t)y^*(t-k)\}$, $y(t)$ – time series.

The most common assessment of PSD is the periodogram:

$$\varphi_p(\omega) = \frac{1}{N}\left|\sum_{t=1}^{N} y(t) e^{-i\omega t}\right|^2.$$

In practice, the frequency variable ω must be discretized and is usually considered $\omega = \frac{2\pi}{N}k,\ k = 0, \dots, N-1$

The use of periodograms has the disadvantage of large fluctuations relative to the true PSD. To solve this problem, smoothing is used using various window functions.

However, the use of window functions can have negative consequences [22]:

- there is the phenomenon of smearing in smoothing. This arises due to the fact that if two peaks of the function $\varphi(\omega)$ are located at a frequency of less than $1/N$, they will appear as one broader peak. Because of this, periodogram-based methods cannot distinguish details in the investigated spectrum that are separated by less than $1/N$ in cycles per sampling interval. Thus, $1/N$ is the limit of spectral resolution for the periodogram method;
- there is the effect of leakage, which is caused by the transfer of power from frequency bands with a large power concentration to bands with less or no power. This leads to a false estimation of the PSD, where the power will be contained at frequencies where it is absent.

Thus, for the most accurate PSD estimation, it is necessary to:

- Select the window length based on a compromise relationship between spectral resolution and statistical variance.
- Select the window type based on a compromise relationship between smearing and leakage effects.

It is possible to solve the above–described problem of selecting smoothing parameters only experimentally for the specifies investigated signal, which requires additional research and is beyond the scope of this work.

Therefore, the Daniell window was used in this work to obtain the PSD estimation because of the simplicity of the software implementation, since this method is based on the idea of reducing the variance by averaging the periodogram over small intervals.

Then the periodogram is calculated as follows [23]:

$$\varphi_D = \frac{2}{f_\Delta N^2}\sum_{m=0}^{M-1}\left|\sum_{n=0}^{N-1} y(t)\cdot e^{-2\pi\frac{n}{N}\left(\frac{f}{f_\Delta/M}+m-\frac{M-1}{2}\right)}\right|^2,$$

where N – number of time series elements; M – averaging factor; f_Δ - frequency resolution equal to M/T ; T – time series length.

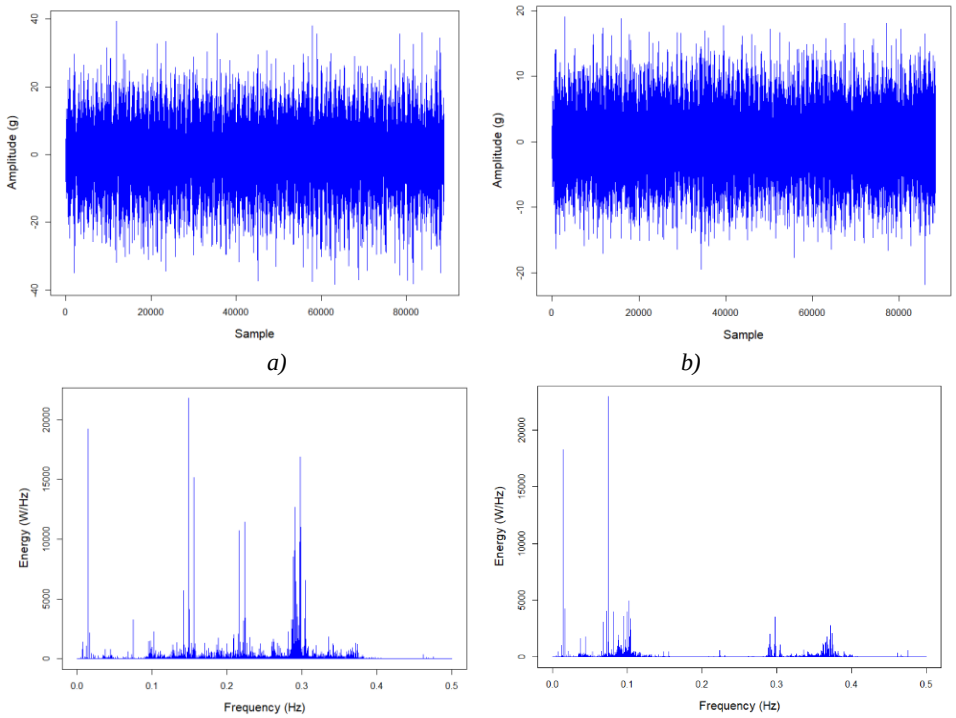
As a result, the PSD for frequencies in the range from 0 Hz to Nyquist frequency is calculated for the investigated time series. Nyquist frequency is the cutoff frequency equal to half of the sampling frequency, i.e., $\frac{f_\Delta N}{M}/2$.

Obtaining periodograms for the initial vibration signals allows identifying the significant frequencies that will characterize the condition of the equipment, as can be noticed the spectral density during normative operation and emergency operation is distributed differently (Figure 1), whereas the periodograms of two signals taken during normative operation are similar (Figure 2).

As can be seen from the above figures, it can be assumed that the degree of equipment deterioration may be identified by comparing the PSD distribution of the recorded signal with the benchmark.

Definition. *The characteristic taken at the first run of the new equipment can be used as a benchmark.*

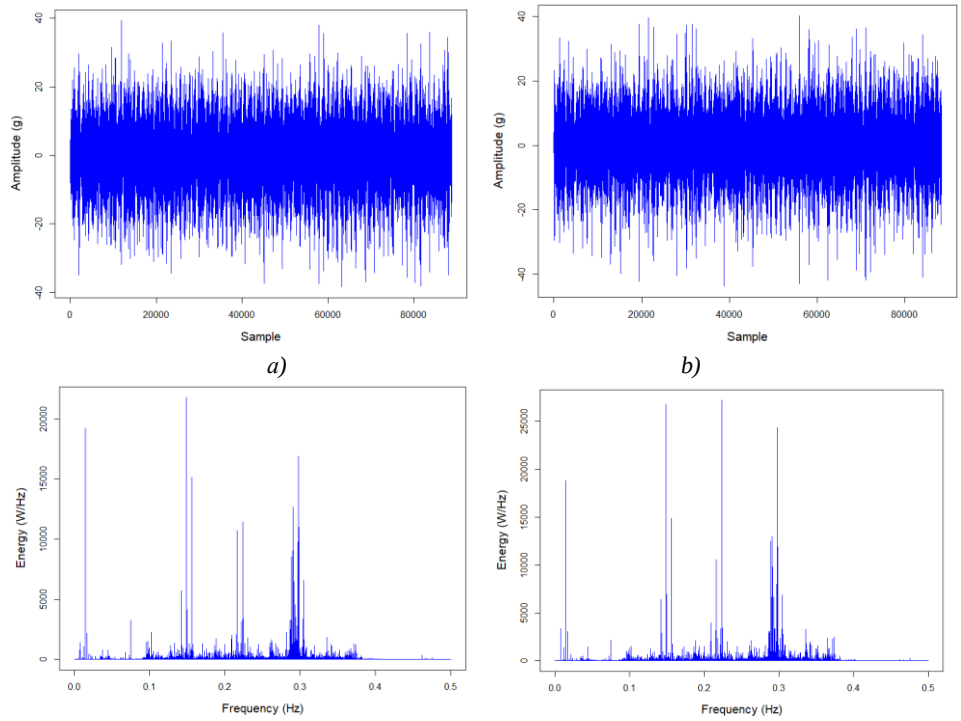
Thus, the condition of each equipment unit will be estimated from the similarity degree of vibration signal PSD estimation with the benchmark one, thereby allowing to track the deterioration dynamics of a particular equipment unit, as well as to compare the deterioration degrees of several equipment units among themselves. This assumption allows us to refuse from singling out the frequencies contributed by each component of the investigated equipment unit [24] and tracking the changes by the set of these frequencies. Such an approach, on the one hand, makes it unnecessary to decompose the signal into separate components contributed by each element of the dynamic system but excludes the possibility of precise identification of the failure cause, which will require stopping and carrying out maintenance of the equipment.



a) *b)*

c) *d)*

Figure 1: a) Initial vibration signal during normative wind turbine gearbox operation; b) initial vibration signal during emergency wind turbine gearbox operation; c) periodogram of the signal a); d) periodogram of the signal b).



a) *b)*

c) *d)*

Figure 2: Periodograms c), d) of vibration signals a) and b) respectively, taken during normative operation of wind turbine gearboxes.

To estimate the similarity of PSD of two time series a cross-spectrum is used [25]:

$$\varphi_{xy}(\omega) = \frac{1}{\pi}\sum_{k=-\infty}^{\infty} r_{xy}(k)e^{-i\omega k},$$

where $0 < \omega < \pi$,

where cross-covariance function $r_{xy}(k) = E\{x(t)y^*(t-k)\}$.

In contrast to the PSD estimation of a single time series by periodogram $\varphi_p(\omega)$, the function $\varphi_{xy}(\omega)$ is complex:

$$\varphi_{xy}(\omega) = c(\omega) - iq(\omega),$$

where the function $c(\omega) = \frac{1}{\pi}\sum_{k=-\infty}^{\infty} r_{xy}(k)\cos(\omega k)$ – co-spectrum, and the function $q(\omega) = \frac{1}{\pi}\sum_{k=-\infty}^{\infty} r_{xy}(k)\sin(\omega k)$ – quadrature spectrum.

As a result of cross-spectrum calculation for each frequency of the investigated vibration signal, the similarity with the benchmark signal is evaluated.

Since the power spectral density is usually measured in decibels, it is necessary to convert $f_{xy}(\omega)$ to $\lg\left(f_{xy}(\omega)\right)$ before performing any operation on the obtained cross-spectrum.

As mentioned earlier, the cross-spectrum allows estimating the similarity at each frequency between two time series, while the criterion for ranking should be one number, which would uniquely characterize the state of the investigated equipment. Since the closer, the current state to the benchmark condition, the greater the cross-spectrum values, and respectively the average value across all frequencies at normative operation will be greater than at faulty operation, and as the equipment wears, the average value of the cross-spectrum will decrease. Accordingly, the average value of the cross-spectrum will be used as a criterion.

Due to the fact that $\varphi_{xy}(\omega)$ is a complex function, the criterion is also a complex value, but, as seen in Figure 3, information about the magnitude of the imaginary and real parts of the criterion (averaged co-spectrum and averaged quadrature spectrum) is not so important for the condition evaluation, but the criterion closeness to the coordinate origin on the complex plane is important, so the complex value of the criterion can be replaced by its modulus.

The dynamic equipment is a complex device consisting of many elements, so the signal is taken not from one vibration sensor, but from several, located on different parts of the equipment. This leads to the task of estimating each piece of equipment condition, based on multiple signals.

The task of evaluating dynamic equipment conditions is reduced to comparing the current states of different pieces of equipment with each other or tracking the deterioration dynamics, in other words, comparing the current state with the previous measurements taken at certain intervals. Then for the task of condition estimation according to the data from several sensors, one of the solutions is the use of outranking methods for multi-criteria ranking of the equipment condition.

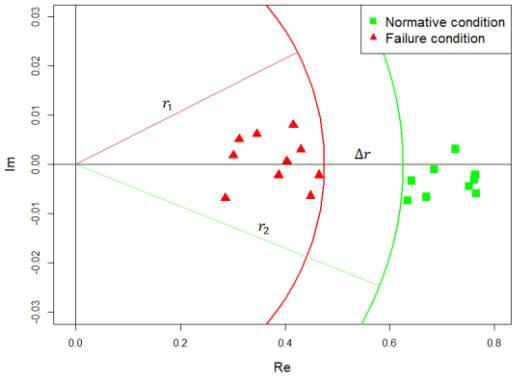


Figure 3: Divergence of wind turbine gearbox condition estimations by the vibration signal in the complex plane.

Among outranking methods, the TOPSIS [26] method is based on the identification of positive ideal (PIS) and negative ideal (NIS) solutions which, in the presence of data on the operating time of failure, will correspond to the conditions of new equipment and out of service (shutdown). Thus, all units of equipment or characteristics of one device removed during the operation will be systematized in relation to these states if we assume that values of criteria monotonically increase or decrease (the technical condition cannot spontaneously improve during operation).

The work of the method can be described in seven steps.

Step 1: Create a scoring matrix consisting of m measurements for n sensors, with scoring values in intersections $P_{ij}; i = 1, \dots, m; j = 1, \dots, n$.

Step 2: Normalize the matrix of P_{ij} values and obtain a matrix R consisting of elements r_{ij} calculated by the formula: $r_{ij} = \frac{P_{ij}}{\sqrt{\sum_{k=1}^{m} x_{kj}^2}}, \forall i, j.$

Step 3: Calculate the weighted normalized decision matrix $t_{ij} = r_{ij} \cdot w_j, \forall i, j,$ where $w_j = W_j \Big/ \sum_{k=1}^{n} W_k, \forall j,$ where W_j – is the initial weight assigned to the j- th criterion (indicator). Obtain the values of the weights satisfying the following equality $\sum_{i=1}^{n} w_i = 1$.

Step 4: Determine the worst (A^-) and the best (A^+) alternatives:
$A^- = \{(\max(t_{ij}) \mid j \in J^-), (\min(t_{ij}) \mid j \in J^+)\} \equiv t_j^+,$
$A^+ = \{(\min(t_{ij}) \mid j \in J^-), (\max(t_{ij}) \mid j \in J^+)\} \equiv t_j^-,$
$\forall j,$
where J^- is the set of indicators an increase in the value of which brings a negative result, J^+ is a set of indicators, an increase in the value of which has a positive result.

Step 5: Calculate the Euclidean distance for the i-the alternative with the worst solution:

$$A^- \ (d_i^- = \sqrt{\sum_{j=1}^{n}(t_{ij} - t_j^-)^2}, \forall i)$$

and with the best solution:

$$A^+ \ (d_i^+ = \sqrt{\sum_{j=1}^{n}(t_{ij} - t_j^+)^2}, \forall i),$$

where d_i^- and d_i^+ are the Euclidean distances to the worst and best solutions.

Step 6: Calculate the closeness to the best or worst state: $s_i^- = {d_i^-}/{(d_i^- + d_i^+)}$ или ${d_i^+}/{(d_i^- + d_i^+)}$.

Step 7: Ranking the alternatives by the values of s_i^- or s_i^+, $\forall i$.

For successful application of such approach, it is necessary: to choose criteria for ranking; to choose the most appropriate ranking method; to determine the internal parameters of the chosen ranking method (criterion weights, maximization/minimization of each specific criterion, etc.). The general scheme of the algorithm for a set of sensors can be represented in Figure 4.

[1]https://drive.google.com/drive/folders/1_ycmG46PARiykt82ShfnFfyQsaXv3_VK

3 EXPERIMENTAL RESEARCH ON THE APPLICABILITY OF THE PROPOSED METHODOLOGY

Due to the presence of moving elements in the dynamic equipment design, constant state degradation of such elements is unavoidable. One of the most common elements in dynamic equipment is bearings, which ensure the rotation or rolling of the connected structural elements with the least resistance, so bearing wear will noticeably affect the operation of the equipment as a whole.

To verify the proposed approach, we will use the data on the failure time of the bearings, which can be downloaded from the link[1]. The availability of the data on the operating time between failures allows estimating the distance to the breakdown condition

The vibration signal is received from two sensors; therefore, the ranking will be performed according to two criteria. In this case, the task is to track how far the technical condition of the equipment is from the emergency condition, so we will take the emergency condition as an ideal-positive solution, respectively, when reducing the criteria, the condition will approach the emergency condition, that is, the criteria must be minimized.

So, at the dynamic condition change, it is impossible to evaluate which of the sensors most clearly shows degradation, the weights will be equal.

According to the obtained results (Figure 6), it can be concluded that the condition is gradually moving from normative to emergency and the rating score is also increasing, as expected.

.

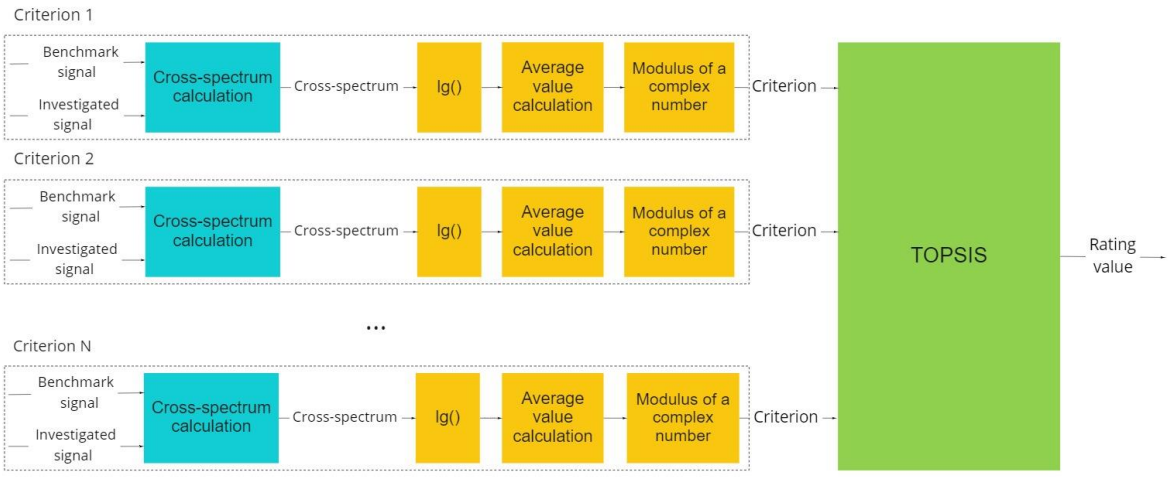
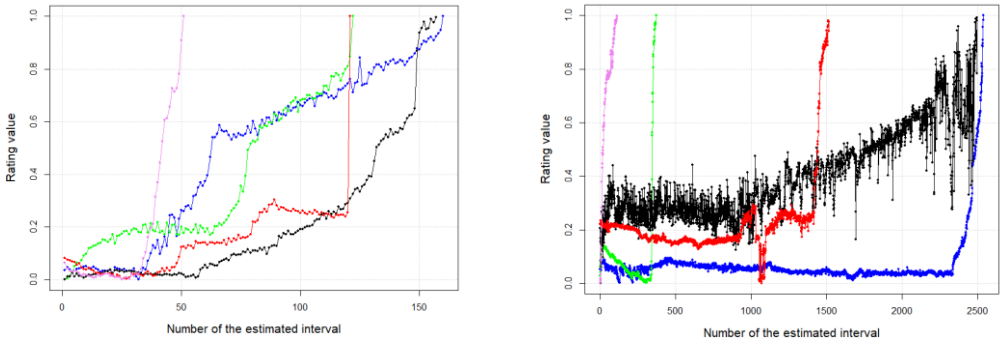


Figure. 4: Algorithm for estimating the condition of the equipment with multiple sensors installed.

The problem. There is no monotony of change and there are fluctuations in the process of tracking the deterioration, and as the bearing rotational speed increases, the fluctuations increase

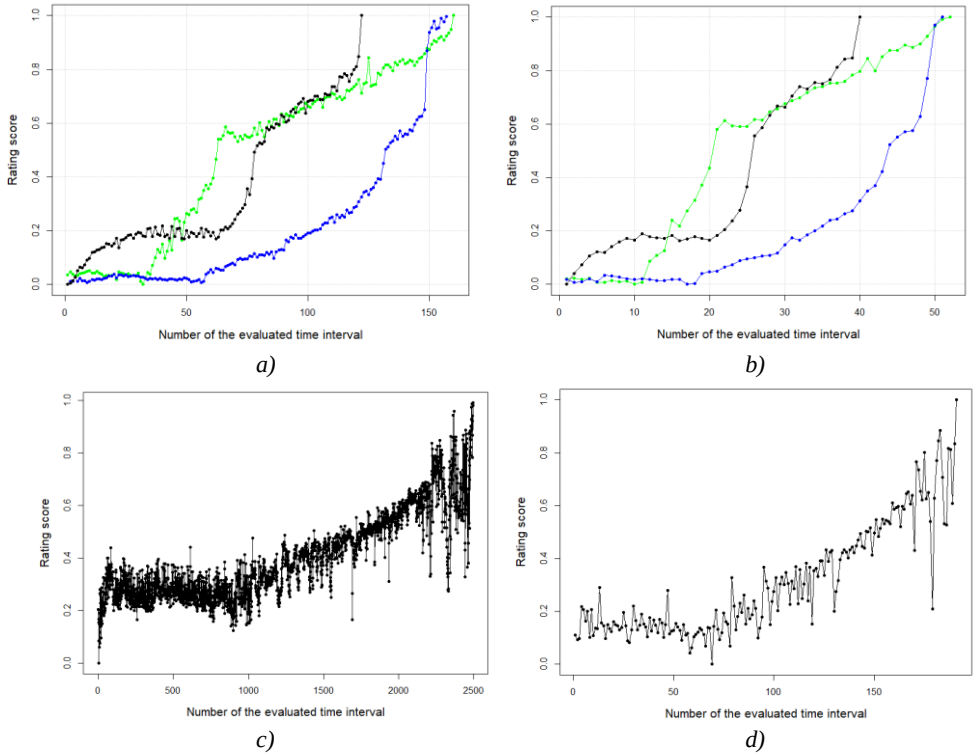
This problem can be caused by the fact, that the experiments were carried out in conditions of accelerated degradation far from the normative ones, assuming sharp random bursts of vibration signal fluctuations, which worsen the method performance. as well as when solving the problem of tracking wear dynamics, an important factor is the choice of the estimation time interval size, the method of smoothing, and the size of the smoothing window.

By increasing the size of the evaluation interval, we can notice a decrease in fluctuations in the dynamics of change in the condition of the investigated equipment (Figure 7), but another problem arises, related to the fact that with too large an interval the probability of untimely detection of critical equipment wear increases.



a) *b)*

Figure 6: Result of 5 bearing deterioration dynamics ranking a) first experiment at 35 Hz rotation speed and 12 kN load, b) third experiment at 40 Hz rotation speed and 10 kN load.



a) *b)*

c) *d)*

Figure 7: Influence of the evaluation time interval on the quality of the ranking. With an initial interval of 1 minute for experiments a), there were small fluctuations, which were noticeably smoothed out in experiment b) when the interval size was changed from 1 minute to 3 minutes. In experiment c) the evaluation was made with an interval of 1 minute and in this case very strong fluctuations and sharp jumps can be observed, when the interval is changed to 13 minutes in experiment d) it is possible to notice a decrease in the number of fluctuations and emissions, but they are still present.

Experiments show that depending on the period of the data taken for analysis, there is an increase or decrease in the magnitude of fluctuations in the analysis data, which indicates the need to choose for each equipment unit the value of time for data acquisition, as well as the periodicity of these operations.

A more complex example is the gearbox in a wind turbine structure. As soon as a failure occurs in the wind turbine gearbox, the efficiency of power generation inevitably decreases, and eventually, unplanned downtime happens [27]. Thus, monitoring the condition of dynamic equipment is necessary not only to prevent emergency breakdown and resulting downtime but also to avoid losses in operational efficiency.

There is open-source data on the performance of generators and gearboxes of wind turbines (data available on the portal of the U.S. National Renewable Energy Laboratory[2], but the results obtained in the currently known studies carried out on them show their effectiveness only on this data [28], which confirms the relevance of the goal.

Since there are four vibration sensors on the investigated equipment, the TOPSIS ranking will be estimated according to four criteria.

As mentioned earlier, and as can be seen from previous calculation results, the criteria decrease as the equipment deteriorates. Since the ranking on this dataset involves comparing several wind turbine gearboxes relative to each other, the larger the criterion, the larger the score should be since it is closer to the benchmark condition, therefore each criterion within the TOPSIS method should be maximized.

Not all sensors can track the wear of equipment, so it is necessary to calculate the matrix of weights W_j so that the criteria from the sensors that show the most obvious wear of equipment have a greater impact on the condition evaluation.

- Calculate the average values for each criterion for each state (normative/emergency):

$$P_j^+ = \frac{\sum_{i \in S^+} P_{ij}}{N^+}, \forall j,$$

where P_j^+ – the average value of j-th criterion during normative operation, S^+ – set of gearboxes operating in normal mode, N^+ – the number of gearboxes operating in normal mode;

$$P_j^- = \frac{\sum_{i \in S^-} P_{ij}}{N^-}, \forall j,$$

where P_j^- – the average value of j-th criterion during emergency operation, S^- – set of gearboxes operating in emergency mode, N^- – the number of gearboxes operating in emergency mode;

- calculate the weights $W_j = \frac{\left|P_j^+ - P_j^-\right|}{max\left\{\left|P_j^+ - P_j^-\right|\right\}_{\forall j}}$.

Thus, the criterion that has on average a large difference between the values at different modes of operation will have a greater weight.

Based on the results of the ranking (Figure 8) it can be seen that the equipment units, the state of which was classified as normative are at the top of the rating with some gap, which confirms the hypothesis about the divergence of equipment estimates, which are in normal and emergency modes.

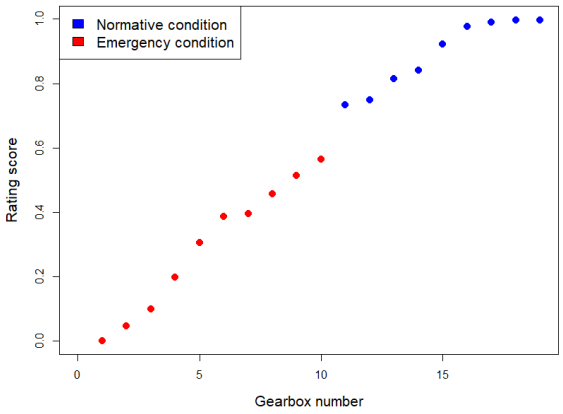


Figure 8: Ranking of wind turbine gearboxes by deterioration using TOPSIS method.

4 DISCUSSION

Thus, we obtain several options for using the approach described in the article, related to the ranking of the same type of equipment by its technical condition to determine the maintenance priorities.

Another way to use the results obtained can be to estimate the change in the condition of the equipment after maintenance or repair. In this way, it is possible to estimate the value of restored lifetime and on the basis of such statistics to solve the problem of selecting a service provider.

The performance quality of the algorithm depends on the accuracy of the ranking, which in its turn depends on such parameters of the algorithm as the time period of equipment vibration measurement, the number of used signals/sensors, the type of the time window function used to build cross-spectrum.

[2]https://openei.org/datasets/dataset/gearbox-fault-diagnosis-data/resource/affa53da-cae6-42f2-b898-ad018ff91641

The proposed algorithm provides the necessity of selecting the internal parameters such as the type of smoothing window, the size of the smoothing window, the size of the estimation time interval, which allows for each case to choose the solutions best suited by the morphological synthesis method (Table 1) and [29].

In the literature, tasks related to failure time prediction are most often considered. The application of the described approach made it possible to reduce the amount of information that contains information about the equipment condition. Thus, the dimensionality of the problem is reduced to one degree of freedom (DOF), which simplifies the task of selecting the estimation/descriptive data function or selecting the ML method for predicting failures [24], and also makes it possible to estimate the condition by an expert method.

Experiments show that when using S-curve regression methods and machine learning methods, the algorithms, correctly, predict trends [30]. In the case of S-curves [31], the inflection point shows the transition from normative operation to non-normative operation (Figure 9), and when using ML algorithms, we predict the time of complete failure and shutdown of equipment unit (Table 2) [32].

Table 1: Example of a morphological table.

	Alternatives				
Parameter	1	2	3	4	**…**
A) Smoothing window type for cross-spectrum calculation	Daniell's Window	Blackman-Harris Window	Hann and Hamming windows	Kaiser-Bessel Window	…
B) Smoothing window size when calculating the cross-spectrum	2	4	6	8	…
C) Size of the time interval for the evaluation of the equipment condition	1 min	2 min	3 min	4 min	…

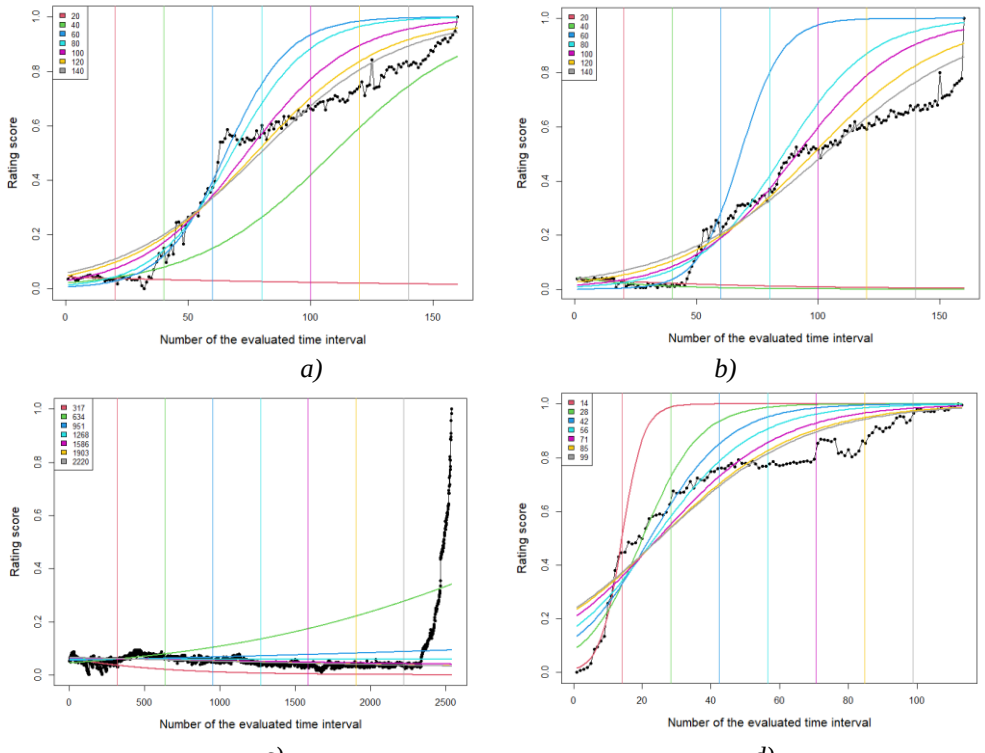


Figure 9: Results of S-curves for a) experiment 1_2, b) experiment 2_2, c) experiment 3_1, d) experiment 3_5.

Table 2: Comparison of real time to failure in estimation intervals with predicted one using models trained by ElasticNet, Ridge, and Lasso methods for experiment 2_2.

Real number of evaluation intervals to failure	Predicted value		
	ElasticNet	Ridge	Lasso
140	90.278476	89.387534	88.718160
130	91.299122	90.818057	90.454385
120	91.353579	90.635841	90.125312
110	84.615569	86.434416	83.321456
100	73.156105	73.173403	72.412751
90	68.560701	68.295393	67.208324
80	67.148999	65.342441	66.392584
70	51.773237	53.533065	53.360640
60	53.740850	50.701370	51.854554
50	47.007266	47.528948	50.048620
40	43.904362	42.878614	45.834927
30	39.439278	40.118334	42.781805
20	36.810094	37.415678	41.023638
10	34.584381	35.482358	39.560329

Even though the conducted experiments show that it is possible to refuse from using all the data collected from the sensors, the accuracy of the algorithms is not high, which indicates the need for additional adjustment associated with the choice of algorithm parameters (see Table 1) and additional research associated with the choice of data description method and/or machine learning method that gives the best results in each specific case [5].

5 CONCLUSIONS

The article presented the research related to the identification of the equipment condition based on vibration signals through vibration diagnostics signal analysis, namely: 1) developed a model of equipment condition identification, distinguished by the use of periodograms of signals coming from vibration sensors; 2) developed a method of equipment condition estimation, distinguished by the use of multi-parameter ranking of equipment condition.

REFERENCES

[1] T. J. Mi S. , Feng Y., Zheng H., Li Z., Gao Y., "Integrated Intelligent Green Scheduling of Predictive," IEEE Access, vol. 8, pp. 45797-45812, 2020, doi: 10.1109/ACCESS.2020. 2977667.

[2] Q. Wang and J. Gao, "Research and application of risk and condition based maintenance task optimization technology in an oil transfer station," J. Loss Prev. Process Ind., vol. 25, no. 6, pp. 1018-1027, Nov. 2012, doi: 10.1016/j.jlp.2012.06.002.

[3] W. S. J. Tautz–Weinert J., "Using SCADA data for wind turbine condition monitoring – a review," IET Renew. Power Gener., vol. 11, no. 4, pp. 382-394, 2017, doi: 10.1049/iet-rpg.2016.0248.

[4] H. Li et al., "Improving rail network velocity: A machine learning approach to predictive maintenance," Transp. Res. Part C Emerg. Technol., vol. 45, pp. 17-26, Aug. 2014, doi: 10.1016/j.trc.2014.04.013.

[5] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," IEEE Trans. Evol. Comput., vol. 1, no. 1, pp. 67-82, Apr. 1997, doi: 10.1109/4235.585893.

[6] M. Sadiakhmatov, "Production planning model in the conditions of changing demand with a stochastic component.," HS Anhalt, 2018.

[7] G. L. Wang B., Lei Y., Yan T., Li N., "Recurrent convolutional neural network: A new framework for remaining useful life prediction of machinery," Neurocomputing, vol. 379, pp. 117-129, 2020, doi: 10.1016/j.neucom.2019.10.064.

[8] L. D. J., "Stupid Data Miner Tricks: Overfitting the S&P 500," J. Invest., vol. 16, no. 1, pp. 15-22, 2007, doi: 10.3905/joi.2007.681820.

[9] B. A. Santos P. , Maudes J., "Identifying maximum imbalance in datasets for fault diagnosis of gearboxes," J. Intell. Manuf., vol. 29, no. 2, pp. 333-351, 2018, doi: 10.1007/s10845-015-1110-0.

[10] A. Paprotny and M. Thess, Realtime data mining: self-learning techniques for recommendation engines. Birke, 2013.

[11] B. Wang, Y. Lei, N. Li, and N. Li, "A Hybrid Prognostics Approach for Estimating Remaining Useful Life of Rolling Element Bearings," IEEE Trans. Reliab., vol. 69, no. 1, pp. 401-412, Mar. 2020, doi: 10.1109/TR.2018.2882682.

[12] L. Mylnikov, B. Krause, M. Kütz, K. Bade, and I. Shmidt, Intelligent data analysis in the management of production systems (approaches and methods). Aachen: Shaker Verlag GmbH, 2018.

[13] K. Nishijima, S. Uenohara, and K. Furuya, "Evaluating Classification Methods in Snore Activity Detection," in Advances in Intelligent Systems and Computing, 2019, pp. 921-926.

[14] K. V. F. Chernyshev S.A., "Vortex ring eigen-oscillations as a source of sound," J. Fluid Mech., vol. 341, pp. 19–57, 1997.

[15] H. Zuo, K. Bi, and H. Hao, "A state-of-the-art review on the vibration mitigation of wind turbines," Renew. Sustain. Energy Rev., vol. 121, p. 109710, Apr. 2020, doi: 10.1016/j.rser.2020. 109710.

[16] N. Gebraeel, M. Lawley, R. Liu, and V. Parmeshwaran, "Residual Life Predictions From Vibration-Based Degradation Signals: A Neural Network Approach," IEEE Trans. Ind. Electron., vol. 51, no. 3, pp. 694-700, Jun. 2004, doi: 10.1109/TIE.2004.824875.

[17] M. Cinelli, S. R. Coles, and K. Kirwan, "Analysis of the potentials of multi criteria decision analysis methods to conduct sustainability assessment," Ecol. Indic., vol. 46, pp. 138-148, Nov. 2014,

doi: 10.1016/j.ecolind.2014.06.011.

[18] J. S. Dyer, "Maut — Multiattribute Utility Theory," in Multiple Criteria Decision Analysis: State of the Art Surveys, New York: Springer-Verlag, pp. 265-292.

[19] N. F. Matsatsinis and A. P. Samaras, "Brand choice model selection based on consumers' multicriteria preferences and experts' knowledge," Comput. Oper. Res., vol. 27, no. 7–8, pp. 689–707, Jun. 2000, doi: 10.1016/S0305-0548(99)00114-8.

[20] J. Xu, X. Ding, Y. Gong, N. Wu, and H. Yan, "Rotor imbalance detection and quantification in wind turbines via vibration analysis," Wind Eng., p. 0309524X2199984, Mar. 2021, doi: 10.1177/0309524X21999841.

[21] Z. Ma, M. Zhao, B. Li, and H. Fan, "A novel blind deconvolution based on sparse subspace recoding for condition monitoring of wind turbine gearbox," Renew. Energy, vol. 170, pp. 141-162, Jun. 2021, doi: 10.1016/j.renene.2020.12.136.

[22] P. Stoica and R. Moses, Spectral Analysis of Signals, vol. 447. Upper Saddle River, New Jersey: Prentice Hall, Inc., 2005.

[23] A. Labuda, "Daniell method for power spectral density estimation in atomic force microscopy," Rev. Sci. Instrum., vol. 87, no. 3, 2016, doi: 10.1063/1.4943292.

[24] T. Wang, Q. Han, F. Chu, and Z. Feng, "Vibration based condition monitoring and fault diagnosis of wind turbine planetary gearbox: A review," Mech. Syst. Signal Process., vol. 126, pp. 662–685, Jul. 2019, doi: 10.1016/j.ymssp.2019.02.051.

[25] O. Marchal, NOTES OF TIME SERIES ANALYSIS, vol. 27. Department of Geology & Geophysics, Woods Hole Oceanographic Institution, 2015.

[26] L. A. Mylnikov. Upravleniye proyektami i sistemami v usloviyakh tsifrovoy ekonomiki. Perm: Izd-vo Perm. nats. issled. politekhn. un-ta. 2021.

[27] Z. Ma, W. Teng, Y. Liu, D. Wang, and A. Kusiak, "Application of the multi-scale enveloping spectrogram to detect weak faults in a wind turbine gearbox," IET Renew. Power Gener., vol. 11, no. 5, 2017, doi: 10.1049/iet-rpg.2016.0722.

[28] N. Yang, S. Liu, J. Liu, and C. Li, "Assessment of Equipment Operation State with Improved Random Forest," Int. J. Rotating Mach., vol. 2021, pp. 1-10, Mar. 2021, doi: 10.1155/2021/8813443.

[29] Zwicky F., Discovery Invention, Research Through the Morphological Approach. McMillan, 1969.

[30] N. Efimov, "Estimation model of the technical condition of dynamic equipment changes based on vibration data," Anhalt University of Applied Sciences, 2021.

[31] A. V. Seledkova, L. A. Mylnikov, and K. Bernd, "Forecasting characteristics of time series to support managerial decision making process in production-And-economic systems," 2017, doi: 10.1109/SCM.2017.7970744.

[32] L. A. Mylnikov. Statisticheskiye metody intellektualnogo analiza dannykh. SPb.: BKhV-Peterburg, 2021.

The Computer Program for the Treatment of Big Data in the Field of Literature Science

Liliia Bodnar[1], Kateryna Shulakova[2] and Olena Tyurikova[3]
[1]*Department of Innovative Technologies and Methods of Teaching Natural Sciences, South Ukrainian National Pedagogical University, 26 Staroportofrankyvska Str., Odessa, Ukraine*
[2]*Department of Computer Engineering and Information Systems, State University of Intellectual Technologies and Telecommunications, 1 Kovalska Str., Odessa, Ukraine*
[3]*Department of Architectural Environment Design, Odessa State Academy of Civil Engineering and Architecture, 4 Didrihsona Str., Odessa, Ukraine*
bodnarl79@pdpu.edu.ua, k.shulakova@onat.edu.ua, tulenanik@mail.ru

Keywords: Program for Big Date Treatment, Zipf's Laws, Evolution of Languages.

Abstract: The problem of processing large databases is important for solving many pressing problems of science and technology. In this paper, we have developed a computer program (Conan 3.0) for processing large text arrays. However, other applications are possible. We have applied the developed program for the analysis of large texts on the basis of Zipf's laws. The task, which was solved in this work, is connected with the laws of the evolution of languages; in particular, correlations in the development of different Slavic languages were traced. It was assumed that an important characteristic of the language is the Zipf's constant. As a result of calculating the changes in the short-circuit over the 18th, 19th and 20th centuries for the Ukrainian, Russian and Polish languages, no significant changes in the short-circuit were revealed. Small fluctuations in the short-circuit for these languages do not correlate.

1 INTRODUCTION

A wide range of research [1, 2] has shown that tools from information theory (e.g. information content/surprises, entropy) are useful tools in addressing questions of linguistic interest. These range from predicting the targets and outcomes of phonological and syntactic processes, to explaining and evaluating models of linguistic data.

Zipf's law is a fundamental paradigm in the statistics of written and spoken natural language. Zipf's law usually refers to the fact $P(s) = Pr\{S > s\}$ that the value S of some stochastic variable, usually a size or frequency, is greater than s, decays with the growth of s as $P(s) \sim s^{-1}$. This in turn means that the probability density functions P (s) exhibits the power law dependence in (1):

$$P(s) \sim 1/s1+m \text{ with } m=1. \quad (1)$$

Zipf's law is strictly valid if randomly if a balanced condition is fulfilled: the sum of all the mechanisms responsible for the growth and decline of firms must vanish on average in a precise sense. Any departure from this requirement yields a departure of the tail index from its canonical value $m=1$. This result can allow one to understand why different tail indexes are reported in the literature for different countries around the world [3].

According to Zipf's law, in a list of word forms ordered by the frequency of occurrence, the frequency of the r th word form obeys a power function of r (the value r is called the rank of the word form). It should be noted that further surveys [4] showed that Zipf's law is roughly realized only for the most frequent words.

In Ref. [5] it was shown that evolution of languages connected with a biological capacity shared by all humans and distinguished by the central feature of discrete infinity – the capacity for unbounded composition of various linguistic objects into complex structures. These structures are generated by a recursive procedure that mediates the mapping between speech- or sign-based forms and meanings, including semantics of words and sentences and how they are situated and interpreted in discourse. This approach distinguishes the biological capacity for language from its many

possible functions, such as communication or internal thought.

The study of language evolution is performed using approaches of "Big Data". Different models of language evolution are expressed in multiple empirical domains. Databases for linguistic structure are available in the Internet [6].

In Ref. [7] it have explored the effectiveness of authorship attribution on works of literature. A certain authors have a highly recognizable style. It was considered usage statistics for the commonly used style markers for two authors. Each number is the number of function occurrences that is the particular function word. It was realize like our ways with usage Natural Language Toolkit (NTLK) library.

Margins, column widths, line spacing, and type styles are built-in. Some components, such as multi-levelled equations, graphics, and tables are not prescribed, although the various table text styles are provided. The formatter will need to create these components, incorporating the applicable criteria that follow.

2 DESCRIPTION OF COMPUTER PROGRAM

The developed computer program (Conan 3.0) was created for analyzing large texts. There is also a possibility to use the program for assessing the quality of literary translations [9].

The program Conan 3.0 is based on the earlier version Conan 2.0, which was written in the C # language using the functionality of MSSQL database like Full-text Search for ordering and processing data which was getting after parsing of a text document and importing them into MSSQL Database. Such a way of processing data was pretty complicated and had a lot of restrictions and as a result of pretty low performance and quality of calculation [8] [9]. This led to the development of a new version of the program, where we abandoned the use of the database functionality in favor of dynamic analysis based on algorithms provided by the NTLK library.

Creating a new version of the application we were following such goals as:

- Make the application more native to the internet surrounding.
- Make the application usage easier for users.
- Increase performance and quality of calculation due to using Python and already well-developed python BigData libraries.

The program is based on the simple text data processing algorithm (Figure 1).

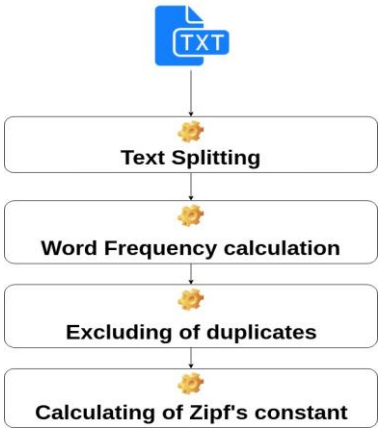


Figure 1: Algorithm for calculation of Zipf's constant.

Our previous programs did not take into account the important features of Zipf's laws. For example, the same word was accounted many times with different suffixes.

In the program Conan 3.0 the probabilistic search was introduced which made the calculations more accurate.

Based on the modern technologies for processing large amounts of data and the Natural Language Toolkit (NTLK) library [10], it was possible to rewrite the program code while using simple algorithms.

The data of this library were used also in the work [7].

NTLK includes extensive software, and documentation that we downloaded from source[1]. We used Natural Language Processing to provide any kind of computer manipulation with natural language. As to programming environment, the choice was done between two scripting languages, Python and JavaScript. At present much of server applications are written in JavaScript and Python languages [12]. Using each of the platforms we easily develop and maintain web applications of any complexity. Event-oriented architecture Node.js, which allows us to handle large streams of data simultaneously, and Python, in turn, is perfect for processing large amounts of data.

[1]http://www.nltk.org/

In this work the results were obtained using some approaches in the field of Big Data [9]. Thus, the chosen approach provided a platform for writing the program as a full-fledged web application for online use.

In the course of the chosen solution, it also became clear that there is no function that calculates the Zipf's constant in its pure form in this library. In the previous versions of the program, the main difficulty of data processing was not the calculation of the Zipf's constant, but the preparation and processing of the text under study. Therefore, the Natural Language Toolkit came up for the intermediate values needed for the Zipf's constant, such as calculating the frequency of occurrences of words. As a result, the algorithm for determining the Zipf's constant has not changed except for using the Natural Language Toolkit library instead of using the Full Text Search technology previously used in the full text search. This greatly accelerated the calculation of the Zipf's constant.

3 ANALYSIS OF SLAVIC LANGUAGES EVOLUTION FROM THE 18TH TO THE 20TH CENTURIES

The study was conducted in order to trace the evolution of the languages of the East Slavic and West Slavic groups: Ukrainian, Russian and Polish.

The representation in terms of the frequency distribution *f (n)* was successfully used to demonstrate the stability of the Zipf's constants for various literary works from the 18th to the 20th centuries with different sizes of texts.

Such works as “Litopis Samovidtsya” (18th century), “Compositions of Shevchenko” (19th century), “Compositions of Dovzhenko” (20th century) and others were investigated in Ukrainian.

Processing the text data for the Ukrainian language, the following results were obtained, shown in Figure 2.

The obtained data indicate that the Zipf's constant for Ukrainian works is within:

- 18th century from 0.049 to 0.072
- 19th century from 0.056 to 0.087
- 20th century from 0.05 to 0.061

The largest variation occurred in the 19th century at 0.031 units. Taking into account the fact that the Zipf's constant is a constant for each language, it can be assumed that this language was intensively modified in a given period of time, which led to a change in the identity of the language. Taking the fact that the value in the 20th century almost returned to the framework of the 18th century and the spread became much smaller, we can assume that destructive changes in the language did not occur and its identity was preserved in one way or another.

Studies of Russian works have shown the results, which are presented in Figure 3. Such works as "Works of Lomonomov" (18th century), "Works of Tolstoy" (19th century), "Works of Bunin" (20th century) and others were investigated in Russian.

The data obtained indicate that the Zipf's constant for Russian works is within:

- 18th century from 0.037 to 0.064
- 19th century from 0.04 to 0.065
- 20th century from 0.021 to 0.05

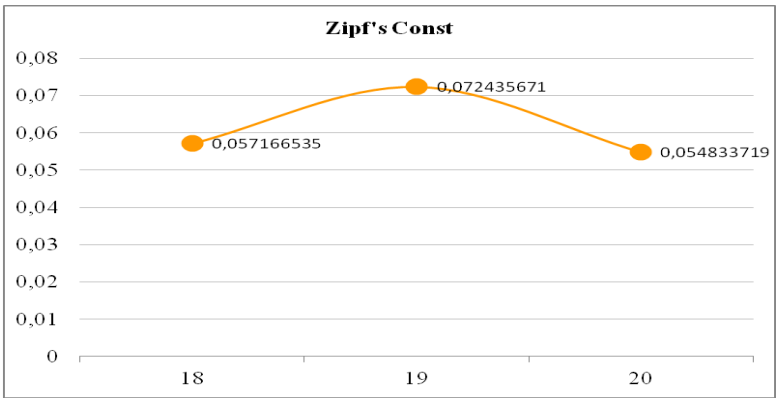


Figure 2: Counting Zipf's constant for Ukrainian.

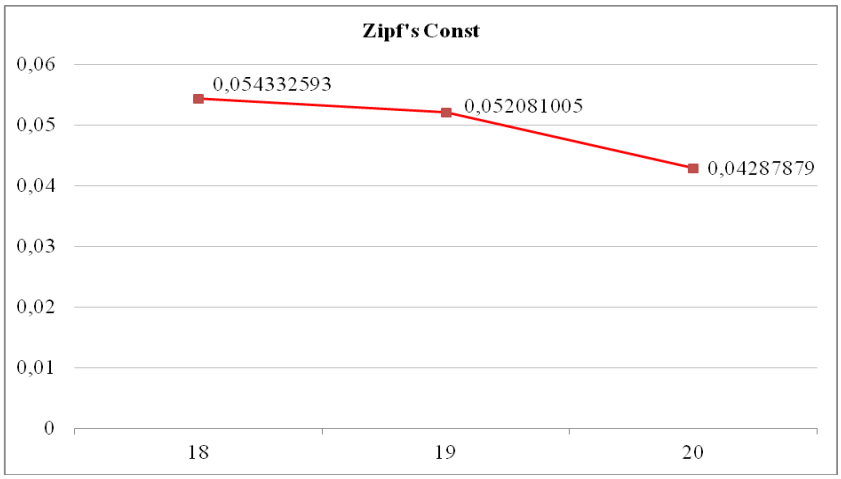


Figure 3: Counting Zipf's constant for Russian.

The results suggest that the language underwent an intense change in the 18th century. Perhaps it was the period of the formation of the language and its foundations. In the 19th century, language changes are minimal, i.e. the language was as stable as possible during this period of time. In the 20th century, the figure shows that the language undergoes an intense change. But given the fact that, for all centuries, the constant is approximately in the same ranges, we can assume that the identity of the language has remained unchanged for three centuries.

Also for comparison were used "Bogurodzica" (18th century), "Creations of Adam Asnyk" (19th century), "Creations of Tadeusz Boy-Żeleński" (20th century) ad other works in Polish. The obtained data can be seen in Figure 4.

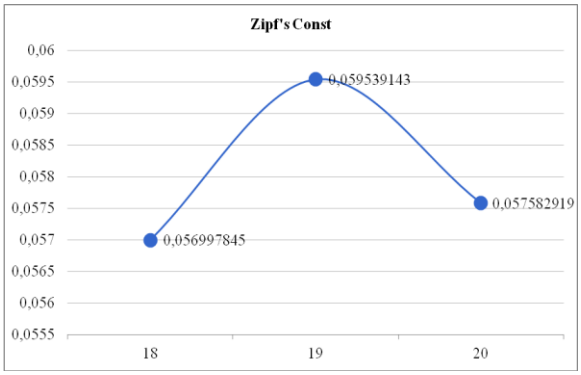


Figure 4: Counting Zipf's constant for Polish.

The data obtained indicate that the Zipf's constant for Polish works is within:

- 18th century from 0.047 to 0.067
- 19th century from 0.058 to 0.064
- 20th century from 0.054 to 0.068

The language has been subject to minimal changes over three centuries and its identity has remained almost unchanged.

4 CONCLUSIONS

Comparing the results, we can assume that the most steady and stable language with the maximum preserved identity among the languages we have chosen was Polish, and the most modified one with the preserved identity is Russian. The Ukrainian language underwent not only structural changes, but also changes in its identity (in the 19th century), but in the 20th century, the results testify to its stabilization and the identity to become close to Polish.

Regarding the practical relevance, this program can be used not only to check the language change over time, but also, for example, to check the authenticity of artistic translations, also to check the language features depending on the social sphere. Can be tracked changes in language depending on the region and migration of the population, which could be one of the points of our next research.

REFERENCES

[1] Á. Corral, G. Boleda, and R. Ferrer-i-Cancho, “Zipf’s Law for Word Frequencies: Word Forms versus

Lemmas in Long Texts", PLoS ONE, vol. 10 (7), 2015.

[2] M. Cristelli, M. Batty, and L. Pietronero, "There is More than a Power Law in Zipf", Scientific Report 2, no. 812, 2012.

[3] A. Saichev, Y. Malevergne, and D. Sornette, "Theory of Zipf 's law and beyond", Lecture Notes in Economics and Mathematical Systems 632, Springer, Heidelberg, Germany, 2010.

[4] V. Bochkarev and E. Lerner, "Calculation of Precise Constants in a Probability Model of Zipf's Law Generation and Asymptotics of Sums of Multinomial Coefficients", International Journal of Mathematics and Mathematical Sciences, vol. 17, 2017.

[5] M. Hauser, Ch. Yang, R. Berwick, I. Tattersall, M. Ryan, J. Watumull, N. Chomsky, and R. Lewontin, "The mystery of language evolution", Frontiers in Psychology, vol. 5, 2014.

[6] W. Fitch, "Empirical approaches to the study of language evolution", Psychonomic Bulletin & Review, vol. 24 (1), 2017, pp. 3-33.

[7] Y. Zhao and J. Zobel, "Search with style: Authorship attribution in classic literature", In Proceedings of the Thirtieth Australasian Computer Science Conference, Association for Computing Machinery, 2007.

[8] K. Shilova, D. Goncharenko, L. Bodnar, O. Britavska, and A. Grechkosiy, "Zipf's laws and translation approaches", Proc. of the 7th Intern, Conference "Information Technologies and Management", Riga, 2009, pp. 61-62.

[9] A. Kiv, D. Goncharenko, Ye. Sedov, L. Bodnar, and N. Yaremchuk, "Mathematical study of evolution of Russian language", Computer Modeling & New Technologies, vol. 12 (1), 2008, pp. 56-59.

[10] A. Kiv, L. Bodnar, O. Britavska, E. Sedov, N. Yaremchuk, and M. Yakovleva, "Quantitative analysis of translation texts", Computer Modeling & New Technologies, vol. 18 (12C), 2014, pp. 260-263.

[11] Analyzing and Interpreting Large Datasets. Atlanta, GA: Centers for Disease Control and Prevention (CDC), 2013.

[12] S. Bird, "Natural Language Processing with Python", O'Reilly Media Inc, 2009, p. 504.

[13] R. Dale, H. Moisl, and H. Somers, "Handbook of Natural Language Processing", Marcel Dekker, 2000.

[14] D. Mertz, "Text Processing in Python", Addison-Wesley, Boston, MA, 2003.

The Application of Multivariate Statistical Methods in Ecotoxicology and Environmental Biochemistry

Halina Falfushynska[1], Oleg Lushchak[2] and Eduard Siemens[3]
[1]*Department of physical rehabilitation and vital sequrity, Ternopil V. Hnatiuk National Pedagogical University, 2 Kryvonosa Str., Ternopil, Ukraine*
[2]*Department of biochemistry and biotechnology, Vasyl Stefanyk Precarpathian National University, 57 Shevchenko Str., Ivano-Frankivsk, Ukraine*
[3]*Anhalt University of Applied Sciences, 57 Bernburger Str., Köthen, Germany*
falfushynska@tnpu.edu.ua, olehl@pu.if.ua, eduard.siemens@hs-anhalt.de

Keywords: Multivariate Statistical Analysis, Principal Component Analysis, Linear Discriminant Analysis, Classification and Regression Tree Analysis, Pesticide Pollution, Ecotoxicology and Environmental Biochemistry.

Abstract: Pesticide pollution of surface- and groundwater are a subject of national importance indeed. However, far too little attention has been paid to find out suitable protocols and algorithms for ecotoxicological data analysis and generalisation. The aim of the present study was to implement Multivariate statistical analysis techniques for prediction of toxicity level of widely-used organophosphate pesticides to living organisms and find out the most appropriate statistical technique out of implemented to integrate biological data. The generalization of the results of biochemical and physiological measurements in zebrafish, Daphnia and Drosophila exposed to widely-used pesticides namely chlorpyrifos, roundup, and malathion have been done using principal component analysis, linear discriminant analysis and classification and regression tree analysis. All of three applied multivariate statistical techniques claimed chlorpyrifos to be the most toxic pesticides out of tested based on responses of living organisms. The importance of battery of biomarkers for risk assessment when compare to individual indices was proved using classification and regression tree analysis and discriminant analysis and daphnia's protein carbonyls level and zebrafish's lactate dehydrogenase activity pertain to the most sensitive indices for group distinguishing. We propose to combine the most widely used in life sciences Principal Component Analysis with classification and regression tree analysis and discriminant analysis to better highlight the important biological entities and reveal insightful patterns in the data.

1 INTRODUCTION

Being rapidly increasing and causing 14,000 deaths of people daily [1], water pollution has to be the matter of global concern. Sure enough that pesticides make a considerable contribution in water pollution. For example, more than 1 billion pounds of pesticides are used annually in the USA and these substances and/or their metabolites are commonly detected in both surface and groundwater bodies[1]. Likewise, 81% of small streams in Germany were accused in pollution by several pesticides in amounts that exceeded permissible concentration, while in 18% of them, these concentrations were simultaneously breached by more than 10 different pesticides [2]. Also, high concentrations of carbendazim, malathion, and diuron were found in the basin of the Llobregat River, Catalonia, Spain [3]. Nevertheless, pesticide contamination become the global threat both for animal and human because its proven toxicity against living organisms, particularly non-targets, far too little attention has been paid to find out suitable protocols and algorithms for ecotoxicological data analysis, integration and generalisation that allow researchers to conjoin efforts regarding water quality monitoring across the world and then implement standardized water quality criteria.

Multivariate statistical analysis is a quickly developed field of classical statistics which can help data processing and analysing associations between the different variables measured[2]. Principal

[1]https://biologicaldiversity.org

[2]https://web.stanford.edu/class/bios221/book/Chap-Multivariate.html

component analysis (PCA), factor analysis, and discriminant analysis belong to the most popular modules of Multivariate statistical analysis and all of them should be very helpful for data generalization, interpretation, and making meaningful conclusions and prediction. It is a valuable tool for identifying factors and sources that may affect quality of surface and groundwater systems [4].

PCA has been widely used to estimate the potential effects of pollutants on water animals and environment. In particular, data processing related to water quality conditions, spatial–temporal changes, and the driving factors in pond and cage aquaculture areas of Zhuanghe area using PCA showed that the most relevant factors of water quality in marine ecosystems are salinity, dissolved oxygen, and antibiotic resistance genes. For the cage aquaculture area chlorophyll a was determined as the additional qualification parameter [5]. Likewise, PCA emphasized that nutrient factor (39.2%), sewage and feacal contamination (29.3%), physicochemical sources of variability (6.2%) and waste water pollution from industrial and organic load (5.8%) affected water quality in the Ganges River [6]. Our prior studies also have noted the importance of PCA for risk assessment and toxic level prediction based on not only water chemical parameters but also biochemical and physiological parameters of living organisms [7]. Moreover, more profound findings come to hand when data analysis using PCA combines with other module of Multivariate statistical analysis namely discriminant analysis or data mining tools [7]. However, a few studies that used different modules of Multivariate statistical analysis and/or data mining in ecotoxicological purpose could be found in literature. Therefore, the aim of the present study was to implement Multivariate statistical analysis techniques for prediction of toxicity level of widely-used organophosphate pesticides to living organisms and find out the most appropriate statistical technique out of implemented to integrate biological data.

2 MATERIALS AND METHODS

For experimental research, we have chosen three species that differ in the level of origin (invertebrate/vertebrate) and ecological needs. We have used the cyprinid fish *Danio rerio*, as a conventional biological model for mechanistic and toxicological studies, which demonstrates universal to vertebrates, responses to stress and toxicity. *Daphnia magna* is widely spread in nature, easy in cultivation, has high sensitivity to various xenobiotics, which makes it very convenient objects for bioindication. The fruit fly *Drosophila melanogaster* was used as an extremely convenient, popular and relatively cheap model object.

Fish were treated with organophosphate pesticides roundup (commercial form of glyphosate), malathion and chlorpyrifos in two concentrations, low and high which could be designated as environmentally relevant due to average concentration of correspondent chemicals in water [8, 9]. Higher tested concentrations correspond to pesticides levels in waste waters and heavily polluted water bodies. In particular, experimental fish were exposed to roundup (RL, 15 $\mu g \cdot L^{-1}$ and RH, 500 $\mu g \cdot L^{-1}$), malathion (ML, 5 $\mu g\ L^{-1}$ and MH, 50 $\mu g\ L^{-1}$) and to chlorpyrifos (CPL, 0.1 $\mu g \cdot L^{-1}$ and CPH, 3.0 $\mu g \cdot L^{-1}$) for 14 days. Daphnia were exposed to the same organophosphate pesticides in the following concentration, Roundup (0.1; 0.5; 1 and 5 $\mu g\ L^{-1}$), chlorpyrifos at concentrations of 0.001; 0.005 and 0.01 $\mu g\ L^{-1}$, and malathion (0.1; 0, 5; 1 $\mu g\ L^{-1}$).

Complex study of zebrafish responses included next parameters: antioxidant defense (total antioxidant capacity, catalase, glutathione total, glutathione transferase); oxidative damage (formation of reactive oxygen and nitrogen species, protein carbonyls and lipid peroxidation); cytotoxicity (stability of lysosomal membranes, lactate dehydrogenase activity in the blood); neurotoxicity (cholinesterase activity); apoptosis (expression of caspase-3, cathepsin D level), DNA damage and repair (DNA fragmentation level, expression of GADD45); immune status (IgM); endocrine disruption (expression of vitellogenin in male, cortisol and triiodothyronine). The impact of pesticides on *Daphnia* was evaluated using parameters of antioxidant defense (superoxide dismutase, catalase, glutathione total); oxidative damage (lipid peroxidation and protein carbonyls). Fruit fly response was evaluated based on the activity of antioxidants (SOD, catalase, aconitase, low molecular weight thiols) and markers of oxidative stress (protein carbonyls, lipid peroxides). All abovementioned parameters were measured according to protocols as described in detail in [10, 11].

Data were tested for normality and homogeneity of variance by using Kolmogorov–Smirnoff and Levene's tests, respectively. Whenever possible, data were normalized by Box–Cox transforming method. For the data that were not normally distributed, non-parametric tests (Kruskall–Wallis

ANOVA and Mann–Whitney *U*-test) were performed. Normalized, Box–Cox transformed data were subjected to principal component analysis (PCA) to differentiate individual specimens by the set of their indices. PCA is the standard tool for extracting components and to visualize the similarities between the biological samples. It serves to find a low dimensional representation of the data which captures most of the variance. In the analysis of animal responses to adverse effects have been considered only variables with correlation >0.60 with the first two dimensions of the PCA (Factor Loadings).

The classification tree in terms of determination the complex interactions among variables and sensitive criterions that distinguish evaluated groups based was built using Classification and Regression Tree (CART)-style exhaustive search for univariate splits using non-transformed studied traits. Canonical discriminant analysis was applied to extract variables that were able to maximize the between-group separation. Scores for each CVA and for each individual were calculated and then plotted in the canonical space. Differences between groups were specified based on the Mahalanobis distance, which reflects the distance between the centroids of each group.

All statistical calculations were performed with Statistica v. 12.0. For all traits and experimental treatment groups, sample size was 8.

3 RESULTS

The generalization of the results of biochemical and physiological measurements in zebrafish, daphnia and Drosophila exposed to organophosphate pesticides by Multivariate statistical analysis confirms the peculiarities of the response of stress-responsive and detoxifying systems of Drosophila on the one hand as a target organism to the effects of insecticides and daphnia / zebrafish as non-target organisms on the other. The first two PCs account for 94 % of the total variation in the dataset. All of the factor loadings in the first PC for biological traits of zebrafish and daphnia have the same sign, so it is a weighted average of all variables, representing 'non-targets' mode of reaction'. Contrary, all groups of fruit fly which is target organism to insecticides and non-target to herbicides have negative factor loadings in the first PC. In general, in depends on the mode of response and adaptation strategy to adverse effects, insects are on the left and cyprinid fish and crustacean which are not intentionally selected for control by pesticides, but which may suffer damage because of exposure to it on the right. The purest intergroup distribution is characteristic for fruit fly, whose specimens based on the oxidative stress markers are on one side of PC 1, and daphnia / zebrafish, whose specimens are located opposite the axis of PC 1 (Figure 1 a). The joint localization of daphnia and zebrafish in the plane of plot area of factor analysis confirms the realization of common strategies of adaptation in non-target organisms to pesticides, regardless of phylogenetic origin.

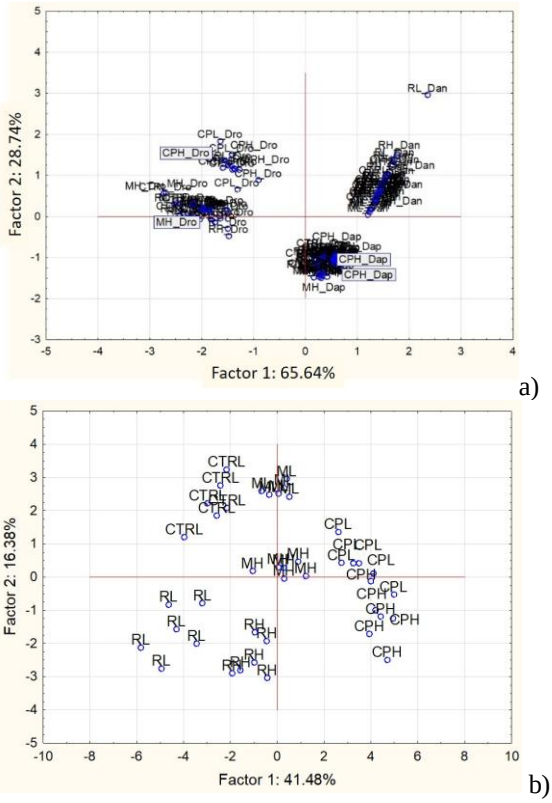


Figure 1: Principal component analysis biplot based on all measured biomarkers of zebrafish *Danio rerio* (Dan), Daphnia (Dap), and Drosophila (Dro) exposed to roundup, chlorpyrifos, and malathion. CTRL – control, RL - roundup low concentration, RH – roundup high concentration, ML – malathion low concentration, malathion high concentration, CPL – chlorpyrifos low concentration, CPH – chlorpyrifos high concentration.

On the other hand, grouping data according to the type of pesticide and its concentration shows a significant dependence of the adaptation strategy on the nature of the acting factor and the depth of its impact (Figure 1 b). Tangentially, groups separation along the axis of Factor2 occurred mainly due to parameters of zebrafish (Table 1) when for Factor1 responses of all three organisms were important.

Chlorpyrifos, is an organophosphate insecticide used to control foliage and soil-borne insect pests, becomes apparent as the most toxic pesticide to living organisms based on the sum of the biochemical and physiological markers of three studied organisms, and roundup (in environmentally relevant concentrations) and malathion manifested to be the least toxic. The specific response to the effects of pesticides in high tested concentrations corresponds to the general oppression of the health status of the organism. However, even in the case of minimal toxicity, signs of oxidative stress, cytotoxicity, inhibition of detoxification processes were observed in zebrafish and daphnia, which can probably reduce the organism's tolerance to stress when organism exposed to additional stressors and lead to irreversible changes in molecular and cellular levels that can potentially appeared themselves over time at both the organismal and population-species levels, remotely affecting biodiversity loss.

Table 1: Factor coordinates of the variables, based on correlations.

Variable	Factor 1	Factor 2
Low-weight molecular thiols (Dro)	0.842451	-0.050171
High-weight molecular thiols (Dro)	0.772227	0.022814
LOOH (Dro)	0.679562	0.200390
CAT (Dro)	-0.857893	-0.023008
Aconitase (Dro)	-0.872867	-0.329750
SOD (Dap)	-0.198962	-0.160684
CAT (Dap)	-0.730681	-0.334843
LOOH (Dap)	-0.724947	0.072561
Low-weight molecular thiols (Dap)	0.341826	-0.417394
High-weight molecular thiols (Dap)	-0.636850	0.483128
Protein Carbonyls (Dap)	-0.762287	-0.502726
ROS (Dan)	0.779434	-0.171246
TAC (Dan)	-0.673789	-0.175300
Glutathione (Dan)	-0.423798	-0.561634
RNS (Dan)	0.408960	-0.611201
GST (Dan)	-0.657437	-0.035853
CAT (Dan)	-0.723644	-0.316562
TBARS (Dan)	0.501207	-0.742588
Protein Carbonyls (Dan)	0.486866	-0.222567
Lactate dehydrogenase (Dan)	0.506237	-0.800050
Caspase 3 (Dan)	-0.693161	-0.365407
Vitellogenin (Dan)	0.260017	-0.650203
Eigenvalues	9.13	3.61

Results of the linear discriminant analysis stated that only 13 out of 22 parameters analysed of zebrafish, daphnia and drosophila were proven important in discriminating studied groups according to pesticide toxicity and adaptation strategy ($F_{(132,89)}$ =22.507 p<0.0000) (Figure 2). The first two discriminant functions accounted for 99.3% of the variation in group separation, being significant, with high Chi–Square value. The distinguishing of chlorpyrifos-treated groups from other groups were primarily due to more prominent signs of cytotoxicity and rate of lipid and protein peroxidation of animals regardless their evolutionary level. These groups locate farthest away from others according to Mahalanobis distance (> 2970) and may well designate as “the most impacted” that corresponds to “High toxicity”. On the oppose, animals have been treated with roundup and malathion locate closest to control (<721) and highly likely belong to “the less impacted” groups that corresponds to “Low toxicity”. The ratio of observed to predicted values of the biological parameters equalled to 100%.

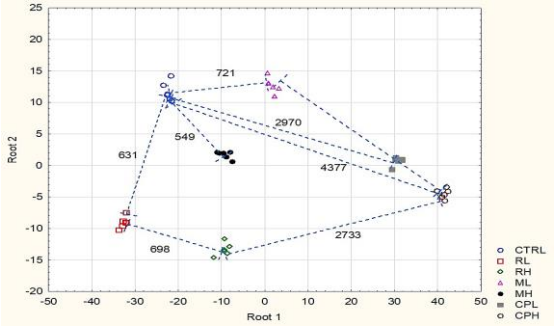


Figure 2: Discriminant analysis biplot of the physiological and biochemical traits of zebrafish *D. rerio*, Daphnia, and Drosophila exposed to roundup, chlorpyrifos, and malathion. Numbers next to the lines indicate the squared Mahalanobis distance between the respective groups. CTRL – control, RL - roundup low concentration, RH – roundup high concentration, ML – malathion low concentration, malathion high concentration, CPL – chlorpyrifos low concentration, CPH – chlorpyrifos high concentration.

It has been recently shown that discriminant analysis could be very helpful to integrate environmental data and establish a set of variables for significant discrimination between affected groups [12]. Several attempts disclosed the power of discriminant analysis to explain the differences between control and polluted vicinities based on water parameters or set of biochemical/ histopathological indices of bioindicator organism [13]. On the other hand, discriminant analysis has been implemented successfully to disclose prepotent biomarkers that allow to separate localities or conditions significantly. In particular, results of the

stepwise discriminant analysis selected nine indices that distinguished studied sites in Hong Kong due to benthic infaunal structure, physical and chemical characteristics of sediment samples, toxicity data and metal accumulation and rate of survival belonged to the most sensitive parameters [14]. We have tried to explain the differences in the mechanisms of toxic mode of pesticides action on the basis of molecular descriptors of exposed organisms which differ in ecological demands and biological organisation. It let us not only to evaluate the biohazards of widely-used pesticides, but also digitize the value of toxicity using Mahalanobis square distance. The PCA, which is more popular in biostatistics, doesn't allow that. Indeed, the combination of different statistical approach for data integration may well optimize a procedure of parameters selection for risk assessment protocols and determined the depths of biohazards. Therefore, despite the small dataset, our results depict the applicability and usefulness of discriminant analysis with close relation with other tools of multivariate statistical analysis in integrating ecotoxicological and biochemical data in terms of water quality monitoring and risk assessment. The predictive capability of the model might take advantage from an increase in sampling parameters, exposure conditions, number of chemical parameters and/or biomarker.

The data analysis using CART algorithm allowed us to identify the valuable parameters that most likely distinguish the groups of interests. The built tree consists of seven terminal leaf nodes and accounts six splits (Figure 3). The set of biomarkers includes lactate dehydrogenase activity of zebrafish blood, which primarily separates the least impacted groups (control and malathion in environmentally relevant concentration), the total antioxidant capacity and caspase 3 activity of zebrafish, as well as protein carbonyls level and catalase activity of daphnia. The most obvious finding to emerge from this study is that none of the Drosophila parameters was included in the classification tree, which clearly indicates the impossibility of using an organism that is targeted to a certain factor as a bioindicative for assessing toxicity and biohazards based on a set of biomarkers. The importance of battery of biomarkers for risk assessment when compare to individual indices was also proved by discriminant analysis and daphnia's protein carbonyls level (F-remove $_{(6,14)}$ = 25.85, $p<0.0001$) and zebrafish's lactate dehydrogenase activity (F-remove $_{(6,14)}$=13.9, $p<0.0001$) pertain to the most sensitive indices for group distinguishing.

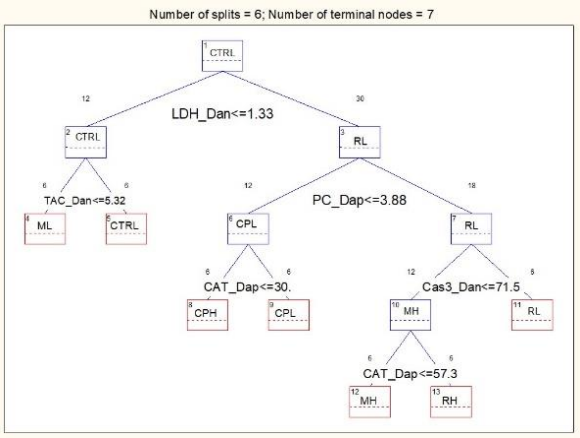


Figure 3: Classification tree of the studied biological traits of zebrafish (Dan), Daphnia (Dap), and Drosophila (Dro) exposed to roundup, chlorpyrifos, and malathion. CTRL – control, RL - roundup low concentration, RH – roundup high concentration, ML – malathion low concentration, malathion high concentration, CPL – chlorpyrifos low concentration, CPH – chlorpyrifos high concentration.

3 CONCLUSIONS

Data analysis is important in processing and generalization data sets in ecotoxicology and life sciences. We have described an approach that uses multivariate statistical techniques namely PCA, discriminant analysis and CART to find variables that are correlated across the samples, separate studied groups of specimens, localities and/or conditions, as well as discover the most sensitive indices/biomarkers that might be helpful to predict outcomes and effects while being treated. Being processed with PCA, discriminant analysis and CART, biochemical and physiological indices of zebrafish, daphnia and drosophila claimed chlorpyrifos to be the most toxic pesticides out of roundup, chlorpyrifos, and malathion. The adaptation strategy of living organisms to adverse effects depends significantly on the nature of the acting factor and the depth of its impact. The importance of battery of biomarkers for risk assessment when compare to individual indices was proved using CART and discriminant analysis and daphnia's protein carbonyls level and zebrafish's lactate dehydrogenase activity pertain to the most sensitive indices for group distinguishing. We propose to combine the most widely used in life sciences Principal Component Analysis with CART and discriminant analysis to better highlight the important biological entities and reveal insightful patterns in the data.

ACKNOWLEDGMENTS

The work was supported by the National Research Foundation of Ukraine (#2020.02/0270), Alexander von Humboldt Stiftung (H.F.), DAAD (DigIn.Net2), and Ministry of Education and Science of Ukraine (#MV-2 and BF/2-2021).

REFERENCES

[1] A. Agrawal, R. Pandey, and B. Sharma, "Water Pollution with Special Reference to Pesticide Contamination in India", Journal of Water Resource and Protection, Vol. 2, 2010, pp. 432-448.

[2] M. Liess, L. Liebmann, P. Vormeier, O. Weisner, R. Altenburger, D. Borchardt, W. Brack, A. Chatzinotas, B. Escher, K. Foit, R. Gunold, S. Henz, K. L. Hitzfeld, M. Schmitt-Jansen, N. Kamjunke, O. Kaske, S. Knillmann, M. Krauss, E. Küster, M. Link, M. Lück, M. Möder, A. Müller, A. Paschke, R. B. Schäfer, A. Schneeweiss, V. C. Schreiner, T. Schulze, G. Schüürmann, W. von Tümpling, M. Weitere, J. Wogram, and T. Reemtsma, "Pesticides are the dominant stressors for vulnerable insects in lowland streams", Water Research, Vol. 201, 2021, pp. 117262.

[3] A. Masiá, J. Campo, A. Navarro-Ortega, D. Barceló, and Y. Picó, "Pesticide monitoring in the basin of Llobregat River (Catalonia, Spain) and comparison with historical data", Sci. Total Environ., vol. 503-504, 2015, pp. 58-68.

[4] J. Liu, D. Zhang, Q. Tang, H. Xu, S. Huang, D. Shang, and R. Liu, "Water quality assessment and source identification of the Shuangji River (China) using multivariate statistical methods", PLoS ONE, vol. 16, 2021, e0245525.

[5] X. Zhang, Y. Zhang, Q. Zhang, P. Liu, R. Guo, S. Jin, J. Liu, L. Chen, Z. Ma, and Y. Liu, "Evaluation and Analysis of Water Quality of Marine Aquaculture Area", Int. J. Environ. Res. Public Health, vol. 17, 2020, pp. 1446.

[6] A. Mishra, "Assessment of water quality using principal component analysis: A case study of the river Ganges", J. Water Chem. Technol, vol. 32, 2010, pp. 227-234.

[7] H. I. Falfushynska, L. L. Gnatyshyna, and O. B. Stoliar, "In situ exposure history modulates the molecular responses to carbamate fungicide Tattoo in bivalve mollusk", Ecotoxicology, vol. 22, 2013, pp. 433-445.

[8] K. L. Newhart, "Environmental fate of malathion. California Environmental, Protection Agency Department of Pesticide Regulation", Environmental Monitoring Branch, 2006, [Online]. Available: http://www.bio-nica.info/Biblioteca/Newhart2006 Malathion.pdf.

[9] A. Ccanccapa, A. Masia, A. Navarro-Ortega, Y. Pico, and D. Barcelo, "Pesticides in the Ebro River basin: occurrence and risk assessment", Environ. Pollut, vol. 211, 2016, pp. 414-424.

[10] B. Rovenko, O. Kubrak, D. Gospodaryov, I. Yurkevych, A. Sanz, O. Lushchak, and V. Lushchak, "Restriction in glucose and fructose causes mild oxidative stress independently of mitochondrial activity and reactive oxygen species in Drosophila melanogaster", Comp. Biochem. Physiol., vol. 187, 2015, pp. 27-39.

[11] O. Bodnar, O. Horyn, I. Khatib, and H. Falfushynska, "Multibiomarker assessment in zebrafish Danio rerio after the effects of malathion and chlorpyrifos", Toxicol. Environ. Health Sci., vol. 13, 2021, pp. 165-174.

[12] P. J.Van den Brink, N. W. Van den Brink, and C. J. F. Ter Braak, "Multivariate analysis of ecotoxicological data using ordination: demonstrations of utility on the basis of various examples", Australasian journal of ecotoxicology, vol. 9, 2003, pp. 141-156.

[13] N. C. Ghisi, E. C. Oliveira, I. C. Guiloski, S. B. de Lima, H. C. Silva de Assis, S. J. Longhi, and A. J. Prioli, "Multivariate and integrative approach to analyze multiple biomarkers in ecotoxicology: A field study in Neotropical region", Sci. Total Environ., Vol. 609, 2017, pp. 1208-1218.

[14] P. K. S Shin and K. Y. S Fong, "Multiple Discriminant Analysis of Marine Sediment Data", Mar. Pollut. Bul., Vol. 39, 1999, pp. 285-294.

Dynamic Scale Adaptation Algorithm of Image Etalon Functions

Mikhail Gavrikov and Roman Sinetsky
Department of Software Engineering, Platov South-Russian State Polytechnic University, 132 Prosvecshenia Str., Novocherkassk, Russia
gmm1000@yandex.ru, rmsin@srspu.ru

Keywords: Image Recognition, Pattern Recognition, Prototype Function, Scale-Adapted Function.

Abstract: An algorithm for large-scale adaptation of prototype functions representing image classes is proposed. The algorithm identifies the parameters of nonlinear scale distortions contained in the functions representing the observed image realizations, and then transforms the original prototype function using the previously proposed model. The algorithm works on a class of images of regular phase processes that have the property of quasi-similarity of shape. Two models of large-scale nonlinear transformations are considered: symmetric and asymmetric. The differences between the models and the practical results of their application are given. The algorithm was experimentally tested on the images of prototypes of fragments of speech signals, electrocardiosignals, and engine cylinder pressure detector signals. Examples and experimental data confirming the effectiveness of the algorithm are given. Conclusions are formulated about the possibility of using the algorithm with both models in practical problems.

1 INTRODUCTION

Most image recognition methods for various physical processes use the operation of comparing the current image with the prototype (etalon) image. For example, speech recognition algorithms use many prototype spectral images of phonemes, algorithms for processing data of technical and medical diagnostics can use prototypes of signals or spectral functions characterizing changes in the parameters of controlled physical processes in various modes and states (normal operation, pathology of certain types, etc.). If the features of images belonging to the same class change little over time (within small limits relative to statistical averages), then the prototype images remain unchanged during the analysis. If the specified parameters can dynamically change within large limits during the processing of the registered image implementations, then for the reliable operation of the recognition algorithm, the parameters of the prototype images must adapt to these changes. That is, the prototype images must be dynamic.

In applied problems, prototype images are often represented in the form of real functions: fragments of signals of finite duration, analytically defined functions approximating these fragments, power spectral densities, etc. In the case of dynamic images, if the image of the prototype is represented by a function $y_0(x)$, and recorded on the interval T ($T \subset X$, X is the real axis) current image of the function $\tilde{y}_T(x)$, then before comparison it is necessary to perform the transformation $y_0(x) \to y_T(x)$ prototype functions $y_0(x)$ in some new function $y_T(x)$ that has the meaning of the adapted prototype with which to compare a function $\tilde{y}_T(x)$ that represents the current image.

Among the physical processes that are the sources of the analysed images, a significant place is occupied by the class of *regular phase processes* (RF-processes) [1, 2, 3]. In particular, such processes include dynamic processes occurring in internal combustion engines, electrical processes of polarization and depolarization of the heart muscle, etc. A characteristic feature of the class of images of RF processes is that all registered implementations of image functions $\tilde{y}_T(x)$ from this class have the property of "similarity of form" (resemblance, but not coincidence). In [4], such images were called "pulsating" and two models for their formation were proposed. In general, these models can be represented in the form of a converter $H(\tilde{s}_k, \tilde{s}_a)$ that performs large-scale transformations of the

prototype $y_0(x)$ into simulated implementations of $\hat{y}(x)$:

$$\hat{y}(x) = H(\tilde{s}_k, \tilde{s}_a)[y_0(x)], \quad (1)$$

where $\tilde{s}_k : x \to kx$, $\tilde{s}_a : y_0 \to ay_0$ is the scale transformations with random parameters k, a. If the parameters of the scale transformations are not constants, but are themselves functions of the argument $x : k = k(x)$, $a = a(x)$, then the scale distortions are nonlinear, and the functions $\tilde{y}_T(x)$ generated according to (1) will have the specified shape similarity property.

The aim of this work is to develop and experimentally test an algorithm designed to identify the parameters of nonlinear scale distortions contained in the observed implementations of a class of functions, and to construct an adapted prototype function representing this class of images.

2 THE ADAPTATION TASK

Let $e(\tilde{y}_T(x), y_0(x))$ be the deviation calculated in some way between the observed function $\tilde{y}_T(x)$ and the prototype $y_0(x)$. The task of adaptation is set as follows:

Having an observable function $\tilde{y}_T(x)$, a prototype function $y_0(x)$, and a model of scale transformations $H(\tilde{s}_k, \tilde{s}_a)$, we need to obtain estimates of the transformation parameters k, a for which the function $\hat{y}(x) = H(\tilde{s}_k, \tilde{s}_a)[y_0(x)]$ satisfies the condition

$$e(\tilde{y}_T(x), \hat{y}(x)) < e(\tilde{y}_T(x), y_0(x)), \quad (2)$$

where $e(\tilde{y}_T(x), \hat{y}(x))$ is the deviation between the corresponding functions.

The model function $\hat{y}(x)$ that satisfies condition (2) can be used as an adapted prototype function $y_T(x) = \hat{y}(x)$, with which the implementation of $\tilde{y}_T(x)$ representing the current image should be compared. A possible modification of such a statement of the problem may consist in replacing condition (2) with condition

$$e(\tilde{y}_T(x), \hat{y}(x)) < e^{\partial},$$

where e^{∂} is the specified value of the permissible deviation. Obviously, in both cases, the problem may have many solutions or not have a solution, and the problem statement itself contains a conceptual scheme for obtaining a solution, if any.

This scheme involves the following steps:

- processing of implementation $\tilde{y}_T(x)$, as a result of which estimates of the parameters of scale transformations $\tilde{s}_k, \tilde{s}_a$ should be obtained;
- substitution of the obtained parameter values into the used model of the converter $H(\tilde{s}_k, \tilde{s}_a)$ and generation of the adapted prototype function $\hat{y}(x) = H(\tilde{s}_k, \tilde{s}_a)[y_0(x)]$;
- checking condition (2) and, if it is fulfilled, replacing the original prototype function $y_0(x)$ with an adapted prototype function $y_T(x) = \hat{y}(x)$.

3 SCALE CONVERTER MODEL

As a model of the transformer $H(\tilde{s}_k, \tilde{s}_a)$, we can use a modification of the relations obtained in [4] to model a set of implementations of $y_n(x)$ by stochastic nonlinear scale distortions of the prototype function $y_0(x)$. In this work, the relations given below do not contain the index *n*, since they are used not for modelling, but for generating an adapted function when performing the second stage of the above conceptual adaptation scheme.

The prototype function is given in the form of a piecewise function

$$y_0(x) = \begin{cases} y_0^{(1)}(x), & x_{(0)}^{\bullet} = 0 \le x < x_{(1)}^{\bullet}, \\ \ldots \\ y_0^{(m)}(x), & x_{(m-1)}^{\bullet} \le x \le x_{(m)}^{\bullet} = b^{\bullet}, \end{cases}$$

defined by some partition

$$\pi_m(D^{\bullet}) = \{d_i \mid d_i = [x_{(i-1)}^{\bullet}, x_{(i)}^{\bullet}),\ i = 1, \ldots, m,\ d_m = [x_{(m-1)}^{\bullet}, x_{(m)}^{\bullet}],$$
$$0 = x_{(0)}^{\bullet} < x_{(1)}^{\bullet} < \ldots < x_{(m)}^{\bullet} = b^{\bullet}\}$$

of the domain of the function definition $D^{\bullet} = [0, b^{\bullet}]$ with nodal points $q(x_{(i)}^{\bullet}, y_{(i)}^{\bullet})$ [5, 6].

Modelling of $\hat{y}(x)$ is performed by linear scale transformations $y_0^{(i)}(x) \to \hat{y}^{(i)}(x)$ of each i-th segment of the piecewise prototype function, but with different values of the scale transformation parameters for different segments.

The functions $\hat{y}^{(i)}(x)$ formed at the output of the converter are determined by the relations:

$$\hat{y}^{(i)}(x) = a^{(i)} y_0^{(i)}\left(\frac{x - x^{\bullet}_{(i-1)}(1-b^{(i)}) - \delta^{(i-1)}}{r\ b^{(i)}}\right) + \\ + y_0(x^{\bullet}_{(i-1)})(1-a^{(i)}) + \xi^{(i)}, \quad (3)$$

where $a^{(i)}, b^{(i)}$ is the coefficients of scale transformations:

$$a^{(i)} = 1 + \frac{\xi^{(i)} - \xi^{(i-1)}}{y_0(x^{\bullet}_{(i)}) - y_0(x^{\bullet}_{(i-1)})},\ b^{(i)} = 1 + \frac{\delta^{(i)} - \delta^{(i-1)}}{x^{\bullet}_{(i)} - x^{\bullet}_{(i-1)}},$$

that depend on the values of random parameters $\xi^{(i)},\ \delta^{(i)}$ with the characteristics:

$$M\{\xi^{(i)}\} = 0,\ \xi^{(i)} \in [-\Delta_y^{(i)}, \Delta_y^{(i)}],$$
$$M\{\delta^{(i)}\} = 0,\ \delta^{(i)} \in [-\Delta_x^{(i)}, \Delta_x^{(i)}],$$

$r = const$ is the coefficient of linear change in the scale of argument x over the entire interval $D^{\bullet} = [0, b^{\bullet}]$.

The use of a constant coefficient r leads to linear compression/expansion of the prototype function $y_0(x)$, and the use of different values of the coefficients of scale transformations $a^{(i)}, b^{(i)}$ for its different segments leads to nonlinearity of scale distortions that will be contained in the resulting function $\hat{y}(x),\ x \in T = [0, \hat{b}],\ \hat{b} = rb^{\bullet}$.

Internal node points $q(x_{(i)}, y_{(i)})$ of the modelled piecewise function (with the exception of the boundary points $q(x_{(0)}, y(x_{(0)}))$, $q(x_{(m)}, y(x_{(m)}))$) satisfy the constraint:

$$(x_{(i)}, y_{(i)}) \in \Theta_{(i)} = [x^{\bullet}_{(i)} - \Delta_x^{(i)},\ x^{\bullet}_{(i)} + \Delta_x^{(i)}] \times \\ \times [y_0(x^{\bullet}_{(i)}) - \Delta_y^{(i)},\ y_0(x^{\bullet}_{(i)}) + \Delta_y^{(i)}] \quad (4)$$

which means that they fall into the *rectangles of movement of nodal points* $\Theta_{(i)}$.

This feature binds the contours of the simulated functions $\hat{y}(x)$ to the contour of the prototype $y_0(x)$ and preserves the similarity property of their shapes. The values of $\Delta_x^{(i)},\ \Delta_y^{(i)}$ set the allowable increments of the ranges of the definition areas and the ranges of values of the corresponding segments of the functions $y^{(i)}(x)$, provided that the adjacent rectangles $\Theta_{(i)}$ should not intersect: $\Theta_{(i)} \cap \Theta_{(i+1)} = \varnothing$. Therefore, $\Delta_x^{(i)},\ \Delta_y^{(i)}$ is defined as:

$$\Delta_x^{(i)} = c_x \min\left\{(x^{\bullet}_{(i)} - x^{\bullet}_{(i-1)}),\ (x^{\bullet}_{(i+1)} - x^{\bullet}_{(i)})\right\},$$
$$\Delta_y^{(i)} = c_y \min\left\{(y_0(x^{\bullet}_{(i)}) - y_0(x^{\bullet}_{(i-1)})),\right. \\ \left.(y_0(x^{\bullet}_{(i+1)}) - y_0(x^{\bullet}_{(i)}))\right\}, \quad (5)$$
$$0 \le c_x, c_y < 0.5.$$

The rectangles $\Theta_{(i)}$ defined in this way are symmetric with respect to the node points $q(x^{\bullet}_{(i)}, y^{\bullet}_{(i)})$ of the prototype $y_0(x)$.

A possible modification of this method is to set the rectangles that are not symmetric in the coordinate while maintaining the condition $\Theta_{(i)} \cap \Theta_{(i+1)} = \varnothing$:

$$\Theta_{(i)} = [x^{\bullet}_{(i)} - \overleftarrow{\Delta}_x^{(i)}, x^{\bullet}_{(i)} + \vec{\Delta}_x^{(i)}] \times \\ \times [y_0(x^{\bullet}_{(i)}) - \Delta_y^{(i)}, y_0(x^{\bullet}_{(i)}) + \Delta_y^{(i)}], \quad (6)$$
$$\overleftarrow{\Delta}_x^{(i)} = c_x(x^{\bullet}_{(i)} - x^{\bullet}_{(i-1)}),\ \vec{\Delta}_x^{(i)} = c_x(x^{\bullet}_{(i+1)} - x^{\bullet}_{(i)}).$$

From relation (3) it follows that for an unambiguous specification of the converter it is necessary to determine the value of r and the set of values $\Omega = \left\{(\xi_n^{(i)}, \delta_n^{(i)}),\ i = \overline{1, m-1}\right\}$, on which the parameters of scale transformations $a^{(i)}, b^{(i)}$ depend.

4 ALGORITHM DESCRIPTION

For modelling functions with a specific shape, it was suggested in [4] to choose as the nodal points of the prototype function not arbitrary points [7, 8], but essential points $q(x, y),\ y = y_0(x)$, where significant changes in the properties of the function occur. The set of formal features used to identify each essential point $q(x, y)$ (for example, the features of the extremums of a function) and determining the behaviour of the function $y_0(x)$ in its neighbourhood determines the type $type(q)$ of this point [1, 2]. The choice of the composition of the types of essential points and the method (algorithm) of their automatic identification depends on the form or other properties of the function and is carried out by the developer of the signal processing system based on knowledge of the subject. The set of node points and their function types $y_0(x)$ will be denoted as $Q^{\bullet}$ and $TYPE^{\bullet}$, respectively:

$$Q^{\bullet} = \left\{q(x^{\bullet}_{(i)}, y^{\bullet}_{(i)}),\ i = \overline{(1, m)}\right\},$$

$$TYPE^{\bullet} = \left\{type\ q(x^{\bullet}_{(i)}, y^{\bullet}_{(i)}),\ i = (\overline{1,m})\right\}.$$

The logic of the algorithm for identifying the parameters of scale distortions of functions given below is based on the following considerations.

The desired scale-adapted function $\hat{y}(x) = H(\tilde{s}_k, \tilde{s}_a)[y_0(x)]$ must have a geometric similarity of shape both with any of the implementations of $y_n(x)$ modelled using model (3), including the prototype function $y_0(x)$, and with the observed function $\tilde{y}_T(x)$. The characteristic features of the form of a function are largely determined by the quantitative relations between the values of the coordinates of its essential points, as well as the order in which their types are located. In order to ensure the specified similarity, the following conditions must be met:

- types of nodal points $\hat{q}(\hat{x}_{(i)}, \hat{y}_{(i)})$ of the function $\hat{y}(x)$ needs to match the types of the nodal points $q(x^{\bullet}_{(i)}, y^{\bullet}_{(i)})$ of the original function prototype $y_0(x)$, and the points to get into the rectangles of movement of nodal points $\Theta_{(i)}$:

$$type\ (\hat{x}_{(i)}, \hat{y}_{(i)}) = type\ q(x^{\bullet}_{(i)}, y^{\bullet}_{(i)}),$$
$$(\hat{x}_{(i)}, \hat{y}_{(i)}) \in \Theta_{(i)};$$

- the transformation $q(x^{\bullet}_{(i)}, y^{\bullet}_{(i)}) \to \hat{q}(\hat{x}_{(i)}, \hat{y}_{(i)})$ of the node points of the prototype function $y_0(x)$ to the node points (with new coordinates) of the function $\hat{y}(x)$ should be performed only if the corresponding rectangle $\Theta_{(i)}$ contains at least one essential point $q(x, \tilde{y})$ of the $type\ q(x^{\bullet}_{(i)}, y^{\bullet}_{(i)})$ of the observed function $\tilde{y}_T(x)$, that is, if the condition is met:

$$(x, \tilde{y}) \in \Theta_i,\quad type\ q(x, \tilde{y}) = type\ q(x^{\bullet}_{(i)}, y^{\bullet}_{(i)}).$$

In addition, the domain $D^{\bullet} = [0, b^{\bullet}]$ of the prototype $y_0(x)$ must be adjusted to the domain $T = [0, \tilde{b}]$ of the observed function $\tilde{y}_T(x)$. To do this, perform a linear transformation of the scale of the argument x and create a prototype:

$$y'_0(x) = y_0\left(\frac{x}{r}\right),\quad r = \tilde{b} / b^{\bullet}. \qquad (7)$$

The coordinates of the node points q'_i of the prototype $y'_0(x)$ will be $(rx^{\bullet}_{(i)}, y'_0\ (rx^{\bullet}_{(i)}))$. These conditions are taken into account in the algorithm when identifying scale distortions and constructing a scale-adapted function $\hat{y}(x)$.

Algorithm A1. Algorithm for dynamic scale adaptation of the function.

Input:

$y_0(x)$ – the prototype function;

$\tilde{y}_T(x)$ – the observed function;

$Q^{\bullet} = \left\{q(x^{\bullet}_{(i)}, y^{\bullet}_{(i)}),\ i = (\overline{1,m})\right\}$ – the set of nodal points for the function $y_0(x)$;

$TYPE^{\bullet} = \left\{type\ q(x^{\bullet}_{(i)}, y^{\bullet}_{(i)}),\ i = (\overline{1,m})\right\}$ – the set of types of nodal points for the function $y_0(x)$.

Output:

$y_T(x)$ – the scale-adapted prototype function.

Step 1. Using the relations (7), create a prototype $y'_0(x)$:

$$y'_0(x) = y_0\left(\frac{x}{r}\right),\ x \in T = [0, \tilde{b}],\quad r = \tilde{b} / b^{\bullet}$$

(The coordinates of the nodal points q'_i of the prototype $y'_0(x)$ will be $(rx^{\bullet}_{(i)}, y'_0\ (rx^{\bullet}_{(i)}))$);

Step 2. Define the set of essential points $\tilde{Q}$ of the observed function $\tilde{y}(x)$ as:

$$\tilde{Q} = \left\{q(x_{(j)}, \tilde{y}_{(j)}) \mid type\, q(x_{(j)}, \tilde{y}_{(j)}) \in TYPE^{\bullet},\ j = (\overline{1,n})\right\}$$

(to determine this set, we can use the algorithm for constructing a piecewise monotone function from [1]);

Step 3. Calculate the deviation $e(\tilde{y}(x), y'_0(x))$ between the observed function $\tilde{y}(x)$ and the prototype $y'_0(x)$:

$$e(\tilde{y}(x), y'_0(x)) = \int_0^b \left|\tilde{y}(x) - y'_0(x)\right| dx$$

Step 4. For each node point of the prototype function $y'_0(x)$, define symmetric:

$$\Theta_{(i)} = [rx^{\bullet}_{(i)} - \Delta^{(i)}_x,\ rx^{\bullet}_{(i)} + \Delta^{(i)}_x] \times$$
$$\times [y'_0\ (rx^{\bullet}_{(i)}) - \Delta^{(i)}_y,\ y'_0\ (rx^{\bullet}_{(i)}) + \Delta^{(i)}_y]$$

or non-symmetric (optional):

$$\Theta_{(i)} = [rx^{\bullet}_{(i)} - \vec{\Delta}^{(i)}_x,\ rx^{\bullet}_{(i)} + \vec{\Delta}^{(i)}_x] \times$$
$$\times [y'_0\ (rx^{\bullet}_{(i)}) - \Delta^{(i)}_y,\ y'_0\ (rx^{\bullet}_{(i)}) + \Delta^{(i)}_y]$$

rectangles, respectively (see relations (5-6));

Step 5. Generate sets C_i of essential points with *type* $q(x^{\bullet}_{(i)}, y^{\bullet}_{(i)})$ of the observed function $\tilde{y}(x)$ falling into rectangles $\Theta_{(i)}$:

$$C_i = \{ q(x_{(j)}, \tilde{y}_{(j)}) \mid q(x_{(j)}, \tilde{y}_{(j)}) \in \tilde{Q},$$
$$(x_{(j)}, \tilde{y}_{(j)}) \in \Theta_{(i)}, type\ \tilde{q}(x_{(j)}, y_{(j)}) = type\ q(x^{\bullet}_{(i)}, y^{\bullet}_{(i)}) \}$$

and determine the cardinalities N_i of these sets (C_i may be empty);

Step 6. Determine the coordinates of the node points $\hat{q}(\hat{x}_{(i)}, \hat{y}_{(i)})$ of the function $\hat{y}(x)$ as

$$(\hat{x}_{(i)}, \hat{y}_{(i)}) = \begin{cases} \left(\frac{1}{N_i} \sum_{j=1}^{\hat{N}_{(i)}} x_{(j)}, \frac{1}{N_i} \sum_{j=1}^{\hat{N}_{(i)}} \tilde{y}_{(j)} \right), \\ \qquad q(x_{(j)}, \tilde{y}_{(j)}) \in C_i, if\ C_i \neq \varnothing, \\ (x^{\bullet}_{(i)}, y^{\bullet}_{(i)}),\ \ if\ \ C_i = \varnothing, \end{cases}$$

and the set Ω of estimates of the scale distortion parameters $(\xi^{(i)}, \delta^{(i)})$ as

$$(\xi^{(i)}, \delta^{(i)}) = \begin{cases} (\hat{x}_{(i)} - r x^{\bullet}_{(i)},\ \hat{y}_{(i)} - y^{\bullet}_{(i)}),\ \ if\ C_i \neq \varnothing, \\ 0,\ if\ C_i = \varnothing; \end{cases}$$

Step 7. Using the relations (3) and the found values of the scale distortion parameters $(\xi^{(i)}, \delta^{(i)})$, construct a piecewise function $\hat{y}(x)$.

Step 8. Calculate deviation

$$e(\tilde{y}(x), \hat{y}(x)) = \int_0^b |\tilde{y}(x) - \hat{y}(x)|\, dx;$$

Step 9. Define a scale-adapted function

$$y_T(x) = \begin{cases} \hat{y}(x),\ \ if\ \ e(\tilde{y}(x), \hat{y}(x)) < e(\tilde{y}(x), y_0(x)), \\ y_0(x),\ \ if\ \ e(\tilde{y}(x), \hat{y}(x)) \geq e(\tilde{y}(x), y_0(x)). \end{cases}$$

End.

5 EXPERIMENTS

In the experimental testing of the algorithm, the prototype images of $y_0(x)$ were used:

a) spectral power density of fragments of speech signals corresponding to certain sounds of human speech (phonemes) [9, 10, 11, 12];

b) fragment of an electrocardiogram (ECG) that corresponds to several complete cycles of the heart at a certain heart rate, seismo- or gyro-cardiogram signals [13, 14, 15];

c) fragment of the function of pressure in the cylinder of a four-stroke internal combustion engine (ICE) versus time (the standards were obtained by analytical calculation at 8900 rpm) [16].

Two types of points were used as the nodal points of the prototype $y_0(x)$: local minima and maxima, which divide the function into intervals $\tilde{y}_T(x)$ of non-strict monotony, on which the function has a positive or negative trend [7]. As observed functions $\tilde{y}_T(x)$ with nonlinear scale distortions, we used implementations $y_n(x)$ generated using the model (3) with 10000 implementations for each case (a) and (b). In the algorithm, both methods of determining the rectangles of movement of the nodal points $\Theta_{(i)}$ were implemented: symmetric and asymmetric.

An illustration of the algorithm is shown in Figure 1, Figure 2 and Figure 2. In Figure 1, the broken line shows the prototype spectrum $y_0(x)$ for the phoneme "E", the dotted line shows the observed spectrum $\tilde{y}_T(x)$, the solid line shows the scale-adapted prototype $y_T(x) = \hat{y}(x)$, as well as the symmetrical rectangles of the movement of the nodal points $\Theta_{(i)}$. Similar ECG graphs are shown in Figure 2.

Figure 1 shows that the nodal points of the observed spectrum $\tilde{y}_T(x)$ fall into the second and third rectangles, so the average peak of the adapted spectrum $y_T(x)$ is pulled up to the average peak of the observed spectrum $\tilde{y}_T(x)$ and in this area differs significantly from the prototype $y_0(x)$.

From Figure 2 it can be seen that the nodal points of the observed ECG cycle fall into four rectangles and the scale adaptation is performed in the area corresponding to the QRS complex and the T wave.

In fig. 3, it can be seen that the key points corresponding to the maximum pressure value that is observed at the moment of ignition of the combustible mixture fall into the movement rectangles, and the standard is adapted for one-cylinder operation cycle.

In all cases, the deviation of the adapted prototype from the observed function is reduced, in particular, for the example shown in Figure 1, the deviation is $e(\tilde{y}_T(x), y_0(x)) = 133476$, $e(\tilde{y}_T(x), y_T(x)) = 119742$, which is 10% less and satisfies the condition (2).

Tables 1, 2 and 3 provide a summary of the results of experiments to evaluate the effectiveness of the algorithm. Two indicators were evaluated:

- *adaptation rate* expressed as the percentage of cases in which the algorithm performed prototype adaptation with a decrease in deviation, relative to the total number of experiments;
- the *relative decrease R of the deviation* $e(\tilde{y}_T(x), \hat{y}(x))$ compared to the deviation $e(\tilde{y}_T(x), y_0(x))$, calculated as

$$R = \frac{e(\tilde{y}_T(x), y_0(x)) - e(\tilde{y}_T(x), \hat{y}(x))}{e(\tilde{y}_T(x), y_0(x))} 100\% .$$

The minimum, maximum, and average *R* values of 10000 implementations are calculated separately.

These indicators are of interest, since they characterize the "degree of relevance" of applying the procedure for adapting etalon images (prototypes) in the problems of parametric identification and pattern recognition in specific

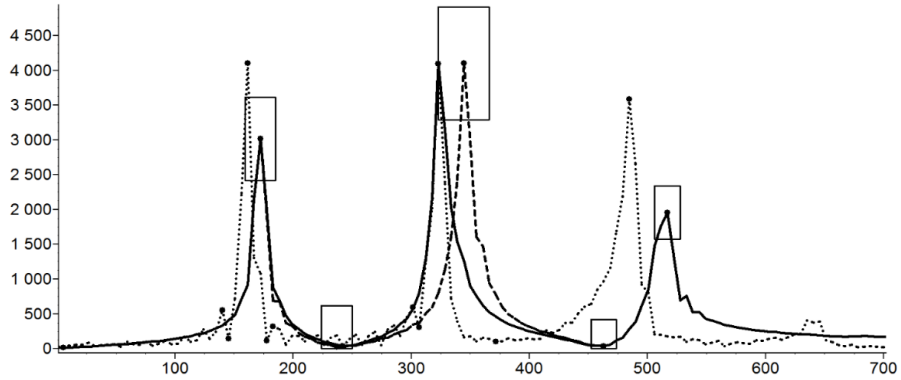


Figure 1: Illustration of the algorithm for the speech signal.

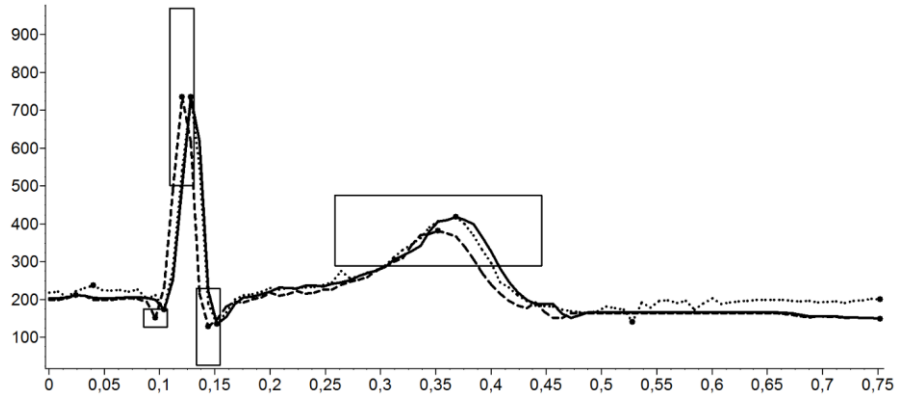


Figure 2: Illustration of the algorithm for the electro cardio signal.

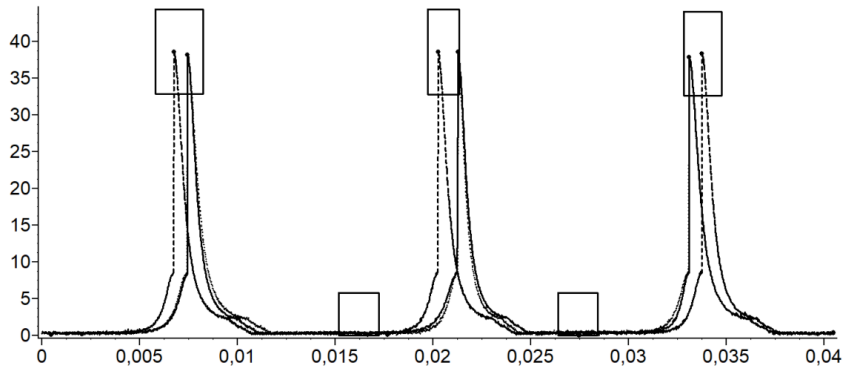


Figure 3: Illustration of the algorithm for cylinder pressure function.

applied areas. The higher the values of these indicators, the higher the "degree of relevance" of applying the adaptation procedure using the above algorithm. These indicators were calculated using the symmetric and asymmetric rectangles algorithm.

Table 1: Results for the speech signal.

Rectangle type	Adaptation rate, %	Relative deviation decrease R, %		
		Min	Average	Max
Symmetrical	65	0,02	5,06	14,24
Asymmetrical	91	0,23	7,95	21,52

Table 2: Results for the electro cardio signal.

Rectangle type	Adaptation rate, %	Relative deviation decrease R, %		
		Min	Average	Max
Symmetrical	94	0,17	29,21	58,95
Asymmetrical	97	0,32	38,06	71,85

Table 3: Results for the cylinder pressure function.

Rectangle type	Adaptation rate, %	Relative deviation decrease R, %		
		Min	Average	Max
Symmetrical	93	0,00	20,97	47,74
Asymmetrical	99	19,6	49,67	77,03

6 CONCLUSIONS

Two main conclusions follow from the analysis of the results obtained.

The frequency of adaptation of the prototype of more than 50% indicates the adequacy of the algorithm of the real situation, which was discussed at the beginning of this work. Static definition of an image class in the form of an immutable prototype function can lead to significant deviations of the functions representing the observed images from this class, due to their scale distortions, and to recognition errors, in which an image belonging to class $\{y_0(x)\}$ is recognized as not belonging to it. Using the prototype adaptation algorithm allows you to significantly reduce these deviations.

When using asymmetrical rectangles, the expected higher adaptation rates are achieved. At the same time, if the value of the constant c_x, which defines the asymmetrical rectangles in the relations (6), is not selected correctly (for example, close to 0.5), a different kind of error may occur, when the prototype $y_0(x)$ can be adapted to the image of another class. In this case, an image that does not belong to class $\{y_0(x)\}$ will be incorrectly recognized as belonging to it.

REFERENCES

[1] M. M. Gavrikov, "Structural approximation and recognition of one-dimensional time images", Concept and applications. Russian Electromechanics, 2003, vol 6, pp. 52-60.

[2] U. Grenander, "Lectures in Pattern Theory", Volume I-III, Springer, 1976-1981, Berlin.

[3] R. van der Vlist, C. Taal, and R. Heusdens, "Tracking Recurring Patterns in Time Series Using Dynamic Time Warping", 2019 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2019, pp. 1-5, doi: 10.23919/EUSIPCO. 2019.8903102.

[4] M. M. Gavrikov and R. M. Sinetsky, "Algorithms for the simulation of signals and spectral functions with a pulsating scale distortions", University news, North-Caucasian region, Technical sciences series, 2013, vol 3, pp. 3-9.

[5] M. M. Gavrikov and R. M. Sinetsky, "Algorithms for segmentation of structural time images and their application in speech signal processing", University news, North-Caucasian region, Technical sciences series, 2010, vol 1, pp. 18-24.

[6] A. Stan, C. Valentini-Botinhao, B. Orza, and M. Giurgiu, "Blind speech segmentation using spectrogram image-based features and Mel cepstral coefficients", 2016 IEEE Spoken Language Technology Workshop (SLT), San Diego, CA, 2016, pp. 597-602, doi: 10.1109/SLT.2016.7846324.

[7] L. S. Fainzilberg, "Information technologies for processing complex waveforms", Theory and practice, Ukraine, Kiev: Naukova dumka, 2008, 336 p.

[8] A. Mazumdar and L. Wang, "Covering arbitrary point patterns", 2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 2012, pp. 2075-2080, doi: 10.1109/Allerton.2012.6483478.

[9] M. M. Gavrikov and R. M. Sinetsky, "Algorithmic and numerical implementation of the structural approximation method for speech pattern recognition", Russian Electromechanics, 2007, vol 2, pp. 52-59.

[10] J. B. Allen, "Short-term spectral analysis, synthesis, and modification by discrete Fourier transform", IEEE Trans. on Acoustics, Speech, Signal Processing, 1997, vol. ASSP-25. N 3, pp. 235-238.

[11] K. Vijayan and K. S. R. Murty, "Analysis of Phase Spectrum of Speech Signals Using Allpass Modeling", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 12, pp. 2371-2383, Dec. 2015, doi: 10.1109/TASLP.2015.2479045.

[12] L. V. Zlatoustova, R. K. Potapova, and V. N. Trunin-Donskoy, "General and applied phonetics", Moscow, MSU, 1986, 304 p.

[13] L. S. Fainzilberg, “Heart functional state diagnostic using pattern recognition of phase space ECG-images”, Proc. 6th European Congress on Intelligent Techniques and Soft Computing (EUFIT '98). Aachen (Germany), 1998, N B-27, pp. 1878-1882.

[14] M. Yochum, Ch. Renaud, and S. Jacquir, “Automatic detection of P, QRS and T patterns in 12 leads ECG signal based on CWT”, Biomedical Signal Processing and Control, Elsevier, 2016, ff10.1016/j.bspc.2015.10.011ff. ffhal-01328478

[15] C. Yang, N. D. Aranoff, P. Green, and N. Tavassolian, “Classification of Aortic Stenosis Using Time–Frequency Features From Chest Cardio-Mechanical Signals”, IEEE Transactions on Biomedical Engineering, vol. 67, no. 6, pp. 1672-1683, June 2020, doi: 10.1109/TBME.2019.2942741.

[16] Q. Wang, T. Sun, Z. Lyu, and D. Gao, “A Virtual In-Cylinder Pressure Sensor Based on EKF and Frequency-Amplitude-Modulation Fourier-Series Method”, Sensors 2019, 19, 3122, [Online]. Available: https://doi.org/10.3390/s19143122.

Information Technology for Land Degradation Assessment Based on Remote Sensing

Nataliia Kussul[1,2], Andrii Shelestov[1,2], Leonid Shumilo[3], Dmytro Titkov[1] and Hanna Yailymova[1,2]
[1]*Educational and Research Institute of Physics and Technology, National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37 Peremohy avenue, Kyiv, Ukraine*
[2]*Department of Space Information Technologies and System, Space Research Institute NAS Ukraine and SSA Ukraine, 40 Glushkov avenue, Kyiv, Ukraine*
[3]*Department of Geography, The University of Maryland, College Park, USA*
{nataliia.kussul, andrii.shelestov}@lll.kpi.ua, lshumilo@umd.edu, dmytrotitkov@gmail.com, anna.yailymova@gmail.com

Keywords: Geospatial Data Analysis, Machine Learning, Land Degradation, Remote Sensing, Land Cover.

Abstract: Since the launch of ESA Copernicus program, satellite data of high resolution became publicly available and methods and tools for their automated processing to solve a wide range of applications have developed rapidly. An important scientific task is to assess land degradation and achieve zero levels of degradation. There are many methods for determining land degradation. Known approaches to the tasks of environmental land monitoring usually use the same methodology for all types of land cover. The paper represents the approach to the calculation of land degradation based on remote sensing data and modelling results taking into account the specifics of land degradation for different land cover and land use types. Our method is based on the classification of different land cover and land use types from satellite imagery and application of different schemes of land degradation assessment for each of them. We consider forest cuts as land degradation for forests and assess them using deep learning models. Land degradation for croplands is estimated by comparison of real leaf area index (LAI) and ideal LAI, calculated with the bio-physical crop development model. And land degradation for grassland is determined with a traditional approach based on vegetation index NDVI extracted from satellite imagery. The proposed approach was implemented for the territory of Ukraine.

1 INTRODUCTION

Since the advent of publicly available satellite data, methods and tools for their automated processing to solve a wide range of applications have developed rapidly [1]. There are a large number of scientific publications on this topic, including [2], [3], [4], etc. An important scientific task is to assess land degradation and achieve zero levels of degradation [5].

There are many methods for determining land degradation [6]. Known approaches to the tasks of environmental land monitoring usually use the same methodology for all types of land cover. In particular, an analysis of existing approaches to the calculation of quantitative indicators (eg sustainable development indicators) shows that all these indicators are calculated in the same way for all territories. It is natural to assume that each of the types of land cover has its own characteristics at the level of the subject area, which is reflected at the level of relevant mathematical models and can potentially be used to improve the performance of methods of processing these data.

For example, the inaccuracies of the productivity map, which can be traditionally built only on the basis of the NDVI index, include the following: in Volyn and the Carpathians (forest regions) productivity dynamics calculated on the basis of the NDVI index will be low, but this result is due to the index NDVI for forests. In the south (Zaporizhzhya and Mykolayiv oblasts), the productivity calculated on the basis of the NDVI index will be the best in Ukraine, but this is due to the fact that most of the fields in these regions are irrigated. Irrigation, in turn, is one of the factors of sustainable development, so it is advisable to consider it for agricultural fields. Therefore, the urgent task is to develop a differentiated approach to assessing the degradation of different types of land cover.

2 METHODOLOGY

This study proposes a comprehensive method for determining land degradation based on Sentinel satellite data [7], which takes into account the specifics of degradation of different types of land cover. The algorithm for determining the level of land degradation consists of the following steps.

1) Based on satellite data, a land cover classification map is built, which includes various types of crops, uncultivated land (meadows, pastures, grassland, etc.), forests, shrubs, man-made objects, water bodies, swamps, bare land. To deliver a crop classification map we use deep learning neural network model which has been trained on time series of satellite imagery and in-situ data [8], [9].

Three groups are separated of the received classes: agricultural land (which includes the main majority crops - cereals, sunflowers, maize, rapeseed, and soybeans), uncultivated lands (meadows, pastures, grassland), and forests. Given that these three groups cover 90.5% of the entire territory of Ukraine, and other lands are artificial and water objects, wetland, and bare land, we will consider that all strategically important territories of Ukraine are considered.

2) Each of the above groups has its own method of analysis.

3) The general map of land degradation is built by reducing the quantitative indicators for each of the groups (diapason of values of indicators in the general case are different) to some common set of values (the same for each group).

Quantitative indicators of agricultural land productivity will be the ratio of the real LAI index (according to satellite data) to the "ideal" for the relevant conditions of the LAI index. To calculate the latter in this study, it is proposed to use biophysical modeling Crop Growth Modeling System based on the WOFOST model. If the values of the real and "ideal" LAI indices differ slightly, the relevant agricultural area is considered not degraded, and if they differ significantly - degraded.

As deforestation is a significant problem, deforestation in some areas can be considered as an indicator of degradation. However, when searching for felling, it is necessary to take into account the fact that the area cut down but planted with new trees cannot be considered as degraded. In this study, it is proposed to use a neural network of U-Net architecture with an Efficientnet B3 encoder to search for fellings.

As practice shows, for the grassland class (uncultivated land), the use of a standard approach based on the NDVI index gives qualitative and adequate results. Degradation of uncultivated land is determined by the negative trend of the NDVI vegetation index.

2.1 Calculation of Land Degradation for Different Land Cover/Land Use Types

The main idea of the construction of the general map of land degradation is to combine the results for each of the groups of classes of the land cover and to build a general map of land degradation on the basis of general quantitative indicators. The procedure for combining the results can be formally presented as follows.

Quantitative indicators of degradation of territories are:

- for agricultural land: $f_{crop}(LAI_{real}, LAI_{perfect})$, where LAI_{real} is the real LAI index for the pixel (x, y), $LAI_{perfect}$ is the "ideal" (simulated) LAI index for the pixel (x, y). Real LAI is extracted from the satellite imagery (monthly composites for the corresponding growing year), and $LAI_{perfect}$ – is simulated with CGMS system based on WOFOST bio-physical model for each crop independently. The simulated $LAI_{perfect}$ was determined for each day, after which the maximum value for each month was calculated. The model utilizes information on meteorological parameters for each day of the vegetation period as well as profiles for different crops and soil types. A quantitative indicator of degradation of agricultural areas will be the difference between satellite LAI and simulated LAI.
- for uncultivated land (grassland): $f_{grassland}(\{NDVI\})$, where $\{NDVI\}$ is the time series of NDVI indices for the pixel (x, y) during the vegetation period. For each year (from 2001 to 2021), the maximum value of NDVI in each pixel is calculated and the trend of its change is analyzed.
- for forests: $f_{forest}(d)$, where $d=0$ in the case of forestcut, and $d=1$ otherwise. Using our own neural network approaches to forest cover, a "change detection" approach was applied for each year separately.

2.2 Unification of Land Degradation Indicators

Let's areas of available values of these functions: $E(f_{crop})$, $E(f_{grassland})$, $E(f_{forest})$, respectively. As described in Section 2.1, we obtained the following input data for each pixel (x, y):

$$I(x, y) = \{LAI_{real}, LAI_{perfect}, \{NDVI\}, d\}, \quad \textbf{3.1}$$

calculated on the basis of available input data (satellite images, meteorological data), or form standard methods (such as time series of NDVI indices).

The general indicator of degradation of the territory to which the pixel (x, y) on the raster map corresponds is f (x, y) with the range of possible values of E(f). The function f (x, y) must satisfy the following conditions: be monotonic, take the minimum value for the most degraded areas, take the maximum value for the areas with the most sustainable development.

Taking into account the above, for each degradation index f_{crop}, $f_{grassland}$, f_{forest} it is necessary to set the appropriate conversion functions:

$$K_{crop}: E(f_{crop}) \rightarrow E(f),$$
$$K_{grassland}: E(f_{grassland}) \rightarrow E(f),$$
$$K_{forest}: E(f_{forest}) \rightarrow E(f).$$

Then the total conversion function K can be written as:

$$K(\varphi(\cdot)) = \begin{cases} K_{crop}(f_{crop}(\cdot)), \varphi = f_{crop}, \\ K_{grassland}(f_{grassland}(\cdot)), \varphi = f_{grassland}, \\ K_{forest}(f_{forest}(\cdot)), \varphi = f_{forest} \end{cases}$$

Given the fact that a set {agricultural lands, grasslands, forests} within the subject is a complete group of events, and that for each element to calculate the quantitative rate of degradation requires only part of the information $I(x, y)$, to reduce the computational complexity software implementation for each pixel (x, y) it is advisable to calculate only part of it, namely:

$$I(x,y) \mid \varphi(x,y) = \begin{cases} \{LAI_{real}, LAI_{perfect}\}, \varphi = f_{crop}, \\ \{NDVI\}_{time\_series}, \varphi = f_{grassland}, \\ d, \varphi = f_{forest}, \end{cases}$$

The total degradation rate can be calculated by the formula:

$$f(x,y) = K\big(I(x,y) \mid \varphi(x,y)\big),$$

and the general map of land degradation is actually a graph f(x, y) on the set X × Y, which corresponds to the area of interest.

3 EXPERIMENTAL RESULTS

3.1 Validation of Classification Map

The efficiency of land degradation assessment is depending on the accuracy of the classification map. That is why the first experiment has been done for validation of the land cover/land use classification map. The main classes on the classification map are maize, winter wheat, soybeans, sunflowers, winter oilseed rape, sugar beet, peas, man-made objects, forests, uncultivated land (grassland), swamps, water bodies, and open ground.

Estimates of the accuracy of the classification map were calculated using a validation dataset, which was not used to train the classifier and construct this classification map. The validation dataset contained 455 polygons with a total area of 5858.16 ha. User and producer accuracy for each class is shown in Table 1.

Table 1: Producer Accuracy (PA) and User Accuracy (UA) for each class at the classification map.

	Class	PA, %	UA, %
1	Artificial	95,9	74,0
2	Winter wheat	96,5	98,8
3	Winter rapeseed	98,4	98,8
4	Maize	98,0	86,2
5	Sugar beet	99,1	100,0
6	Sunflower	98,2	94,0
7	Soybean	72,7	97,0
8	Forest	99,7	99,2
9	Grassland	96,8	66,0
10	Bareland	61,4	100,0
11	Water	100,0	100,0
12	Wetland	83,7	100,0
13	Peas	100,0	99,8
	Overall accuracy, %	94,2	
	Kappa	0,93	

3.2 Productivity Map Based on NDVI

The traditional remote sensing approach to degradation assessment is based on the dynamic of vegetation index NDVI, extracted from satellite imagery. During the last 5 years, the best data source for NDVI calculation is Sentinel 2 imagery, because of its high resolution (10 meters) and good coverage (each point on the Earth's surface is visited every 6-12 days). In our study, we use this data source, but only for nonagricultural lands. The lands, where NDVI has negative trend over the years we consider as degraded. Figure 1 demonstrates the productivity map based on NDVI dynamics for Ukraine for 2021.

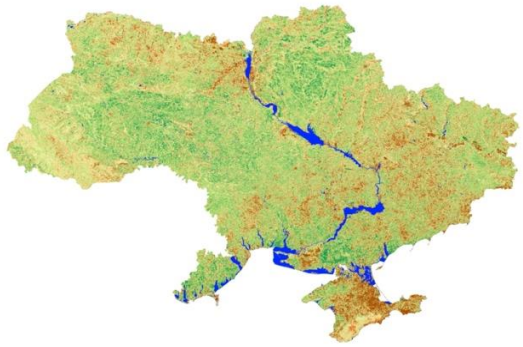


Figure 1: Productivity map based on vegetation index.

3.3 Land Degradation Maps

Figure 2 shows land degradation for the 2020 year. Red color depicts degraded lands, while green – is sustainable and productive land according to the complex methodology of land degradation assessment. As we can see, most of the territory of Ukraine stays sustainable. Most land degradation is observed on croplands due to ecology unfriendly agricultural practices.

Figure 2: Land degradation for Ukraine for 2021.

4 CONCLUSIONS

In this study, we have developed the geospatial method of land degradation assessment based on remote sensing data, neural networks, and bio-physical modelling. It takes into account different land cover/ land use classes and provides a specific way for land degradation assessment for each of them. Due to the high computation complexity of the method, it is reasonable to implement it in a cloud environment [10]. According to our study, most of the territory of Ukraine stays sustainable. Most land degradation is observed on croplands due to ecology unfriendly agricultural practices.

The developed technology is flexible and applicable to different climatic zones, because during biophysical modeling according to the WOFOST model it takes into account precipitation, temperature, as well as the main stages of crops growth - seedlings, maturation, maturity. Obtaining annual degradation maps according to the described methodology, it is possible to analyze changes for the better or worse, analyze degraded areas and their distribution, as well as make appropriate management decisions to prevent and regulate land quality in Ukraine.

ACKNOWLEDGMENTS

This research was partly funded by the National Research Foundation of Ukraine within the project 2020.02/0284 «Geospatial models and information technologies of satellite monitoring of smart city problems» and Horizon 2020 e-shape project (https://e-shape.eu/).

REFERENCES

[1] A. Lehmann, Y. Guigoz, N. Ray, E. Mancosu, K. C. Abbaspour, E. R. Freund, and G. Giuliani, "A web platform for landuse, climate, demography, hydrology and beach erosion in the Black Sea catchment", Scientific data, 2017, vol. 4 (1), pp.1-15.

[2] A. Kolotii, N. Kussul, A. Shelestov, S. Skakun, B. Yailymov, R. Basarab, and V. Ostapenko, "Comparison of biophysical and satellite predictors for wheat yield forecasting in Ukraine. International Archives of the Photogrammetry", Remote Sensing and Spatial Information Sciences - ISPRS Archives, 2015, vol. 40 (7W3), pp. 39-44.

[3] N. Kussul, et al., "Crop inventory at regional scale in Ukraine: developing in season and end of season crop maps with multi-temporal optical and SAR satellite imagery", European Journal of Remote Sensing 51.1, 2018, pp. 627-636.

[4] A. N. Kravchenko, N. N. Kussul, E. A. Lupian, V. P. Savorsky, L. Hluchy, and A. Y. Shelestov, "Water

resource quality monitoring using heterogeneous data and high-performance computations", Cybernetics and Systems Analysis, 2008, vol. 44(4), pp. 616-624.

[5] G. Giuliani, P. Mazzetti, M. Santoro, S. Nativi, J. Van Bemmelen, G. Colangeli, and A. Lehmann, "Knowledge generation using satellite earth observations to support sustainable development goals (SDG): A use case on Land degradation", International Journal of Applied Earth Observation and Geoinformation, 2020, no. 88, p. 102068.

[6] G. Giuliani, B. Chatenoux, A. Benvenuti, P. Lacroix, M. Santoro, P. Mazzetti, "Monitoring land degradation at national level using satellite Earth Observation time-series data to support SDG15–exploring the potential of data cube", Big Earth Data, 2020, vol. 4(1), pp. 3-22.

[7] M. Claverie, J. Ju, J. G. Masek, J. L.Dungan, E. F. Vermote, J. C. Roger, C. Justice, "The Harmonized Landsat and Sentinel-2 surface reflectance data set", Remote sensing of environment, 2018, no. 219, pp. 145-161.

[8] N. Kussul, A. Shelestov, M. Lavreniuk, I. Butko, S. Skakun, "Deep learning approach for large scale land cover mapping based on remote sensing data fusion", Geoscience and Remote Sensing Symposium (IGARSS), 2016, pp.198-201.

[9] N. Kussul, G. Lemoine, J. Gallego, S. Skakun, and M. Lavreniuk, "Parcel based classification for agricultural mapping and monitoring using multi-temporal satellite image sequences," 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2015, pp. 165-168.

[10] A. Shelestov, M. Lavreniuk, V. Vasiliev, L. Shumilo, A. Kolotii, B. Yailymov, H.Yailymova, "Cloud approach to automated crop classification using Sentinel-1 imagery", IEEE Transactions on Big Data, 2019, vol. 6(3), pp. 572-582.

.

Comparative Analysis of Methods of Forecasting the Consumer Price Index for Food Products (on the Example of the Altai Territory)

Stepan Mezhov and Maxim Krayushkin
Altai State University, 61 Lenin avenue, Barnaul, Russia
megoff@mail.ru, kramaks-97@mail.ru

Keywords: Consumer Price Index, Time Series, Neural Network, Decision Tree, Error Back Propagation, Gradient Boosting, Forecast.

Abstract: At the moment, there are no uniform universal methods for forecasting regional indicators of economic development in general and the consumer price index in particular. But depending on how accurate and reasonable the forecasts of the consumer price index will be, the budget of the region will be drawn up so correctly and the parameters of the forecast of socio-economic development, in the calculation of which this indicator is used, will be accurately predicted. The article presents a comparative analysis of methods for forecasting the consumer price index for food products. First, the most popular methods of forecasting the consumer price index were identified. Then models of time series, neural networks and decision trees were built, as well as retro-forecasts of the consumer price index for food products based on them. It is revealed that neural networks provide higher accuracy of forecasts compared to other models. The result of the work was the forecast of the consumer price index for food products in the Altai Territory for 2021 based on the constructed neural network model. The constructed neural network models can be used in relevant organizations to increase the accuracy of the forecast of this indicator. In addition, such an approach can be used as a basis for forecasting other indicators that characterize the socio-economic development of regions.

1 INTRODUCTION

Forecasts of the socio-economic development of the region, including the forecast of the consumer price index (hereinafter also – CPI), are sent to the Ministry of Finance of the region and to the Ministry of Economic Development of the Russian Federation. Using these data, the Ministry of Finance of the region develops the main parameters of the regional budget, and the Ministry of Economic Development of the Russian Federation clarifies the forecast of socio-economic development of the Russian Federation and monitors the socio-economic development of the region.

At the moment, there are no uniform universal methods for forecasting regional indicators of economic development in general and the consumer price index in particular.

The following are identified as the most popular methods used to predict the consumer price index:

- Based on time series analysis.
- Based on artificial neural networks.

This conclusion was made based on a meaningful analysis of 30 scientific papers.

Also, the construction of forecasts of the consumer price index based on decision trees is currently gaining popularity. In 4 out of 30 cases they were used [1, 2, 3, 4], and in 6 cases other methods: [5, 6, 7, 8, 9, 10].

With regard to other methods, it should be noted that such as regression analysis, factor analysis, and a method based on the construction of a system of balanced indicators were used to predict this indicator.

As factors for the construction of CPI models, the authors usually used such as: the price index of manufacturers of industrial products, retail trade turnover, the volume of paid services to the population, the dollar exchange rate, the price index for agricultural products and others.

As specific information technologies for building models, the authors used both the programming languages: R, Python and the like, as well as modern software packages for statistical data analysis, for example, the Statistica software package.

2 CONSTRUCTION AND COMPARATIVE ANALYSIS OF FORECAST CPI MODELS FOR FOOD PRODUCTS IN THE ALTAI TERRITORY

2.1 Construction of a Forecast Model of the CPI for Food Products in the Altai Territory Based on Time Series Analysis

As it has already been revealed, the method based on time series analysis is very popular in forecasting the consumer price index. To build such models, a representative sample was first determined. It is believed that it is advisable to train models with data from 2017, since similar economic conditions have been formed since that time.

That is, first, data on the CPI for food products in the Altai Territory from January 2017 to December 2020 were taken. Then the analysis of the time series was carried out.

The analysis was carried out based on the expert analysis of a number of data, and then using the construction and analysis of the correlogram.

Based on this analysis and the specifics of this group of products, it was determined that a number of data have all deterministic components (trend and seasonality). It is believed that if they are available, it is advisable to build a Holt-Winters model [11].

It should be noted that to build a forecast using a time series model for one year, it usually takes at least 2 years to train it [11].

We consider it appropriate to train the model first with data from 2017 to 2018 and test it on 2019 data, and then train it with data from 2018 to 2019 and test it on 2020 data.

Schematically, the Holt-Winters model can be shown as follows (1):

$$IPC_t = A + B * t + S_i, \quad (1)$$

where: IPC_t – forecast value of CPI for food products, A – constant, B – angular coefficient, t – forecast period, S_i – seasonality coefficient.

It should be noted that when constructing the Holt-Winters model, special coefficients characterizing the sensitivity of a data series to components are usually also estimated.

Initially, the Holt-Winters model of the CPI for food products was built according to data from 2017 to 2018. Her work on training and test data is presented below (Figure 1).

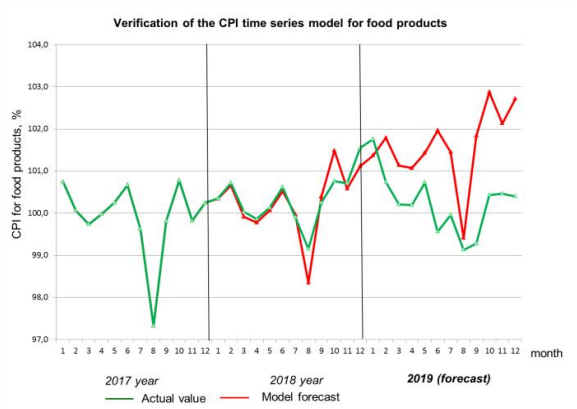


Figure 1: Operation of the CPI time series model for food products trained from 2017 to 2018.

The value of the middle absolute error (hereinafter also referred to as MAE) of the model on the training data is 0,23 percentage points (hereinafter also referred to as p. p.), which is unacceptable. It should be noted that for 8 months of 2019, the forecast for the MAY model was 1,73 p. p.

The choice in favor of the MAE indicator was made due to the fact that in order to solve our task, it is important that the deviation of real data from the predicted values obtained by the model on the training data was on average no more than 0,05 percentage points, and on the forecast average no more than 0,35 percentage points. Otherwise, in our case, it is impractical to use this model in the future for forecasting.

After that, a model of time series of CPI for food products was built according to data from 2018 to 2019. The work on training and test data is shown in Figure 2.

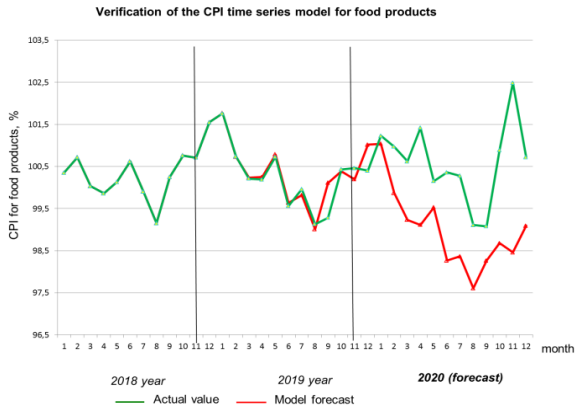


Figure 2: Operation of the CPI time series model for food products trained from 2018 to 2019.

The value of the average absolute error (MAE) of the model on the training data is 0,19 percentage points, which is unacceptable. Note that for 6 months of 2020, the forecast for the model MAY was 2,02 percentage points.

Thus, in our case, this model is impractical to use in forecasting.

2.2 Construction of a Forecast Model of the CPI for Food Products in the Altai Territory Based on Neural Networks

As already noted, it is advisable to train a neural network model on a data set starting in 2017.

Based on the analysis of Pearson pair correlation coefficients and expert analysis of possible factors affecting the CPI for food products, it was found that it would be advisable to take the CPI for food products as factors for training the neural network model: The CPI of the Russian Federation for food products, the producer price index for agricultural products sold in the Altai Territory, the producer price index for industrial goods according to the type of economic activity «Food Production» in the Russian Federation. This is due to the consideration of real economic processes that provide the necessary correlation.

The training of neural networks with a teacher was chosen as a paradigm, the training rule was error correction, the architecture was a multilayer neural network, the learning algorithm was reverse propagation. The choice was made during the analysis of various learning algorithms for solving the forecasting problem.

The Deductor Studio software environment was used to build neural networks. This is due to a number of factors: the simplicity of building a model, the possibility of additional training and a user-friendly interface for a user who does not have high qualifications [12].

When using the learning algorithm with a teacher, the weighting coefficients of the neural network are adjusted in such a way as to minimize deviations of the predicted values from the values of the test sample [13].

Of course, within the framework of solving our task, it is important that the modal value of the CPI be as close as possible to statistical, that is, the model error should be small enough.

Numerous publications on industrial applications of multilayer networks with a learning algorithm by the method of back propagation of errors have confirmed its fundamental operability in practice [14].

It should be noted that it is important when building models of neural networks that it is necessary to achieve the absence of the effect of retraining. After all, then the model works well on training data, but the forecast when using this model turns out to be inaccurate. Therefore, it is important to correctly approach the selection of network parameters. That is, it is necessary to find the optimal number of hidden layers, the activation function, the optimal number of training epochs so that the neural network is not retrained.

The scheme of operation of this algorithm is presented below (Figure 3).

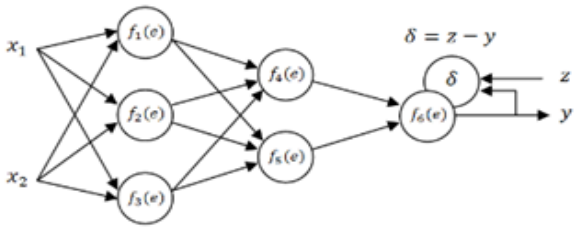


Figure 3: Diagram of the operation of the error back propagation algorithm.

When using this algorithm, the weighting coefficients are found at the beginning of the first epoch of training. Then they are adjusted from the output to the inputs, and not vice versa, in order to reduce the error of training and testing.

Neural networks have been trained for epochs. At first, 10,000 epochs were taken. Then it is determined on which of the epochs the error on the test sample is minimal. This served as a criterion for stopping learning.

A two-year sample for training a neural network with a prediction period of up to 1 year is quite enough.

Note that the constructed neural networks have 3 hidden layers of 8 neurons in each of them, and the activation function of neurons is sigmoidal.

The graph of neural networks can be represented as follows (Figure 4).

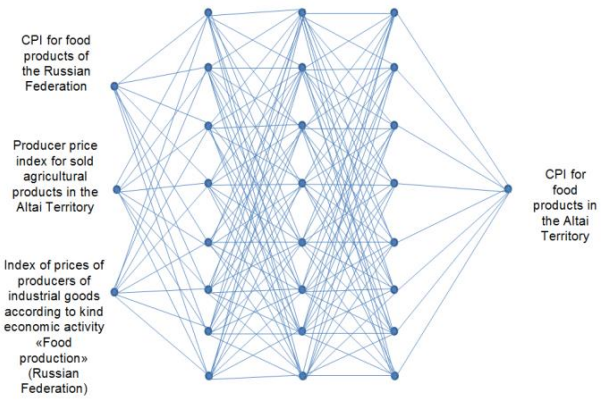


Figure 4: Graph of constructed neural networks of CPI for food products.

It should be noted that in our case, three hidden layers of 8 neurons in each of them are enough. This conclusion is made because it is believed that the number of hidden layers should not exceed the number of network inputs, and the number of neurons should not exceed the number of observations.

First, a neural network of CPI for food products was built according to data from 2017 to 2018. The results of her work on training and test data are shown in Figure 5.

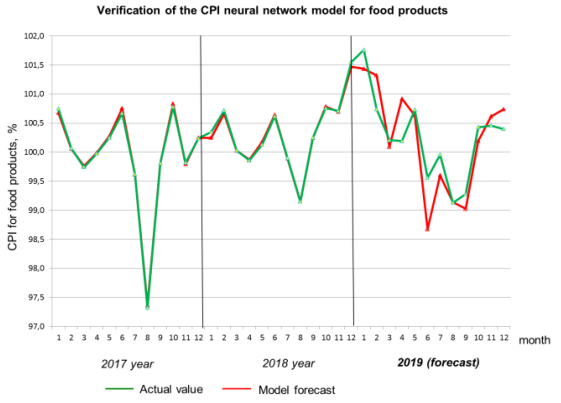


Figure 5: Verification of a neural network for predicting CPI for food products, trained on 2017-2018 data.

The value of the average absolute error of the model (MAE) was 0,03 percentage points, which is generally acceptable. Note that the forecast error obtained using the neural network model is 0,29 p. p.

Then the CPI neural network for food products was built according to data from 2018 to 2019. The results of her work on training and test data are shown in Figure 6.

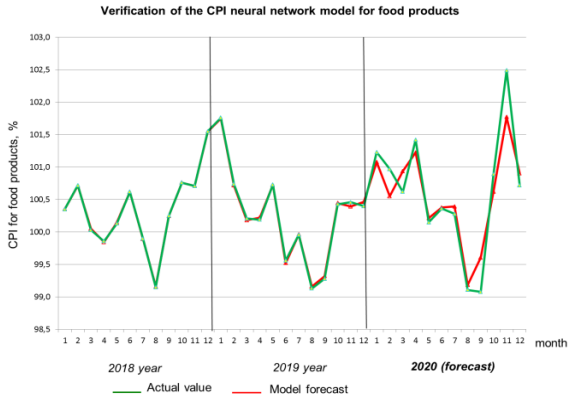


Figure 6: Verification of a neural network for predicting CPI for food products, trained on 2018-2019 data.

The value of the average absolute error of the model (MAE) in this case was 0,02 percentage points, which is also acceptable. The forecast error for 6 months of 2020 based on the neural network model is 0,31 percentage points.

These two examples confirm the applicability of neural network models in predicting CPI, which significantly increase the accuracy of the forecast.

2.3 Forecasting the Consumer Price Index for Food Products in the Altai Territory Based on Decision Trees

The next fairly popular approach to constructing forecasts of the consumer price index is the approach based on the construction of decision trees.

The process of building a decision tree can be represented as the following steps:

- Definition of input and output parameters.
- Building a tree.
- Evaluation of the quality of the decision tree.

We will use the same input and output parameters as for training neural networks.

At the second stage, we will build the tree itself. The ported version of the Statistica software package is used, because of the possibility of a fairly fast and convenient construction of a decision tree in it by the «Gradient Boosting» method, which is widely used in forecasting.

The idea of «Gradient Boosting» is an iterative process of sequential tree construction. The new tree is trained using information about errors made at the previous stage. This technique uses the idea that the next model will learn from the mistakes of the previous one [15].

Models have an unequal probability of appearing in subsequent models, and those that give the greatest error will appear more often. That is, many trees are being built, each of which can be more accurate. The process ends when the required accuracy is reached or the accuracy begins to fall due to retraining [15].

In the Statistica software package, it is possible to build different decision trees using different methods. But it is for forecasting that it is better to use the «Gradient Boosting» method when constructing them.

A lot of trees are being built in this package. As a result of «Gradient Boosting», the most accurate tree is determined.

In addition, the second most popular method for predicting using decision trees is the «Random Forest».

The essence of this method is to use an ensemble of decision trees. By itself, the decision tree provides an extremely low quality of classification, but due to the large number of them, the result is significantly improved.

Schematically, a fragment of the decision tree model can be represented as follows (Figure 7).

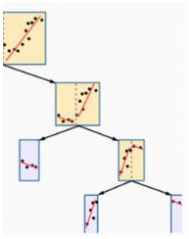


Figure 7: A fragment of the decision tree.

Initially, a decision tree of the CPI for food products was built according to data from 2017 to 2018 (Figure 8).

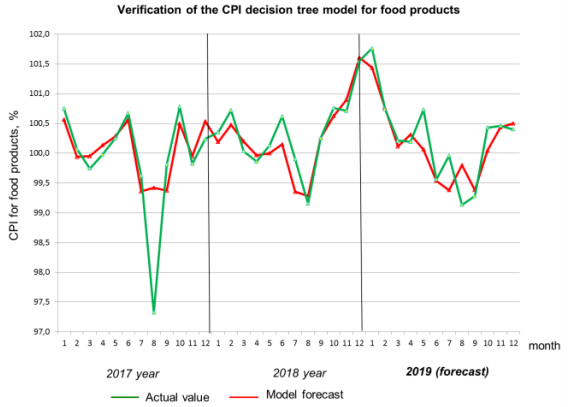


Figure 8: The work of the CPI decision tree for food products trained from 2017 to 2018.

The value of the average absolute error (MAE) of the model on the training data is 0,28 p. p. Note that for 8 months of 2019, the MAE forecast for the model was 0,32 p. p.

After that, a decision tree of the CPI for food products was built according to data from 2018 to 2019 (Figure 9).

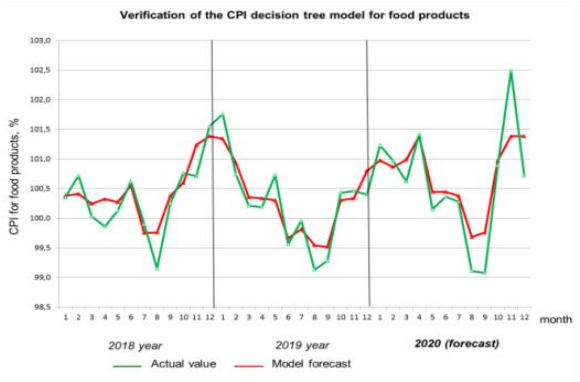


Figure 9: The work of the CPI decision tree for food products trained from 2018 to 2019.

The value of the average absolute error (MAE) of the model on the training data is 0,24 p. p. Note that for 6 months of 2020, the value of the average absolute error of the forecast for the model is 0,53 p. p.

2.4 Comparative Analysis of Forecast Models and the Construction of a Forecast of the CPI for Food Products in the Altai Territory for 2021

In this section, the forecast of the CPI for food products in the Altai Territory was built on the basis of an optimal model. Initially, the optimal model was selected based on the absolute verification of retro-forecasts (Table 1).

Table 1: The fact of absolute verification of retro-forecasts.

Training period	Forecast period	MAE retro-forecast, p. p.		
		Time Series model	Neural network model	Decision Tree model
from 2017 to 2018	May-December 2019	1,73	0,29	0,32
from 2018 to 2019	July-December 2020	2,02	0,31	0,53

Table 1 shows that neural network models are the most accurate.

Then it was on their basis that the forecast of the CPI for food products in the Altai Territory was built. The forecast value of December 2021 by December 2020 was 109,3 percent.

3 CONCLUSION

Forecasting the consumer price index is of great importance for regional development. The correctness of the formation of the budget of the region will depend on how accurate and reasonable the forecast of this indicator is.

During the analysis of scientific papers devoted to the prediction of the CPI, it was revealed that the most common methods of forecasting this indicator are an approach based on time series analysis, forecasting based on the construction of artificial neural networks and decision trees.

Then models of time series, neural networks, and CPI decision trees for food products were built. It should be noted that neural network models were chosen as the best, on the basis of which the forecast

of the CPI for food products for 2021 was built. When constructing econometric models, programs such as R-Studio, Deductor Studio, Statistica were used.

REFERENCES

[1] V. A. Baykov and A.Y. Tarasova, "An integrated approach to forecasting the inflation rate", Bulletin of the Moscow University of Finance and Law, 2019, no. 7, p. 49-57.

[2] N.N. Karabutov, "Interrelation of indices determining the level of inflation", Economy, Taxes, Law, 2020, no. 10, pp. 40-47.

[3] A. A. Skrobotov and A.V. Tsarev, "Forecasting of macroeconomic indicators of the Russian economy", Economic development of Russia, 2020, no. 18, pp. 45-55.

[4] S. G. Shulgin, "Selection of variables for instability analysis and forecasting using gradient boosting models", System monitoring of global and regional risks, 2018, no.8, pp. 115-153.

[5] E. V. Balatsky, N. A. Ekimova, and M. A. Yurevich, "Short-term forecasting of inflation based on marker models", Forecasting problems, 2019, no. 5, p. 28-40.

[6] I. A. Vakhrushev, "Forecasting the trend dynamics of the stock market based on macroeconomic factors using a diffuse index", Scientific Journal of ITMO Research University, The series «Economics and Environmental Management», 2020, no. 5, pp. 42-48.

[7] F. E. Huseynova, "Analysis of medium-term trends in the dynamics of inflationary processes of the Russian economy", Skif. Questions of student science, 2019, no. 7, p. 12-20.

[8] I. N. Dementieva, "Application of the index method in studies of consumer sentiment of the population", Economic and social changes, 2019, no. 1, pp. 153-173.

[9] E. B. Mitsek and S. A. Mitsek, "Analysis of factors of dynamics of the main macroeconomic variables of the Russian Federation", Questions of management, 2020, no. 1, pp. 47-61.

[10] A. Yu. Yakimchuk, A. I. Teplenko, and M. N. Konyagina, "The influence of the key rate on inflation rates in modern Russia", Bulletin of the Academy of Knowledge, 2020, no. 37, pp. 490-497.

[11] I. N. Dubina, Mathematical and statistical methods in empirical socio-economic research: textbook, Moscow, Finance and Statistics, 2010, 415 p.

[12] Deductor - platform capabilities. Official website of BaseGroup Labs, 2021, [Online]. Available: https://basegroup.ru/deductor/description.

[13] F. Wasserman, "Neurocomputer technology", Theory and practice: a textbook, Moscow, 1992.

[14] S. A. Terekhov, Lectures on the theory and applications of artificial neural networks: textbook, - Snezhinsk, 1998.

[15] Database of examples of solving specific management tasks in the statistica system. Official website of Statsoft, 2021, [Online]. Available: http://statsoft.ru/solutions/ExamplesBase/tasks.

Advanced Method of Land Cover Classification Based on High Spatial Resolution Data and Convolutional Neural Network

Andrii Shelestov[1], Bohdan Yailymov[2], Hanna Yailymova[1], Leonid Shumilo[3], Mykola Lavreniuk[2], Alla Lavreniuk[1], Sergiy Sylantyev[2] and Nataliia Kussul[1]
[1]*Institute of Physics and Technology, NTUU "Igor Sikorsky Kyiv Polytechnic Institute", 37 Peremohy avenue, Kyiv, Ukraine*
[2]*Department of Space Information Technologies and System, Space Research Institute NAS Ukraine an SSA Ukraine, 40 Glushkov avenue, Kyiv, Ukraine*
[3]*Department of Geography, The University of Maryland, College Park, Maryland, USA*
{andrii.shelestov, yailymov, anna.yailymova, shumilo.leonid, alla.lavrenyuk, sylantyev, nataliia.kussul}@gmail.com, nick_93@ukr.ne

Keywords: Convolution Neural Network, Probability Classification, Land Cover Map, Urban Atlas, Smart City.

Abstract: Based on modern satellite products Planet with high spatial resolution 3 meters, authors of this paper improved the neural network methodology for constructing land cover classification maps based on satellite data of high spatial resolution using the latest architectures of convolutional neural networks. The process of information features formation for types of land cover is described and the method of land cover type classification on the basis of satellite data of high spatial resolution is improved. A method for filtering artificial objects and other types of land cover using a probabilistic channel is proposed, and a convolutional neural network architecture to classify high-resolution spatial satellite data is developed. The problem of building density maps for the quarters of the city atlas construction is solved and the metrics for estimating the accuracy of classification map construction methods are analyzed. This will make it possible to obtain high-precision building maps to calculate the building area by functional segments of the Urban Atlas and monitor the development of the city in time. This will make it possible to create the first geospatial analogue of the product Copernicus Urban Atlas for Kyiv using high spatial resolution data. This Urban Atlas will be the first such product in Ukraine, which can be further extended to other cities in Ukraine. As a further development, the authors plan to create a methodology for combining satellite and in-situ air quality monitoring data in the city based on the developed Urban Atlas, which will provide high-precision layers of PM_{10} and $PM_{2.5}$ concentrations with high spatial and temporal resolution of Ukraine.

1 INTRODUCTION

The image of remote sensing data is a matrix of "pixels", which are the smallest unit of the image and contain the values of the measured spectral reflectance. Traditionally, pixel-based approaches are used to classify images, as a result of which each pixel of the image is mapped to a class according to its spectral properties [1]. Object characteristics include spectral characteristics, shape, size, texture, and context properties. Signs usually determine the upper and lower limits of the range of measured characteristics of objects. Image objects within certain contours belong to a certain class, and those outside them belong to other classes [2]. An important task is to identify the most relevant features for each class in order to effectively perform the classification of images with high accuracy [3]. The following features are most often used in the literature for the classification of urban areas: brightness, average value of the reflection spectrum and standard deviation of the spectra [4]. According to the average value and the standard deviation of the features, the threshold values of the selected features are determined, which are used to assign a class label to each pixel [32]. To understand the relevance of the research problem and answer key issues that researchers face when classifying land cover using high-resolution data, we analyzed existing approaches related to the pixel and object-oriented approach using convolutional neural networks

(CNN), which gave us the opportunity to improve our own algorithm for classifying urban areas on Sentinel-2 satellite data with a spatial resolution of 10 meters, as well as to develop a new algorithm for classifying satellite data with a high spatial resolution (for example, Planet data with a spatial resolution of 3 meters).

Within the object-oriented approach, the smallest spatial unit for mapping is geospatial objects, which look like vector polygons extracted from high-resolution remote sensing images using segmentation techniques. These internally homogeneous geographical objects are used as the basic units of the classification map. Conditionally multi-scale segmentation is used to select homogeneous image objects and create appropriate vector polygons [5].

The purpose of this article is to consider the main problems of monitoring and classification of land cover in Kyiv on the basis of our proposed neural network algorithm, which provides an opportunity to update annually for city's land cover maps with high spatial resolution, which is crucial information in management decisions when planning the cities development in the long run.

In recent years, segmentation methods have been proposed with further classification based on CNN to extract the boundary information of objects using self-learning functions from images [6]. A study by Dong, Wu, Luo, Sun, and Xia (2019) proposed a method for determining geospatial objects based on CNN [7]. The method consists of the following stages.

Initially, road and river landfills on the historical map of the earth's crust are used to zone the target image into several subregions. Subsequent object classification of subregions can be performed in parallel mode. In the second stage, a probability map is built for each subregion, using a modified convolutional network VGG16 [8]. This network uses five convolution layers, the output of which is combined by the upsampling method [7]. In the third stage, a vector probability map of object boundaries is built. The subregion images are then extracted from the landfill boundary of each geospatial object. The results of all subregions are combined to form a structured map of geospatial vector classification.

After determining the shape of geospatial objects for each of them the procedure of extraction of features is performed. Traditionally, each class of geospatial objects has a set of features that are determined on the basis of satellite data. Let's focus on the following image characteristics: spectrum, shape and texture. The spectral characteristics the average pixel brightness for the selected object, the standard deviation and maximum differences of the spectral signals, the normalized differential vegetation index (NDVI) and the normalized differential water index (NDWI) are usually used. These indices are calculated based on satellite data with high spatial resolution (one value per pixel), and the average values of the corresponding pixels that make up the geospatial object are used as features of the object itself. Indicators of the shape of geospatial objects are the size of the object, the ratio of length to width, pixel index of shape and the overall index of shape [9]. The textural features of geospatial objects use the index PanTex of the presence of buildings [10], and the degree of matrices of the gray-level co-occurrence matrices (GLGM) [11]. Based on the data of the digital terrain model, you can also calculate the terrain characteristics, including the average topographic height and steepness for each pixel.

The corresponding features of the object are calculated as the average values of the pixels that are part of each geospatial object. In Table 1 lists the features of the classes used in deep learning algorithms are presented.

Table 1: The List of Features Used in the Classification by Deep Learning.

Spectrum characteristics	Texture features	Topographic features
The average value of the spectrum signals in different satellite channels	Homogeneity	Height
Standard deviation of spectral signals in different satellite channels	Contrast	Incline
Brightness of spectral signals	Unlikeness	
Maximum differences of spectral signals	Entropy	
Normalized vegetation index (NDVI)	Correlation	
Normalized water index (NDWI)		

In Figure 1 presents the NDVI values for selected types of land cover and different geospatial objects. You can use additional geospatial data from different sources to obtain additional informational features.

Different land cover classification methods based on high spatial resolution data can be very useful for different land classification problems under different urban agglomerations, making it possible to obtain an accurate geospatial vector urban atlas like the Copernicus Urban Atlas. The use of satellite data with a spatial resolution of 10 meters does not make it possible to accurately identify artificial objects in suburban areas.

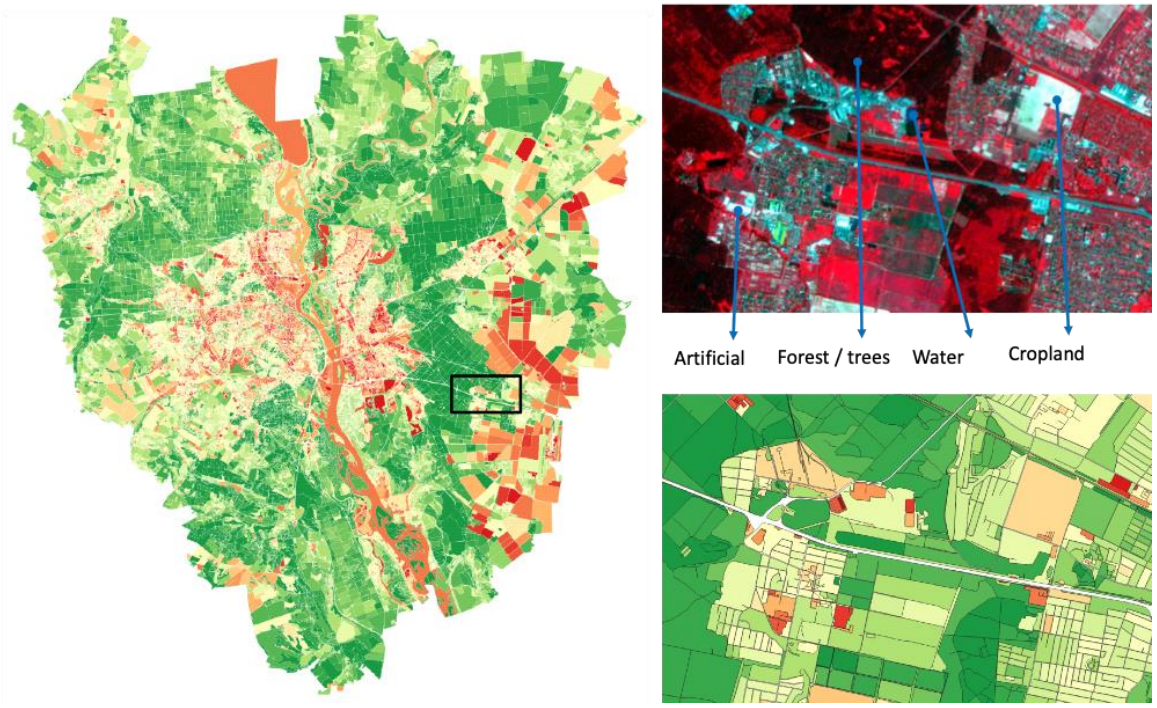


Figure 1: Examples of NDVI values for selected types of land cover and various geospatial objects of the Kyiv city and its environs based on Planet data.

In recent years, the Object-based image land cover classification has attracted considerable attention based on objects with high spatial resolution data and a convolutional neural network for mapping the earth's surface using remote sensing images. Numerous studies over the past decade have examined a wide range of issues related to land mapping [1] – [8]. However, most of these studies focused on agricultural land and crop types. In our case, it is important to accurately identify artificial objects, the identification of which on the satellite data of high spatial resolution may cause problems with shadows from large buildings, which belong by the neural network to the incorrect class. This is our research - to improve existing and develop our own neural network algorithm for the classification of urban areas.

2 METHOD

2.1 Study Area

Mapping urban land use and land cover is a fundamental task of urban planning and management. Very high resolution (VHR) satellite images, such as images from QuickBird, IKONOS, GeoEye, WorldView-2/3/4 and GaoFen-2, have shown a great advantage in urban land monitoring conditions due to high spatial detailing, compared to free data. In recent years, many studies have been conducted in the world on the classification of urban land use and soil cover based on VHR images [12], [34], [35], [37]. High-resolution classification methods can generally be divided into two classes - pixel-based methods and object-based methods [13]. The first defines the classes for individual pixels mainly on the basis of spectral information. In VHR images, pixel-by-pixel methods can be problematic because the spectral values of individual pixels on the earth's surface do not reflect the characteristics of the object. Object-oriented methods have been proposed to solve this problem. Unlike pixel-by-pixel methods, object-oriented methods combine adjacent pixels into objects using image segmentation, such as Multi-Resolution, Mean-Shift or Quadtree-Seg, and the objects are considered as units of classification [14]. Object-oriented classification methods greatly help to reduce problems with pixel-by-pixel approaches, while the quality of image segmentation significantly affects the accuracy of classification [13]. In addition, analysts must first randomly select a set of object features or design representative features as input to

the classification using design techniques methods. The effectiveness of the selected features also affects the accuracy of the classification. Objects selected in a complex urban area are unlikely to be representative of all types of land cover. To solve this problem, it is necessary to provide automatic selection of features on images from remote sensing satellites instead of using manually selected features.

2.2 Classification Procedures

Artificial intelligence methods for image recognition and computer vision, such as K-means, neural networks (NN), support vector machines (SVMs), and random forests (RF), are not effective ways to classify Earth remote sensing images. A study by Fu, Zhao, Li, and Shi (2013) proposed for those purposes the idea of deep learning [14]. Compared to traditional machine learning methods such as NN, SVM and RF, deep learning models provide automatic feature extraction from large data sets. They learn to remove key features in the modeling process; thus, do not require prior selection of signs. In computer vision communities, deep learning, including convolutional neural networks (CNNs), has been used successfully to categorize images, identify goals, and understand scenes [15]. As a rule, typical CNNs include convolutional and aggregation layers, a nonlinearity activation function followed by fully bound layers as classifiers. CNN's first application of remote sensing was the allocation of road networks and buildings by Mnih and Hinton (2010) [16]. CNN has recently been used for pixel-by-pixel semantic labeling (or classification) of high-resolution remote sensing images. Mnih and Hinton (2010) proposed the CNN architecture for aerial imaging. Wang et al. (2015) used a three-layer CNN structure and a finite state machine to identify the road network [17]. Hu et al. (2015) used a pre-trained CNN network to classify different scenes in high-resolution remote sensing images [18]. Längkvist and others (2016) used CNN for the pixel-by-pixel classification of 0.5 m aerial photographs in natural colors into five classes of land cover, including vegetation, land, roads, buildings, and water, using the Digital Surface Model (DSM) [19]. Maltezos (2016) used CNN to identify buildings based on ortho-photo images using object height information [20]. Pan and Zhao (2017) proposed an extended CNN model for land cover classification based on 4-channel GaoFen-2 satellite images of rural areas [21]. Long at al. (2015) proposed a fully convolutional neural network (FCN) [22]. In FCN, fully connected layers in CNN are replaced by layers with upward convolution and merged with a shallow layer. Because the standard CNN model is based on the "image label" principle, the FCN's "start-to-end" labeling mode is more suitable for pixel-based image classification, and assigning a specific class to each pixel. The structure of the FCN has also shown great potential in the classification of remote sensing images. Sherrah (2016) proposed the FCN structure without decreasing sampling for semantic marking based on reference data from the International Society for Photogrammetry and Remote Sensing (ISPRS) in Vaihingen and Potsdam, which are publicly available aerial photographs in conventional colors with spatial resolution 9 cm and accompanied by DSMs derived from Lidar [23]. Based on the same data sets, implemented a multi-scale FCN [24]. The ensemble FCN model is offered [25]. Maggiori at al. (2017) compared CNN and FCN models for aerial photography and presented a multilayer perceptron (MLP) structure based on the FCN model for large-scale aerial photography classification [26]. Our review of the literature shows that most of the existing studies, with exception of [24], were conducted on basis of test sets ISPRS. Although these publicly available datasets provide images and reference sources for more efficient modeling and model comparisons, lower-spatial satellite imagery is often used for operational land classification and land use in large urban areas, and accurate DSMs are not always available. The U-Net model proposed in [27] is an improved FCN model, characterized by symmetrical U-shaped architecture consisting of an encoder and a decoder. This model combines low-level features with detailed spatial information with high-level features with semantic information to improve segmentation accuracy and achieves promising results in segmentation problems of individual classes, such as biomedical image segmentation, aerial road network definition, building identification and sea and land segmentation in Google Earth images. For multi-class classification tasks such as land cover classification, contextual information at different scales is important, as the characteristics of different types of land cover or terrestrial objects usually have different scales [27]. However, this information is not included in the original form of the U-Net model. Convolution is an important step in the CNN and FCN models because it allows models to distinguish features of different scales and degrees of abstraction.

2.3 NN Architecture

The initial form of the U-Net architecture consists of a compression path and a mirror expansion path. The convolution path determines high-level features through convolution and merges operations, while the spatial resolution of object maps is reduced. The expansion path (decoder) attempts to restore the resolution of object maps using up convolution operations. For each level of the compression path, feature maps are transmitted to a similar level in the decoder, which allows you to distribute contextual information over the network. For our experiments, we use one of the most accurate deep learning architectures for semantic segmentation tasks, the U-Net model. The architecture of our U-Net model traditionally consists of convolutional and de-convolutional parts, which are interconnected by a concatenation operation.

Traditionally, cross-entropy (CE) or weighted cross-entropy is chosen as a loss function (1) for multilayer perceptron training and deep learning models [5]:

$$CE = -\sum_{k=1}^{N}\sum_{c=1}^{C} \alpha_c y_c \log(p_c), \quad (1)$$

where C is the number of classes, N is the number of elements in the sample, y is the target vector, in our case we choose one-hot coding, p is the output of the last layer of the neural network, α_c is the coefficient to control class influence, $\alpha_c = \frac{N_c}{N}$.

Batch normalization is applied before each Relu activation function. Batch normalization scales the input for layers, typically mini-packets, using mean value and variance. This scaling eliminates the internal covariance shift and thus speeds up the learning process [28]. A 2 × 2 max merge operation is used between the top and bottom layers on the compression path, which reduces the resolution of object maps. The size of the objects in the bottom layer of the compression contour is reduced to 1/64 of the original image. The bottom layer of the standard U-Net model corresponds to the same structure of the bottom layer, that is it includes two consecutive convolutions, batch normalization, and Relu operation.

2.4 Filtration Method

Given the high segmentation of the terrestrial cover map obtained as a result of the classification of high-resolution satellite data, it is necessary to improve the quality of the classification map by post-processing, namely the filtering of the resulting map. To do this, it is proposed to use an additional channel, which contains the recognition probabilities of each pixel according to each type of land cover.

The neural network used to construct the terrestrial cover map uses the Softmax layer to convert objects to the probability that a pixel belongs to each type of land cover. This function reduces the K-dimensional vector z with arbitrary values of the components to the K-dimensional vector σ (z) with the actual values of the components on the interval [0, 1] giving a sum of one. The function is set as follows (2), (3):

$$\sigma: \mathbb{R}^K \to [0,1]^K, \quad (2)$$

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}. \quad (3)$$

The probability that an object x_i with a class y_i will occur in the sample is (4):

$$b(x_i)^{[y_i=+1]}(1 - b(x_i)^{[y_i=-1]}). \quad (4)$$

Therefore, we will record the plausibility of the sample (the probability of obtaining such a sample in terms of algorithm) (5):

$$Q(a, X) = \prod_{i=1}^{l} b(x_i)^{[y_i=+1]}(1 - b(x_i)^{[y_i=-1]}), \quad (5)$$

where X is the space of objects x, $y_i = +1$ is the identifier that object x belongs to class +1.

This plausibility can be used as a functional for learning the algorithm, with an amendment that is more convenient to optimize its logarithm (6):

$$-\sum_{i=1}^{l}([y_i = +1]\log b(x_i) + [y_i = -1]\log(1 - b(x_i))) \to min. \quad (6)$$

This loss function allows you correctly to predict the probabilities. Due to the need to filter the resulting mask of artificial objects from unreliable areas that erroneously fell into this mask due to classification, the method of threshold filtering (according to the principle of threshold classification [29]) was used with a threshold of 80% probability channel for artificial class objects.

2.5 Accuracy Assessment

Due to the limited resolution of satellite images, the density of buildings within quarters on Sentinel classification maps will be higher than they really are. That is why within the research framework, Planet satellite data with a spatial resolution of 3 meters are used to build a land cover classification map for the city of Kyiv. However, even in the absence of such data, for the construction of an urban atlas, the accuracy of classification according to Sentinel images is more than 90% [29], [30].

The condition for the creation of the Urban Atlas is the need to determine the percentage of buildings in each of the city's neighborhoods. This uses a map of the land cover and a vector layer with quarters of the city, built on the principle described above. To determine the percentage of the area of a certain type of land use (n - land cover class number) for each quarter k we will use the (7):

$$\forall k = \overline{1, QN}, n = \overline{1, CN}: \; P_k^n = \frac{S_k^n}{S_k} \cdot 100\%, \qquad (7)$$

where QN - the number of quarters; CN - the number of earth cover classes on the map obtained as a result of neural network training; P^k_n - the ratio of the area of class n to the area of quarter k; S^n_k - the area of class n in quarter k; S_k - the area of quarter k. To determine the percentage of buildings, we use formula (7) for the class of land cover "artificial objects" in each of the quarters of the city of Kiev. According to the above formula, the percentages of building density by quarters of the city of Kyiv are calculated.

2.6 Results and Discussion

To build a map of the land cover for the Kyiv city, cloudless satellite images of Planet with a spatial resolution of 3 meters and 4 multispectral channels (Blue, Red, Green, and Near-Infrared), obtained on September 6, 2020, and June 11, 2020 (Figure 2), over urban areas of Kyiv were used. Python programming language with a set of libraries that work with geospatial objects, including GeoPandas, was used to analyze spectral channels and highlight geospatial features. For each of the types of land cover, based on the training data, raster images are created for each of the image channels of the Planet satellite.

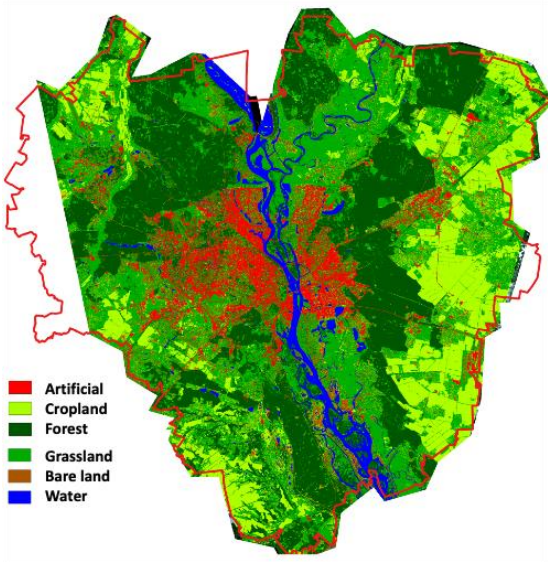


Figure 2: Classification map for the city of Kyiv and its environs according to Planet (2020).

Using the proposed neural network architecture [31], a map of the land cover classification for Kyiv and its environs was built according to Planet data with a spatial resolution of 3 meters (Figure 2). The overall accuracy of the obtained map is 95%, but there are some disadvantages due to the high spatial resolution of the data used.

3 CONCLUSIONS

The article proposes models for the classification of urban land cover by images with high spatial resolution. Like other supervised learning classification methods, the classification procedure consists of a learning phase and a testing phase. Unlike conventional learning strategies, which use randomly selected pixels or image objects as a training sample, the training patterns in our model are pairs of image fragments and their corresponding base chains, with each pixel marked with a specific class. Optimal sets of parameters are studied using inverse propagation and iterations [36]. In the classification phase, the trained models are executed on the input image to predict the class for each pixel.

The improved method of classification of land cover types on the basis of satellite data of high spatial resolution is described. The design process of information features for land cover types, the architecture of convolutional neural network for classification of satellite data of high spatial resolution is developed. The method of filtering artificial objects and other types of land cover using the probabilistic threshold method is proposed, and the problem of creating building density maps for city atlas quarters is solved.

Information features based on satellite data of high spatial resolution for different land cover types for land cover classification of the city of Kyiv are singled out and investigated. The architectures of neural networks used in the world for similar classification tasks are studied, the architecture of the neural network for segmentation and classification of satellite data of high spatial resolution for the city of Kyiv is developed. Using additional probabilistic information on the recognition of land cover classes, a method of filtering the obtained land cover classification map was developed. A method for estimating the density of the buildings within urban neighborhoods based on the obtained map of land cover classification has been developed. Metrics for

the accuracy estimation of the obtained results are determined. All these methods and metrics are implemented in the Python programming language.

ACKNOWLEDGMENTS

This research was funded by the National Research foundation of Ukraine within the project 2020.02/0284 «Geospatial models and information technologies of satellite monitoring of smart city problems», which won the competition "Leading and Young Scientists Research Support".

REFERENCES

[1] P. Casals-Carrasco, S. Kubo, and B. Babu, "Madhavan Application of Spectral Mixture Analysis for Terrain Evaluation Studies", Int. J. of Remote Sensing, 2000, vol. 21, pp. 3039-3055.

[2] R. Manandhar, I. Odeh, and T. Ancev, "Improving the accuracy of land use and land cover classification of landsat data using post-classification enhancement", Remote Sens, 2009, vol. 1, pp. 330-344.

[3] T. Blaschke, "Object based image analysis for remote sensing", ISPRS J. Photogramm, Remote Sens, 2010, vol. 65, pp. 2-16.

[4] D. C. Duro, S. E. Franklin, and M. G. Dubé, "A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery", Remote Sens, Environ, 2012, vol. 118, pp. 259-272.

[5] Z. Y. Lv, P. L. Zhang, and J. A. Benediktsson, "Automatic object-oriented, spectral-spatial feature extraction driven by Tobler's first law of geography for very high-resolution aerial imagery classification", Remote Sens, 2017, vol. 9, p. 285.

[6] D. Marmanis, K. Schindler, J. D. Wegner, S. Galliani, M. Datcu, and U. Stilla, "Classification with an edge: Improving semantic image segmentation with boundary detection", ISPRS J. Photogramm, Remote Sens, 2018, vol. 135, pp. 158-172.

[7] W. Dong, T.J. Wu, J.C. Luo, Y.W. Sun, and L. G. Xia, "Land-parcel-based Digital Soil Mapping of Soil Nutrient Properties in an Alluvial-diluvia Plain Agricultural Area in China", Geoderma, 2019, vol. 340, pp. 234-248.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7-9 May 2015, pp. 1-14.

[9] D. Lu and Q. Weng, "A survey of image classification methods and techniques for improving classification performance", Int. J. Remote Sens, 2007, vol. 28, pp. 823-870.

[10] M. Pesaresi, A. Gerhardinger, and F. Kayitakire, "A robust built-up area presence index by anisotropic rotation-invariant textural measure", IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens, 2008, vol. 1, pp.180-192.

[11] D. J. Marceau, P. J. Howarth, J. M. Dubois, and D. J. Gratton, "Evaluation of the grey-level co-occurrence matrix method for land-cover classification using SPOT imagery", IEEE Trans. Geosci, Remote Sens, 1990, vol. 28. pp. 513-519.

[12] E. Hussain and J. Shan, "Object-based urban land cover classification using rule inheritance over very high-resolution multisensor and multitemporal data", GISci, Remote Sens, 2016, vol. 53, pp. 164-182.

[13] D. Li, Y. Ke, H. Gong, and X. Li, "Object-based urban tree species classification using bi-temporal WorldView-2 and WorldView-3 images", Remote Sens, 2015, vol. 7. pp. 16917-16937.

[14] G. Fu, H. Zhao, C. Li, and L. Shi, "Segmentation for High-Resolution Optical Remote Sensing Imagery Using Improved Quadtree and Region Adjacency Graph Technique", Remote Sens, 2013, vol. 5, pp. 3259-3279.

[15] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, W. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding", In Proceedings of the IEEE conference on computer vision and pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016, pp. 3213-3223.

[16] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images", In Computer Vision—ECCV 2010, Lecture Notes in Computer Science; Daniilidis, K., Maragos, P., Paragios, N., Eds., Springer: Berlin/Heidelberg, Germany, 2010, vol. 6316, pp. 210-223.

[17] J. Wang, J. Song, M. Chen, and Z. Yang, "Road network extraction: A neural-dynamic framework based on deep learning and a finite state machine", Int. J, Remote Sens, 2015, vol. 36, pp. 3144-3169.

[18] F. Hu, G. S. Xia, J. Hu, and L. Zhang, "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery," Remote Sens, 2015, vol. 7, pp. 14680-14707.

[19] M. Längkvist, A. Kiselev, M. Alirezaie, and A. Loutfi, "Classification and segmentation of satellite orthoimagery using convolutional neural networks", Remote Sens, 2016, vol. 8, p. 329.

[20] E. Maltezos, "Deep convolutional neural networks for building extraction from orthoimages and dense image matching point clouds", J. Appl, Remote Sens, 2017, vol. 11, pp. 1-22.

[21] X. Pan and J. Zhao, "A central-point-enhanced convolutional neural network for high-resolution remote-sensing image classification", Int. J, Remote Sens, 2017, vol. 38, pp. 6554-6581.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 5-7 June 2015, pp. 3431-3440.

[23] J. Sherrah, "Fully convolutional networks for dense semantic labeling of high-resolution aerial imagery", arXiv, 2016, arXiv:1606.02585.

[24] N. Audebert, B. Le Saux, and S. Lefè vre, "Beyond RGB: Very high resolution urban remote sensing with multimodal deep networks", ISPRS J. Photogramm, Remote Sens, 2018, vol. 140, pp. 20-32.

[25] X. Sun, S. Shen, X. Lin, and Z. Hu, "Semantic Labeling of High Resolution Aerial Images Using an Ensemble of Fully Convolutional Networks", J. Appl, Remote Sens, 2017, vol. 11, p. 042617.

[26] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "High-resolution aerial image labeling with convolutional neural networks", IEEE Trans. Geosci, Remote Sens, 2017, vol. 55, pp. 7092-7103.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation", In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5-9 October 2015, Springer: Cham, Switzerland, 2015, pp. 234-241.

[28] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", arXiv, 2015, arXiv:1502.03167.

[29] M. Lavreniuk, N. Kussul, and A. Novikov "Deep learning crop classification approach based on coding input satellite data into the unified hyperspace", 38th International Conference on Electronics and Nanotechnology (ELNANO), 2018, pp. 239-244, doi: 10.1109/ELNANO.2018.8477525.

[30] M. Lavreniuk, N. Kussul, and A. Novikov "Deep Learning Crop Classification Approach Based on Sparse Coding of Time Series of Satellite Data", In IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium, 2018, pp. 4812-4815, doi: 10.1109/IGARSS.2018.8518263.

[31] A. Shelestov, H. Yailymova, B. Yailymov, L. Shumilo, and A. Lavreniuk "Extension of Copernicus Urban Atlas to non-european countries", 2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2021, Brussels (virtual format), pp. 6789-6792, doi: 10.1109/IGARSS47720.2021.9553546.

[32] G. M. Bakan and N. N. Kussul, "Fuzzy ellipsoidal filtering algorithm of static object state", Problemy Upravleniya I Informatiki (Avtomatika), 1996, no. 5, pp. 77-92.

[33] A. N. Kravchenko, N. N. Kussul, E. A. Lupian, V. P. Savorsky, L. Hluchy, and A. Y. Shelestov, "Water resource quality monitoring using heterogeneous data and high-performance computations", Cybernetics and Systems Analysis, 2008, vol. 44(4), 616-624. doi:10.1007/s10559-008-9032-x.

[34] N. Kussul, A. Shelestov, B. Yailymov, H. Yailymova, M. Lavreniuk, L. Shumilo, and Y. Bilokonska, "Crop monitoring technology based on time series of satellite imagery", 2020 IEEE 11th International Conference on Dependable Systems, Services and Technologies, DESSERT 2020, May 2020, pp. 346-350.

[35] N. Kussul, A. Shelestov, H. Yailymova, B. Yailymov, M. Lavreniuk, M. Ilyashenko, "Satellite Agricultural Monitoring in Ukraine at Country Level: World Bank Project", 2020 IEEE International Geoscience and Remote Sensing Symposium, IGARSS 2020, 26 September 2020, pp. 1050-1053.

[36] A. Shelestov, A. Kolotii, S. Skakun, B. Baruth, R.L. Lozano, and B. Yailymov, "Biophysical parameters mapping within the SPOT-5 take 5 initiative", European Journal of Remote Sensing, vol. 50, Issue 1, 2017, pp. 300-309.

[37] N. Kussul, A. Kolotii, A. Shelestov, B. Yailymov, and M. Lavreniuk, "Land degradation estimation from global and national satellite based datasets within un program", 2017 IEEE 9th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS 2017, vol. 1, 3 November 2017, pp. 383-386.

Learning Management Systems as a Platform for Deployment of Remote and Virtual Laboratory Environments

Anastasiia Sapeha[1], Aleksandra Zlatkova[2], Marija Poposka[2], Filip Donchevski[2], Kirill Karpov[1], Zdravko Todorov[2], Danijela Efnusheva[2], Zhivko Kokolanski[2], Andrej Sarjas[3], Dusan Gleich[3], Marija Kalendar[2] and Eduard Siemens[1]

[1]*Department of Electrical, Mechanical and Industrial Engineering, Anhalt University of Applied Sciences 55 Bernburger Str., Köthen, Germany*

[2]*Faculty of Electrical Engineering and Information Technologies, Ss Cyril and Methodius University 18 Rugjer Boshkovik Str., Skopje, R. N. Macedonia*

[3]*Faculty of Electrical Engineering and Computer Science, University of Maribor, 15 Smolskov trg., Maribor, Slovenia {anastasiia.sapeha, kirill.karpov}@hs-anhalt.de, {aleksandraz, fdoncevski, poposkam, todorovz,danijela,kokolanski, marijaka}@feit.ukim.edu.mk, eduard.siemens@hs-anhalt.de, { andrej.sarjas, dusan.gleich}@um.si*

Keywords: LMS, Moodle, Open edX, WebLab-Deusto, RDP, EjS/EjsS, Distance Learning, Remote Laboratory.

Abstract: The constant development in ICT has not omitted the field of learning activities, digitalizing the concept of learning and making it more effective. This became even more obvious during the Covid-19 Pandemic when all educational activities were forced to be transferred remotely and online, with the help of existing Learning Management Systems. Motivated by this, the paper focuses on reviewing different LMSs and comparing their capabilities towards implementation of remote and virtual laboratories, providing analytical and empirical investigations. The goal is to assess the possibility of implementation and deployment of remote access, firstly to some existing on-campus laboratories hosting lab exercises based on real hardware; and secondly to virtual software platforms also usually available in the laboratories. Using such a set up enables the students to get the look and feel of what it means to be in presence in a real lab, nevertheless virtually. This immersive remote lab experience shall hereby be integrated into an already existing and widely used LMS for better operability and manageability. This work has been done in the course of the European project UbiLab ("A ubiquitous virtual laboratory framework"), conducted jointly between the Ss. Cyrill and Methodius University in Skopje, North Macedonia, Anhalt University, Koethen, Germany, and University of Maribor, Slovenia. In the course of the project a remote collaboration platform for hardware-in-the loop will be developed and deployed, and at the same time learning experience in using it in the course of the education process of electrical, system automation, and computer engineers shall be gathered. As an outcome of the work done, it was found that there is no ready-made solution for carrying out remote laboratory work that completely covers the goals of the UbiLab project. Thus, a custom solution based on the compilation of listed platforms components should be developed.

1 INTRODUCTION

The previous couple of years nave been very challenging in many aspects for the modern world. They have brought changes in the usual functioning of many areas of business, life, and education as well. Due to these challenges, the presence of students at the university campuses, and most of all, in the university laboratories has been significantly limited. We can argue that the most affected aspect of students' study experience is the limited access to the laboratory experiments that used to be conducted on site, in the lab facilities of the universities. Even though, the research of virtual and remote laboratories is not something that is emerging at the moment, today's challenges brought by the Covid-19 Pandemic, have pushed forward the need to intensively explore down this road and provide improved virtual and remote laboratories and implement novel ideas in this area. The main goal of the "A ubiquitous virtual laboratory framework - UbiLAB" project, supported by the Erasmus+ framework in the special "Coronavirus response: Extraordinary Erasmus+ calls to support digital

education readiness and creative skills", [1], is exactly aiming in this direction. The UbiLAB project aims to design a complete Virtual Laboratory Framework enabling the process of designing and implementing more realistic remote and virtual laboratory exercises, and at the same time supporting the entire process of university study, as well as the collaborative learning experience of students. Thus, taking into account the learning experience that educators used to provide to their students, we have identified two most difficult challenges: a nearly realistic substitute for learning experiences in the actual physical laboratories; and a nearly realistic substitute for the social element of collaborative learning and "making friends" in the process [2, 3]. These challenges are the main focus of the UbiLAB project.

On the other hand, today it is almost impossible to imagine any kind of training without the support of LMS (Learning Management Systems) platforms, especially when everything takes place at a distance. Two of the most popular and widely used LMS platforms, that we decided to focus on, are Moodle and Open edX [3, 4, 5, 6]. These LMS platforms, despite the fact of being most widely used on one hand, are open-source on the other. Additionally, due to our focus on remote and virtual laboratories and the need to provide remote access to actual hardware, we add to the review the specialized RLMS (Remote Laboratory Management System) Weblab - Deusto [7]. Reviewing these platforms defines their strengths and weaknesses, and subsequently will enable building upon those previous experiences to produce a novel, strengthened and a more suited UbiLAB platform. This platform will be addressing current modern challenges reinforced by the ongoing Covid-19 pandemic, but will nevertheless be practical and usable in general for the currently developing era of digital living, work and education. The review of the platforms will focus on the elements required by the students and teachers, always having in mind the almost realistic, but nevertheless virtual, experiences in the educational, practical and social aspects.

The rest of the paper is organized as follows: The second section describes each reviewed platform from an architectural and feature viewpoint. The third section is focused on the possibilities to implement laboratory environments and the special tools and features for virtual and remote laboratory experiments. The fourth section compares all the reviewed platforms, especially focusing towards the features expected from the UbiLAB platform. Finally in the last section we summarize and conclude the paper.

2 ARCHITECTURE AND FEATURES OF THE REVIEWED PLATFORMS

Even though many LMS platforms exist and are widely used, this section will focus on the three chosen (R)LMS systems. The first two (Moodle and Open edX) are complete full featured LMS platforms and both have an active developer community that constantly improves the platform by adding new elements and modules (tasks of specific formats, new analytic tools, and similar). The third platform, Weblab-Deusto, is a specialized Remote Laboratory Management System focused on providing the elements for developing and implementing remote and virtual laboratories, but also enabling some of the features of a usual LMS system. This section will focus on all three platforms, explaining their architecture, main features, and their extensibility.

2.1 Moodle LMS

Moodle (Modular Object-Oriented Dynamic Learning Environment) is an open-source LMS mainly intended to improve online and remote learning[8]. The first version of Moodle was introduced in 2002 to provide students and teachers with the needed technology for interactive and cooperative distance learning. Moodle has evolved to offer access to a great variety of learning materials, opportunity to connect with peers, various tools to support online learning activities (discussion forums, chats, assignments, quizzes, grading books and a plethora of interesting community contributed add-ons). As such, Moodle contributes to improving the effectiveness of online learning [9]. As seen in the following sections, Moodle is not natively intended to support virtual and remote laboratories.

2.1.1 Moodle LMS Architecture

Moodle has been developed based on the very popular open-source LAMP architecture (Linux, Apache, MySQL, PHP), presented in figure 1. Nevertheless, a wide range of operating systems, database systems, web servers, and programming languages can be supported due to Moodle's open-source and modular nature.

As Figure 1 presents, the architecture of Moodle utilizes the most common elements of today's web applications: a request made by the user; consecutively passed to the web-server that calls the PHP module responsible for the call; the PHP module calls the database with an action that returns the

requested data in the form of HTML code to the web-server. Finally, the information is displayed back to the user.

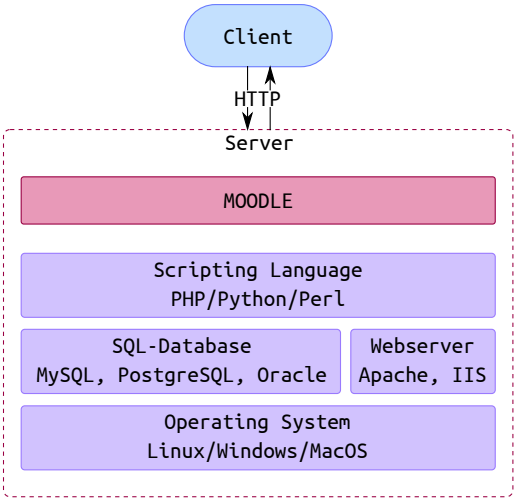


Figure 1: Simplified architecture of the Moodle platform.

2.1.2 Features of Moodle LMS

This section presents a selection of Moodle's features. Some of them are especially interesting for the possibilities to design virtual and remote laboratories; keeping track of students' progress; and the collaborative study experience:

- **Open source:** Moodle is full-feature open source project, so the source code can be downloaded and tailored to the needs.
- **Security and privacy:** A very important feature enabling complete control over the stored and used data.
- **Flexible learning:** Moodle provides its own building education platform that supports deep collaborative learning, through group activities, exploration and experimentation, Calendar, Notifications and Messaging.
- **Mobile learning:** Users can access all content, activities and assignments through a Moodle App, in order to improve the user experience on modern touch screen devices.
- **Scheduling:** Moodle provides extensive ability to set up timed schedules for resources and modules, as well as creating deadlines for tasks.
- **Course access control:** The administrator staff can configure the number of students who have access to the material, and can provide access to resources to defined groups and/or students.
- **User tracking:** Moodle has elaborate resources for tracking student activities in the course and in the separate activities, to enable different analytic and grading. Moodle has a quite elaborate grading system implemented.
- **Administration:** The Moodle platform uses modular and detailed administration tools on many levels: site, course, user administration.
- **Easy integration and extensibility:** Moodle can be integrated with various other programs and platforms to suit different needs. Moodle is a highly modular system. Thus besides the core components, Moodle comprises a wide variety of modules and plugins. Some are incorporated in the Moodle plugin database. However, many community-contributed plugins exist that enable extending Moodle in versatile ways. Moodle also supports the most popular standards in e-learning: IMS, AICC, and SCORM.

All the features present Moodle as a full-featured LMS platform with all the needed elements for course creation and modeling. On the other hand, as seen from the usual features, Moodle lacks the elements for building and connecting virtual and remote laboratories. Nevertheless, Moodle's extensibility and modularity make it an obvious candidate for an easily extensible platform that can be modeled to incorporate needed features, even though they have not been implemented yet.

2.2 Open edX LMS

Open edX is an open-source educational platform that allows organizing online learning for various educational tasks: an online campus, instructor courses, group training programs and one-off training courses.

Open edX includes two main components - Studio for developing course content and LMS. Studio is a tool that manages xBlocks, individual pieces of content from which courses can be composed in the desired sequence. Each xBlock is a Sharable Content Object Reference Model (SCORM) compliant, so material in Open edX can be easily imported from different compatible systems (Camtasia, Articulate Storyline, Adobe Captivate), giving users more flexibility in composing and organizing content [6].

2.2.1 Open edX LMS Architecture

The Open edX platform consists of LMS components with which students interact directly, and Studio designed components used for learning management and course authoring.

Figure 2 shows a simplified architecture of the Open edX platform and its main components [10].

The LMS uses multiple data stores. For example, MongoDB is used to store courses, MySQL stores student data, and videos are hosted on Amazon S3. Studio is the course development environment which enables creating and updating educational material. Studio writes its courses to MongoDB. The server-side code in Open edX is written in Python. Django is used as a framework for web applications.

On the other hand, Open edX courses are made up of modules called XBlocks. Open edX provides the ability to write own XBlocks, which gives the instructors great flexibility in presenting course information and extending the platform.

However, there are several other ways to expand the functionality of the course. For example, course authors can embed Learning Tools Interoperability (LTI) components to integrate third-party learning tools into an Open edX course. JavaScript components can be integrated using JS Input. Furthermore, because the Open edX platform is built primarily using Python, the authors can use built-in Python code to present a problem or evaluate a student's response. The Python code written by the instructor is executed in the secure CodeJail environment. Elasticsearch was also successfully integrated into Open edX and provides a more flexible search for information on the platform.

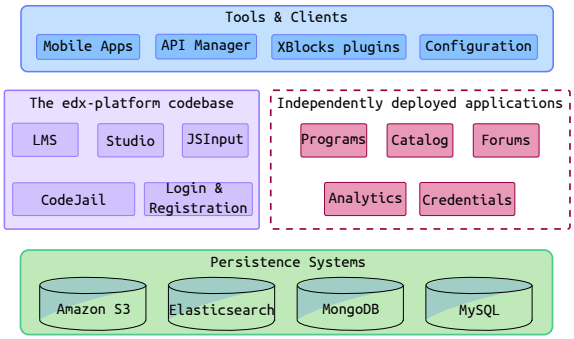


Figure 2: Simplified architecture of the Open edX platform.

2.2.2 Features of Open edX LMS

Open edX [11], like any other LMS, provides some standard functionality, such as tracking student progress, a platform for group discussion or sharing materials, as well as the availability of various types of quizzes to check and evaluate lecture material. The following platform features can be noted:

- **Open source and free**. Similar to Moodle, the source code can be downloaded and freely modified.
- **Scheduling:** Open edX enables setting up release schedules for certain Units, and creating deadlines for completing tasks.
- **Course access control:** The number of students who have access to the material can be limited, and access to lectures can be provided to specific groups, or even students.
- **User tracking:** Open edX also has extensive user tracking capabilities on course level and study modules level as well.
- **Administration:** The Open edX platform includes a specialized administration panel and module for course and user administration.
- **Intuitive interface:** The interface allows students to take classes easily, and tutors to create a course without even looking at the manual.
- **Certificates:** According to the results of the course, the student can receive a certificate. It can be generated from the default template of the Open edX platform, but lecturers have the ability to modify and upload their own templates as well.
- **Integration with Google:** Users can add content from both the Google Calendars and Google Drive apps to the course. Students can view the class schedule and files uploaded by the tutor.
- **Flexible search:** Elasticsearch allows users to search by keywords not only in the list of courses or teachers, but also in the comments database.
- **Extensible:** The Open edX platform is made up of XBlocks, that can be modified as needed, or written by course developers.

Considering the building components of the Open edX platform and the extension possibilities, it can be concluded that Open edX is a full-featured LMS system used typically as a self learning course environment with many possible extensions, but mostly in the virtual software (laboratory) learning domain.

2.3 WebLab-Deusto - A Specialized RLMS

WebLab-Deusto is an open-source RLMS used to develop and manage remote (physical) and virtual laboratories [7]. This platform was developed at the University of Deusto in the early 2000s. At first, it was used mainly in classes for CPLDs and FPGAs, but its acceptance contributed to the development and implementation of various other laboratories [12]. The usefulness and usability of remote laboratories was researched, and the results show that the platform is functional and useful, and accordingly, students get a good feeling of control over the hardware/equipment [13, 14].

2.3.1 Architecture of Weblab-Deusto RLMS

The remote laboratories developed in Weblab-Deusto can be managed or unmanaged laboratories [13]. The architecture of Weblab-Deusto is presented in Figure 3.

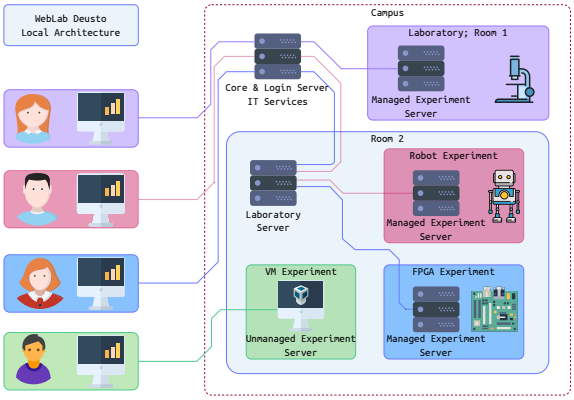


Figure 3: Simplified architecture of the Weblab-Deusto platform.

The managed laboratories are developed using the application programming interface (API) of Weblab-Deusto. This enables bypassing the development of communications, security, and information stored in databases regarding the experiment server because it is being handled through the API of the platform. Weblab-Deusto supports and provides API libraries for multiple programming languages: Python, Node.js, Java, .NET, C++ and C. Managed laboratories consist of two parts: a client and a server; the developer does not manage the communication; and the data is stored in a database that can be used for user tracking.

On the other hand, unmanaged laboratories are those where the communication does not go through WebLab-Deusto. An unmanaged laboratory handles the experiment server communications directly, running a custom server-client app without regard to Weblab-Deusto API. This enables the flexibility of using any programming language, framework, or protocols not supported by Weblab-Deusto (WebSockets, virtual machines, SSH/VNC/Remote Desktop). However, Weblab-Deusto is responsible for the authentication, authorization, scheduling and user tracking in the time slot during the establishment of the communication between the experiment and client. The client (student) using HTTP with JSON connects to the main server responsible for the authentication, authorization, scheduling, user tracking; then he chooses the laboratory, and the main server sends the request to the laboratory. If the authorization is verified, the student gets the privilege to access the laboratory. When another user uses the laboratory, the student will wait in a queue.

2.3.2 Features of Weblab-Deusto RLMS

Weblab-Deusto is a specialized platform for remote laboratories (virtual and physical). The benefits of this platform can be observed through the following useful features [13]:

- **Open Source:** Weblab-Deusto is also an open source platform, and it requires exclusively open source technologies. It also supports proprietary technologies for optional extension.
- **Extensible:** WebLab-Deusto is designed to ease the development of remote laboratories. It provides APIs and different approaches to include new or existing remote labs.
- **Administration:** WebLab-Deusto provides extensive web administration tools.
- **LMS integration:** WebLab-Deusto laboratories can be accessed from LMSs, for example Moodle, relying on the LMS for authentication and authorization.
- **LightWeight:** Weblab-Deusto can run at low-cost devices, but it must be supported by the requirements for the laboratory.
- **Federation:** Weblab-Deusto enables the educational institution to share the designed laboratory with other educational institutions.
- **Scheduling:** WebLab-Deusto guarantees exclusive access to the laboratories through a priority queue subsystem. It can be customized to support multiple concurrent (optionally collaborative) users to a single laboratory.

Considering the features, WebLab-Deusto is an open-source, extensible RLMS platform incorporating some of the usual LMS platforms features (scheduling, user tracking- in the course of using the laboratory), but mainly specialized in the design of remote and virtual laboratories.

3 ADD-ONS AND CAPABILITIES FOR DEPLOYMENT OF REMOTE AND VIRTUAL LABORATORIES

This section willl present the add-ons (modules) enabling development and implementation of remote and virtual laboratories by extending the functionalities of the usual full-featured LMS platforms lacking native support for such laboratories. Additionally, the native capabilities

of WebLab-Deusto for building the remote and virtual laboratories will also emphasized. Finally, the section will present some example remote and virtual laboratories developed in each platform.

3.1 Easy Java/JavaScript Simulation in LMS

Easy Java/JavaScript Simulation (EjS/EjsS) is also a free and open-source standalone tool (not an LMS) used for creating virtual and remote laboratories[15]. This tool is being independently developed for more than a decade, and has been extensively used to create a great number of simulations and remote labs, mostly in the Physics domain (ComPADRE-OSP digital library [16]), as well as many virtual and remote labs in the automatic control field (UNILabs network [17, 18]). Eventhough most of the previous applications were based on Java and deployed as Java applets, EjsS now focuses on building Javascript simulations.

An example of a simple Physics experiment with mass and spring is shown in Figure 4.

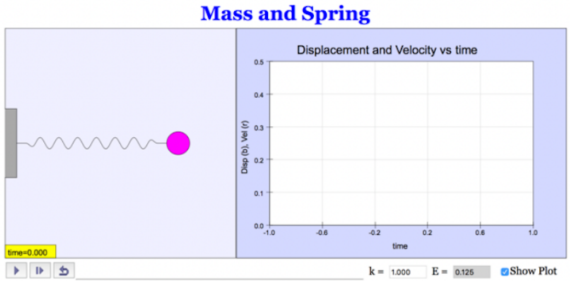


Figure 4: EjS Mass and Spring Simulation Example.

The EjsS is a tool that provides a full platform of elements and libraries to design and implement simulations conveying the laws of Physics in many different fields of study. The full set of already available features makes it simple, not only for scientists with computer skills, to build such realistic simulations comprising many predesigned graphic components specialized for a given type of visualization or user interaction, in a sophisticated HTML interface. Through these front-end elements, the user has full control over the simulation (reset to initial state, change the variables through input fields and/or slider elements, run, pause, reset or run the simulation step by step). The EjS/EjsS platform also enables elements for extending the simulations and accessing remote physical laboratories and communicating with real equipment [19, 20].

3.1.1 Moodle Extension for EjS/EjsS Simulations

Since the EjS/EjsS platform only provides the modules for designing simulations, virtual and remote laboratories, in order to enable a full set of classroom/laboratory experience as intended in the UbiLab Project, an additional LMS, such as Moodle, would be required. Since Moodle is the most widely used open-source LMS, and highly extensible by developing plugins or add-ons, the EjS/EjsS platform has utilized this possibility and has developed a specialized and fully featured Moodle plugin called EJSApp[21].

This plugin enables all EjS/EjsS applications to be embedded into the Moodle LMS. By embedding the applications in the Moodle LMS, they benefit from the integration with the usual and additional Moodle LMS features: integrating a booking system for controlling the access to the RL, multi-language support, saving data and image files from the virtual or RL application to users' file repository in the LMS, grading, monitoring the time spent by users working with the experiment, as well as backup and restore options.

Thus, preparing an EjS/EjsS application (simulation or remote lab) beforehand, and then implementing it into Moodle LMS using the EJSApp plugin [22] is a very good option that enables many possibilities for virtual and physical remote laboratory experiments, while accompanying them with the usual students' tracking and learning options provided by the LMS platform.

Figure 5 presents the architecture and features overview of the EJSApp plugin incorporated within Moodle LMS.

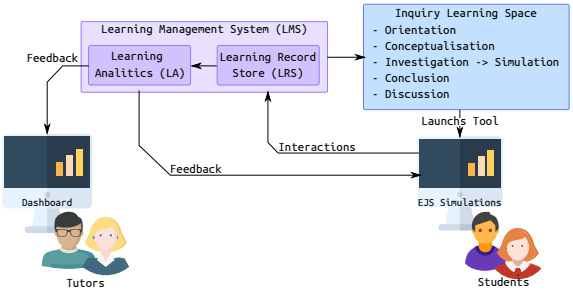
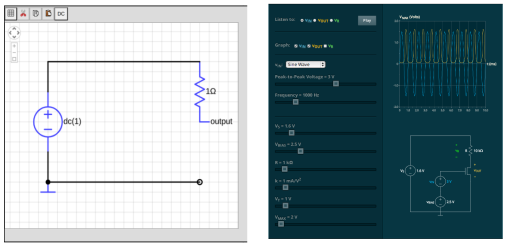


Figure 5: EJSApp Moodle LMS plugin and its features.

3.2 Open edX Laboratory Extensions

As already noted, the Open edX LMS is built upon the XBlocks framework that enables designing interactive tasks for students [10]. The XBlocks



(a) (b)

Figure 6: Examples of Using XBlocks to Create Labs. (a) The Circuit Schematic Builder, (b) The Conditional XBlock.

framework enables creating not only quizzes, but more complex tasks as well. XBlocks components support compilers for various programming languages. Some examples of using XBlocks in the labs are presented in (Figure 6).

The Circuit Schematic Builder (Figure 6(a)) allows students to create virtual circuits by placing elements such as voltage sources, capacitors, resistors, and transistors on an interactive grid. The system evaluates a permanent variable or enables transient circuit analysis. Flowchart-based Conditional XBlocks (Figure 6(b)) allows students to change the input while watching the output change.

One of the most customizable blocks is the iFrame. This comprises HTML code used to embed interactive media, as well as third-party pages/code into a web-page. iFrame creates a separate window in the html document, which is located inside a regular document, and it allows the user to load other independent documents, videos and interactive media files into the page in an area of given size.

These functionalities of the XBlocks framework present a quite flexible possibility for implementing different types of virtual and remote laboratories and presenting various and intuitive user interfaces.

3.3 Weblab-Deusto Remote Laboratories

The WebLab-Deusto platform, as presented before, is responsible for the educational, organizational and technological objectives [13] of building remote laboratories. It provides defined API modules to ease the process of laboratory development. Consequently, WebLab-Deusto does not require additional extensions for enabling remote and vrtual laboratory development. In addition, there are many ready-made laboratories implemented on this platform[13], [23] that prove the flexibility and modularity of Weblab-Deusto for building additional and diverse laboratories. For example, one of the most popular implemented physical experiments in the laboratory is Arduino Robot (Figure 7) where students can practice their robot programming skills remotely, but at the same time see the results through a live streaming camera. The task is to program the robot to follow a line or avoid walls.

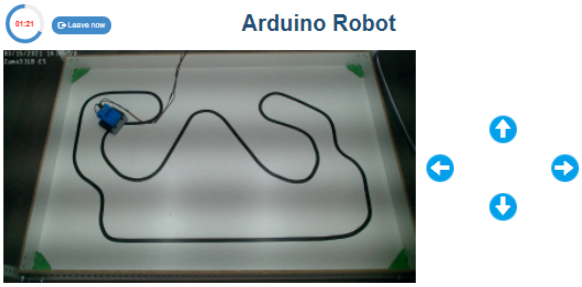


Figure 7: Line Follower Arduino robot.

The development of a new laboratory in WebLab-Deusto is mostly left to the creativity of the developers, since there is no specific way of developing a WebLab-Deusto laboratory. The process consists of developing a client-side web based application using common technologies (JavaScript, CSS and HTML). Next, the communication between the client and the laboratory needs to be implemented by utilizing the already available Weblab-Deusto API functions. In order to enable interaction between the student and the laboratory hardware, it is also necessary to implement a server side software using the existing callback functions from the Weblab-Deusto API.

3.4 Apache Guacamole as an LMS Extension

Apache Guacamole is a client-side remote desktop gateway that supports standard protocols such as Virtual Network Computing (VNC) protocol, Remote Desktop Protocol (RDP), and Secure Shell (SSH) protocol. It accesses remote desktops through a web browser, which means no plugins or additional software is required. It should be noted that this tool can also be used to extend the capabilities of other platforms. The architecture of Apache Guacamole is presented in Figure 8.

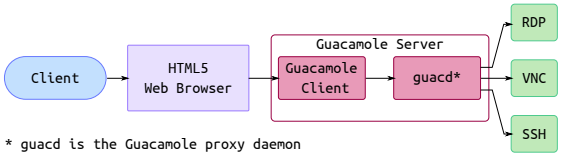


Figure 8: The architecture of Apache Guacamole.

Apache Guacamole supports LDAP database integration, which provides students with a smoother connection to remote machines. This tool is especially useful when using an iFrame in an LMS, and as such it can be noted as a universal extension for LMS platforms such as Moodle and Open edX when remote laboratory work needs to be carried out. Figure 9 shows a diagram of a remote desktop connection using Open edX LMS, an iFrame tool, Apache Guacamole, and a student account (StA) in LDAP. Figure 10 shows an example of using Apache

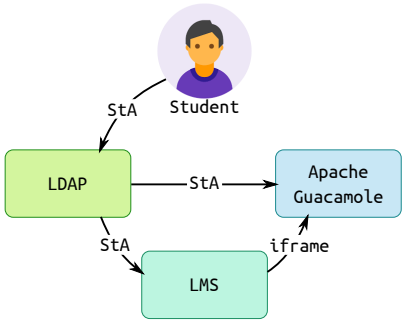


Figure 9: Learning system authentication blocks.

Guacamole to perform remote lab work in a Matlab application.

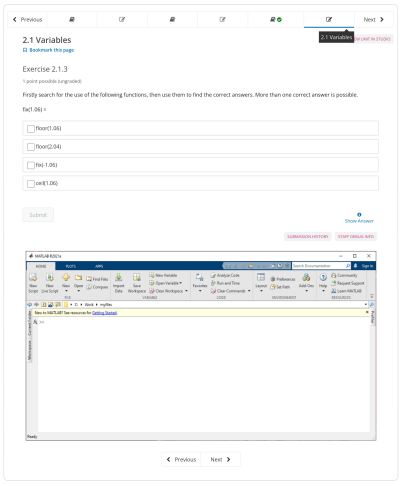


Figure 10: Open edX assignment with remote desktop.

4 PLATFORMS REVIEW REGARDING VIRTUAL AND REMOTE LABORATORY EXTENSIONS

This section presents the comparison of the different reviewed platforms features regarding the possibilities for implementing virtual and remote physical laboratories, as well as the possibilities for extending the platforms with additional own collaborative learning social elements. Table 1 presents a summary of the evaluation of the platforms' features.

Table 1: General comparison.

	Moodle	Open edX	WLDeusto
Virtual Simulations	Yes	Yes, through XBlocks	Limited
Real Time Simulations	Yes	Yes, through XBlocks	Yes
Scheduling System	Yes with plugin	Yes	Yes
User tracking	Limited	Yes	Yes
Grading System	Yes	Yes	No
Programming Language	PHP/JS	Python	Python
No. of concurrent laboratory users	Limited by lab type	Limited by lab type	30
Video call integration	Limited	Yes, through XBlocks	Limited
Mobile learning	Yes	Yes	Yes
Integration System	Moodle dependent	Stand-alone	Stand-alone
No. of concurrent platform users	Limited by hardware	Limited by hardware	Limited by hardware
Remote access to desktop device	N/A	Limited by lab type	Limited
Laboratory sharing across universities	Yes	Yes	Yes

As presented, the platforms are being evaluated according to the following features as they represent the most important elements that would provide insight for the future building of the unified UbiLAB virtual and remote laboratories platform. Each feature is elaborated according to the importance for building the virtual and remote laboratories.

1) **Virtual Simulations** The user can run and observe virtual (software) simulations.
2) **Real-time Simulations** The user can run and observe real-time (physical) simulations.
3) **Scheduling System** The user can check if a laboratory is available and schedule the laboratory.
4) **User tracking** The platform tracks the user progress with the laboratory exercise.
5) **Grading System** Incorporation of student grading in the system.
6) **Programming Language** The programming language used for the development of simulations.
7) **No. of concurrent laboratory users** The maximum number of students (incl. the professor) in a laboratory.
8) **Video call integration** Incorporation of video call system during laboratory exercises within the platform.
9) **Mobile learning** Support for learning from mobile devices (tablets and smartphones)
10) **Integration System** The system on which the LMS platform runs.

11) **No. of concurrent platform users** The maximum number of concurrent users using the LMS.
12) **Remote access to desktop device** The lowest system level of access of the user to the platform.
13) **Laboratory sharing across universities** Sharing laboratory resources between universities.

The features: **ease of new simulation development** and **ease of new simulation implementation**, quantify the difficulty level for a new user to develop or implement a new simulation. For all reviewed platforms we concluded that for an average new user, the difficulty level for development is medium, and for implementation is easy.

According to the features and comparison of the reviewed platforms we can argue that all of the platforms mainly possess the basic features for enabling a complete full-featured LMS system focused towards development and implementation of virtual and remote laboratories and collaborative learning experience. Nevertheless, each platform presents some unique features making it stronger on one side or the other.

The WebLab-Deusto platform is primarily intended for the development of laboratories with flexibility in used programming languages and the design of the experiment. The tools provided (API interfaces for communication, administration, scheduling, reservation) are mainly aiding the process of laboratory development. The features from a full-featured LMS system are generally lacking.

On the other hand, Open edX has an advantage for self-paced learning and for offering flexible extension capabilities by adding or developing specialized XBlocks and/or iFrame modules. On the downside, the Open edX standard course creation toolkit does not natively include virtual and remote lab development tools.

Finally, Moodle LMS, as a long existing and constantly developing LMS, even though does not provide native tools for laboratory development, has the biggest number and greatest modularity of additionally developed tools, effortlessly integrated in the platform. Due to these extension capabilities, Moodle can be easily upgraded with existing and novel laboratory development tools, as well as innovative collaborative learning modules.

5 CONCLUSIONS

Different LMSs differ not only in functionality, but also in what problems they can solve. Therefore, there is no universal solution to the LMS market and to the goals of the UbiLAB Framework, as well. By choosing to focus on open-source platforms, we have intentionally sought the flexibility of upgrading the platforms (improving the elements and modules that lack or do not have the full features required by the envisioned UbiLAB framework).

Thus, all platforms are good candidates for being the core of the UbiLAB Framework, since they present strengths in different elements. WebLab-Deusto is scored on the top for having full set of ready-made elements for developing new and versatile virtual and physical laboratories. Moodle is put on the top for including most elements for student interaction and success tracking, student materials, scheduling tasks, and student collaboration. Open edX is top-scored for its self-paced learning process and great possibilities for designing modular, versatile, and interactive tasks for the students.

Due to these conclusions, we can propose basing our UbiLAB framework on either Moodle or Open edX platform (as a solid basis for a complete LMS system) and upgrading and developing the additional modules envisioned in the UbiLAB framework. On the other hand, since WebLab-Deusto is the best choice for developing virtual and remote laboratories - an important part of the UbiLAB framework, and because WebLab-Deusto already has a Moodle integration developed, we would focus in this direction, too. Nevertheless, regarding the virtual software environments and interaction with the students, Open edX seems superior. Thus we would also consider integrating its advantages into the final full-featured UbiLAB framework, mostly for enabling the envisioned virtual software laboratories.

6 ACKNOWLEDGMENTS

This work has been supported by the European Erasmus+ project ”A ubiquitous virtual laboratory framework” (UbiLAB, ERASMUS+ Key Action 2, ref: 2020-1-MK01-KA226-HE-094548).

REFERENCES

[1] Coronavirus response: Extraordinary erasmus+ calls to support digital education readiness and creative skills. [Online]. Available: https://erasmus-plus.ec.europa.eu/news/coronavirus-response-extraordinary-erasmus-calls-to-support-digital-education-readiness-and-creative-skills-0.

[2] I. Gustavsson et al., “On objectives of instructional laboratories, individual assessment, and use of

collaborative remote laboratories," IEEE Transactions on Learning Technologies, vol. 2, no. 4, pp. 263–274, 2009.

[3] D. Galan et al., "Automated assessment of computer programming practices: The 8-years uned experience," IEEE Access, vol. 7, pp. 130 113–130 119, 2019.

[4] F. J. G. Clemente et al., "Development of learning analytics moodle extension for easy javascript simulation (ejss) virtual laboratories," 2019.

[5] N. Cavus, "Distance learning and learning management systems," Procedia - Social and Behavioral Sciences, vol. 191, pp. 872–877, 2015, the Proceedings of 6th World Conference on educational Sciences. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877042815028712.

[6] Y. Chaabi et al., "Development of a learning analytics extension in open edx," in 2021 International Conference of Modern Trends in Information and Communication Technology Industry (MTICTI), 2021, pp. 1–6.

[7] P. Orduña et al., "The weblab-deusto remote laboratory management system architecture: Achieving scalability, interoperability, and federation of remote experimentation," in Cyber-Physical Laboratories in Engineering and Science Education, M. Auer et al., Eds. Springer, Cham, 2018, ch. 2, pp. 17–42. [Online]. Available: https://doi.org/10.1007/978-3-319-76935-6 2.

[8] N. Simanullang et al., "Learning management system (lms) based on moodle to improve students learning activity," Journal of Physics: Conference Series, vol. 1462, p. 012067, 02 2020.

[9] N. H. S. Simanullang et al., "Learning management system (LMS) based on moodle to improve students learning activity," Journal of Physics: Conference Series, vol. 1462, no. 1, p. 012067, feb 2020. [Online]. Available: https://doi.org/10.1088/1742-6596/1462/1/012067.

[10] "Open edX Developer's Guide," Circuit Sub PBLLC, Developer's Guide, January 2022. [Online]. Available: https://openedx.org/.

[11] "Open edX Learner's Guide," Circuit Sub PBLLC, Tech. Rep., February 2022. [Online]. Available: https://openedx.org/.

[12] J. Garcia-Zubia et al., "Acceptance, usability and usefulness of weblab-deusto from students point of view," in 2008 Third International Conference on Digital Information Management, 2008, pp. 899–904.

[13] J. Garcia-Zubia et al., "Addressing software impact in the design of remote labs," IEEE Transactions on Industrial Electronics, vol. 56, p. 4757 – 4767, 12 2009.

[14] J. Garcia-Zubia et al., "Application and user perceptions of using the weblab-deusto-pld in technical education," in 2011 First Global Online Laboratory Consortium Remote Laboratories Workshop, 2011, pp. 1–6.

[15] Esquembre et al., "Easy java simulations: A software tool to create scientific simulations in java," Computer Physics Communications, vol. 156, pp. 199–204, 01 2004.

[16] F. J. García Clemente, "Ejss: A javascript library which makes computational-physics education simpler," 08 2014.

[17] J. Sáenz et al., "Open and low-cost virtual and remote labs on control engineering," IEEE Access, vol. 3, pp. 805–814, 2015.

[18] D. Chaos et al., "Virtual and remote robotic laboratory using ejs, matlab and labview," Sensors (Basel, Switzerland), vol. 13, pp. 2595–612, 02 2013.

[19] R. Pastor Vargas, "Web-based virtual lab and remote experimentation using easy java simulations," vol. 38, 07 2005, pp. 2289–2289.

[20] J. Chacón et al., "Enhancing ejss with extension plugins," Electronics, vol. 10, p. 242, 01 2021.

[21] L. de la Torre Cubillo et al., "Providing collaborative support to virtual and remote laboratories," IEEE Transactions on Learning Technologies, 10 2013.

[22] L. de la Torre et al., "Easy creation and deployment of javascript remote labs with ejss and moodle," in 2016 13th International Conference on Remote Engineering and Virtual Instrumentation (REV), 2016, pp. 260–261.

[23] Garcia-Zubia et al., "An integrated solution for basics digital electronics: Boole-deusto and weblab-deusto," 02 2013, p. 1–5.