International Journal of Advanced Statistics and IT&C for Economics and Life Sciences

Volume 14
Number 1
December 2024

ISSN 2559 - 365X
eISSN 2067 - 354X

DOI: 10.2478/ijasitels-2024-0010

Transforming System Test

STAN Cătălin[1], TOGHE Răzvan[2], VĂIDIAN Iulia[3]

[1]Marquardt Schaltsysteme SCS, Research and Development Department, Software Engineer

[2]Marquardt Schaltsysteme SCS, Research and Development Department, Software Test Engineer

[3]OMiLAB NPO, Berlin, Germany, Community of Practice Manager

Abstract

During the product System Test phase, it is necessary to determine the passed/failed status of tests by analyzing network signals. A second layer of verification is introduced by using a camera to take pictures triggered by these signals. An image recognition model is used to identify the successful operation of the product (tests passed), by processing the captured images. The main goal of the proposed method is the improvement of quality assurance techniques. The possibility of automated defect detection in the manufacturing process will lead to improved productivity, higher product quality, and transparency in reporting.

Keywords: System Test, image recognition, verification, conceptual modelling, resilience, sustainability, human-centric, industrial processes, automotive

1. Introduction

In today's rapidly evolving industrial landscape, resilience has become a critical characteristic for companies to ensure the continuity, reliability, efficiency, and quality of their industrial processes. Moreover, sustainability in the context of industrial processes is typically associated with reducing waste, conserving resources, and minimizing the environmental impact of production activities. By optimizing internal processes, a company can ensure its long-term operational efficiency. People are at the center of these processes. Therefore, a human-centered approach to process re-design is a must. It places the needs and experience of the people – engineers, operators, and stakeholders – at the forefront of process design and/or optimization and acknowledges that while automation and advanced technologies are crucial, the ultimate goal is to empower human workers.

In this paper we will present the optimization of a system testing process within Marquardt Company, following a conceptual modeling workshop facilitated by OMiLAB NPO and with participation of professors from the "Lucian Blaga" University of Sibiu. Marquardt Company is a German automotive supplier which considers the three pillars of resilience, sustainability, and human-centricity for its processes in order to keep up with the rapidly evolving industrial landscape.

1 Transformation of the industrial processes

1.1 Resilience of industrial processes

For the proposed use case, within the product verification activities at System Test, the introduction of a second layer of verification supports the resilience of Marquardt Company. The System Test phase plays a crucial role in verifying the functionality and quality of products before they are released to the market. Traditionally, this verification phase relies heavily on network signal analysis to determine the pass/fail status of tests. Nevertheless, relying on a single verification method introduces vulnerabilities, such as undetected defects that could compromise the quality of the product. Therefore, the introduction of a second layer of verification also acts as a means of increasing the trust level for test results and business trust between supplier and customer, with Marquardt being a supplier that focuses on pro-active testing. Automation of the prototyping process and industrial production series process also supports resilience.

1.2 Sustainability of industrial processes

As stated above, by optimizing internal processes, a company can ensure its long-term operational efficiency. The presented use case has no direct influence/impact on the changing market conditions, but it maintains Marquardt’s internal operational efficiency and flexibility. As the image recognition approach matures and becomes a stable mechanism to validate tests, the network signal analysis approach could be phased out, at least in some cases. This will reduce waste and preserve resources. Further, it will also free up valuable test time (by removing the physical wiring, there is no need to solder additional wires or execute additional test-box preparation activities), resulting in much faster testing cycles.

1.3 Human-Centered approach for the industrial processes

This case study introduced engineers to Design Thinking and Conceptual Modelling techniques and tools as a proven approach towards finding an optimal solution. Using paper figures as actual anchors for discussion, engineers from Marquardt, professors at ULBS (“Lucian Blaga” University of Sibiu) in partnership with OMiLAB NPO conducted a joint workshop within this scope. The workshop had the following objectives:

- to encourage engineers to understand the operator’s needs when designing processes,
- to foster creativity by encouraging engineers to explore possible solutions together before actually implementing them; the exploration proved to be iterative starting with the physical paper figures, digitalizing the co-created scenes and continuing to refine them .
- to collaborate across company departments, but also outside company, having invited engineers with different backgrounds and professors alike to work together to solve the given problem.

The workshop started with a presentation of the problem in focus, a setup for the tool environment by OMiLAB , followed by the actual debate/collaboration around a central table capturing the attention of all participants. The scenes created with the help of paper figures were captured and digitalized using the Scene2Model tool. The result of this workshop will be detailed in the next chapter.

Further, the implementation of this case study at Marquardt Company will help engineers to improve their programming skills and knowledge within image processing and computer vision domain. This is a win-win situation as there is an ever-growing need for engineers to expand their expertise in these emerging fields.

Another significant benefit derived from the automation introduced in System Test is the ability to free up engineers from repetitive and time-consuming tasks. They could be re-allocated to other higher-value work (e.g. optimizing current processes, innovating) resulting in greater employee job satisfaction.

2 The second layer of verification

As stated before the integration of a human-centered design approach was facilitated by a joint workshop between Marquardt, ULBS, and OMiLAB. In this section, we present the solution using screenshots of the digital scenes captured during the actual workshop. The human interaction and discussions, which were happening around physical paper figures, were digitally documented using the Scene2Model tool from OMiLAB. This tool captures the transition from the actual state (the “as-is” problem) to the future state (the desired solution), presented as a visual story consisting of four scenes (see Fig. 1).

There are many similar products under discussion, grouped by their features into what is referred to as the Switch Panel Product Family, which contains hundreds of variants. These products include, but are not limited to, dashboard panel controls, side door panels, or steering wheel panels. Each of these variants has specific testing requirements, therefore, the solution must be flexible enough to accommodate them.

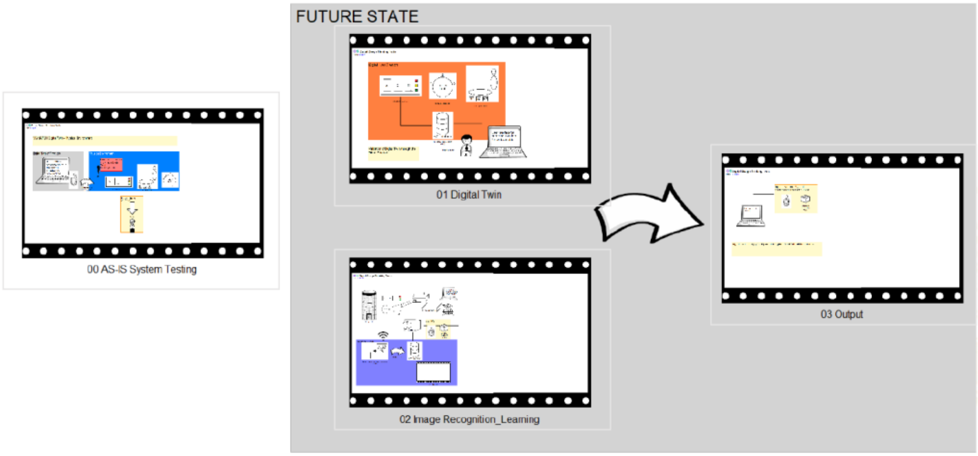


Figure. 1. STORYBOARD: Image Recognition for System Test

The concepts of Digital Twin and Digital Shadow are rather new to the engineers at Marquardt, but daily System Test activities show that the proprietary CANoe Network Simulator, together with the product’s signals descriptions and values, is a good candidate for creating a Digital Shadow of these products. By adding the visual aspects of the product's various operating states, synchronized with signal values, this Digital Shadow can evolve into a Digital Twin.

The current System Test process involves testing the complete product using a Hardware-In-the-Loop (HIL) test bench connected to the network simulator via Fast Data eXchange (FDX) protocol. By using HIL, the solution fully integrates real-world hardware and simulated environments. Based on the signal values recorded during these tests, the test engineer validates the product's expected behavior. In addition, a human visual inspection is performed and a Test Report is generated as the final output (see Fig.2). Automating this process with image recognition further reduces the need for human visual inspections, thereby reducing the human error. As a result, the consistency and reliability of the test results is improved, enhancing business trust and overall product quality.

A first step towards the future state is the creation of a database containing images of the product's operating modes. Building this database will be time consuming at first, but we expect the effort to be paid-off after 3-4 test cycles. For example, the seat temperature button, which has three levels - low, medium, and high – would correspond to three distinct images showing one, two, or three LEDs illuminated respectively. Similarly, for a rotary knob, the different positions of the knob,

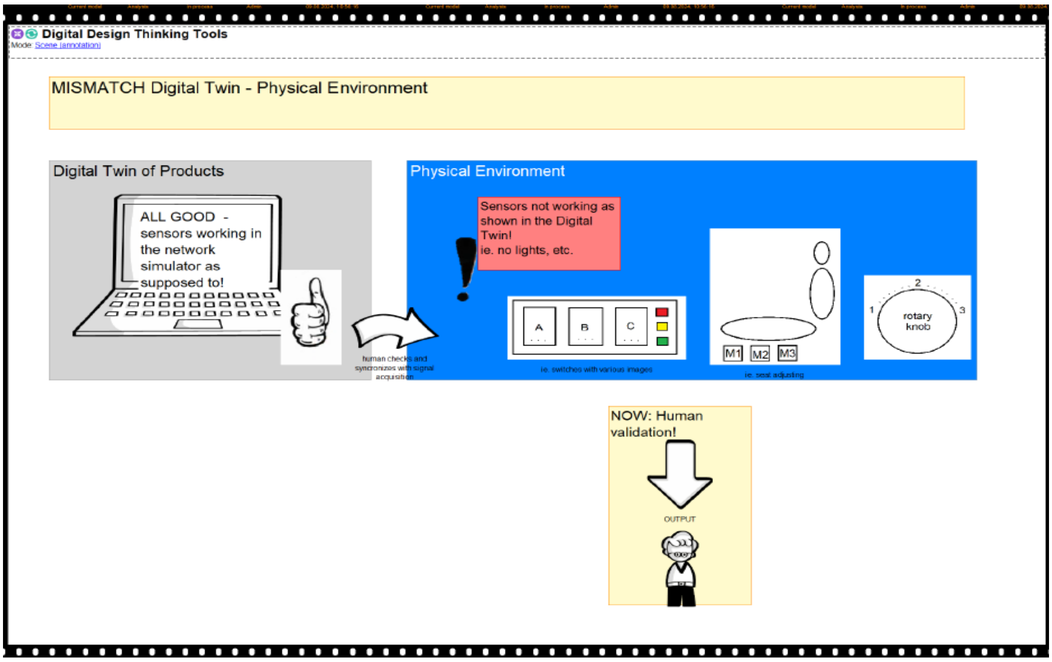

Digital Design Thinking Tools
Mode: Scene (annotation)
MISMATCH Digital Twin - Physical Environment
Digital Twin of Products
ALL GOOD - sensors working in the network simulator as supposed to!
human checks and syncronizes with signal acquisition
Physical Environment
Sensors not working as shown in the Digital Twin!
ie. no lights, etc.
A
B
C
ie. switches with various images
M1
M2
M3
ie. seat adjusting
1
2
3
rotary knob
NOW: Human validation!
OUTPUT

Figure. 2. AS-IS: System Testing

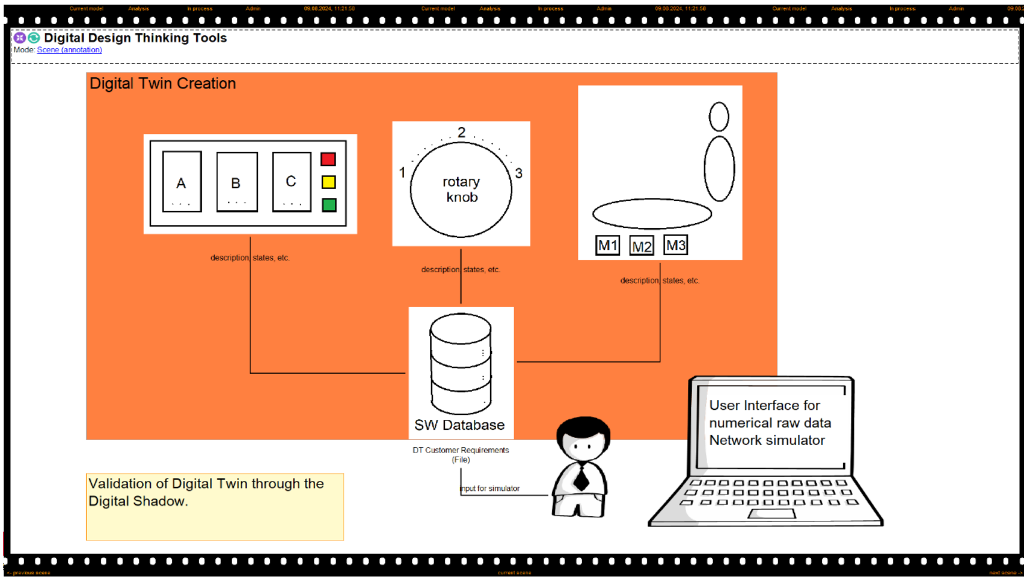

Current model
Analysis
In process
Admin
09.08.2024, 11:21:58
Digital Design Thinking Tools
Mode: Scene (annotation)
Digital Twin Creation
A
B
C
1
2
3
rotary knob
M1
M2
M3
description, states, etc.
description, states, etc.
description, states, etc.
SW Database
DT Customer Requirements (File)
Input for simulator
User Interface for numerical raw data Network simulator
Validation of Digital Twin through the Digital Shadow.
<- previous scene
current scene
next scene ->

Figure. 3. FUTURE STATE: Digital Twin

indicate different functions, or different intensities. They would also be associated with distinct images.

Next, the solution then requires a mapping of these pictures to the product's signal values within the same database (see Fig.3). This will ensure that each signal state is accurately reflected by its corresponding visual representation.

These images form the basis for later training of a Machine Learning (ML) algorithm. The ML model will be used to analyze and correctly identify the images captured during real-world tests (see Fig.4). It will be trained on a diverse set of images, allowing it to generalize well across different product variants. This ensures that the solution remains effective, even for updated versions of the existing products. We should remember that one of the goals is to automate the verification of the product's response based on the simulation inputs.

In the future System Test setup, most of the existing equipment will remain in place, such as the product storage space, the mounting fixture, the data acquisition equipment, the HIL test-bench and the PC running network simulator. However, a camera will be added to capture images of the product (in JPG format) as it responds to the simulation inputs.

For the image recognition process, YOLOv5 will be used. YOLOv5 is well regarded in the open-source community and its use is well documented, making it an ideal choice. It has advanced object detection capabilities allowing it to quickly and accurately identify key features of the product. Its open-source nature also means it can be customized for Marquardt's specific internal use cases, offering flexibility as testing requirements evolve.

The image recognition will be handled by a separate PC, distinct from the one controlling the HIL test-bench. This dedicated machine, equipped with a powerful graphics processing unit (GPU), will handle the computational load, preventing the HIL system from being overburdened. . In this way it is ensured that the entire system remains stable enough and responsive, even when handling large image datasets. It also allows the HIL to focus solely on controlling the physical hardware. We can also appreciate the modularity of this solution as it provides flexibility for future expansions (e.g. adding hardware for more complex tests). The detection results will

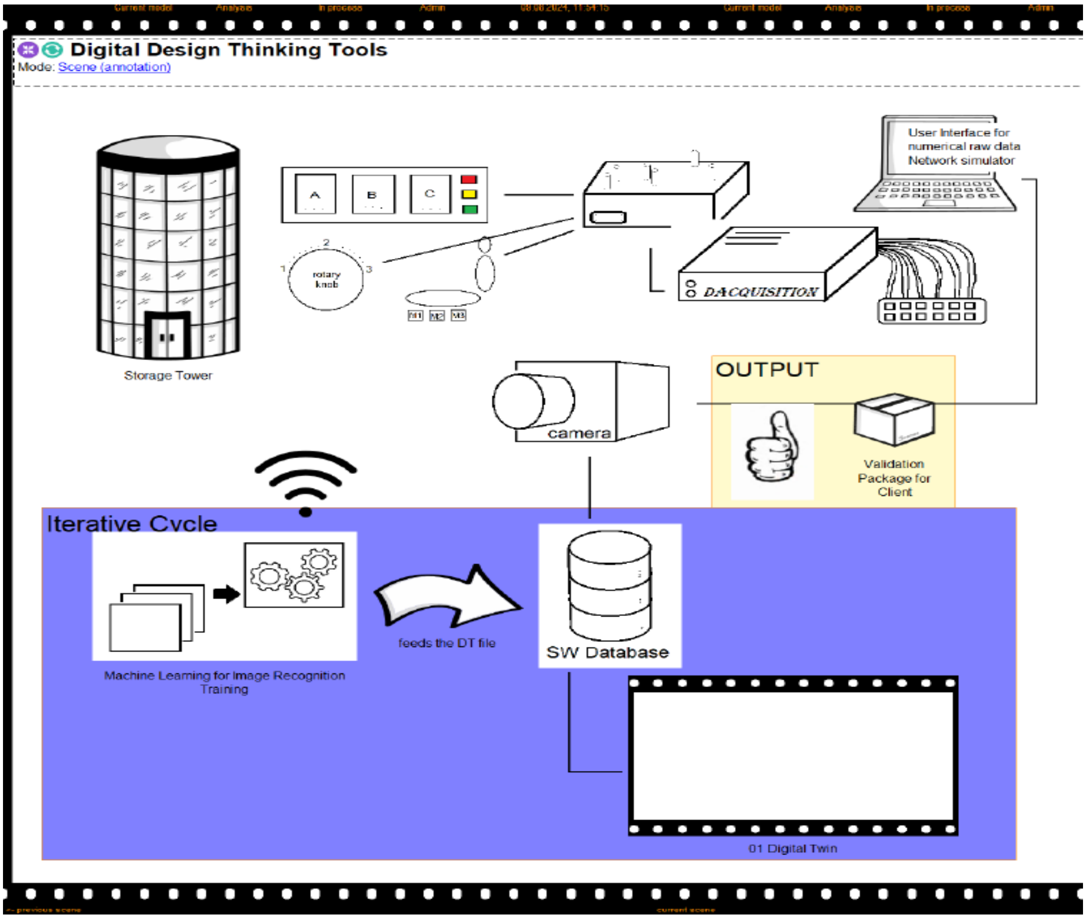
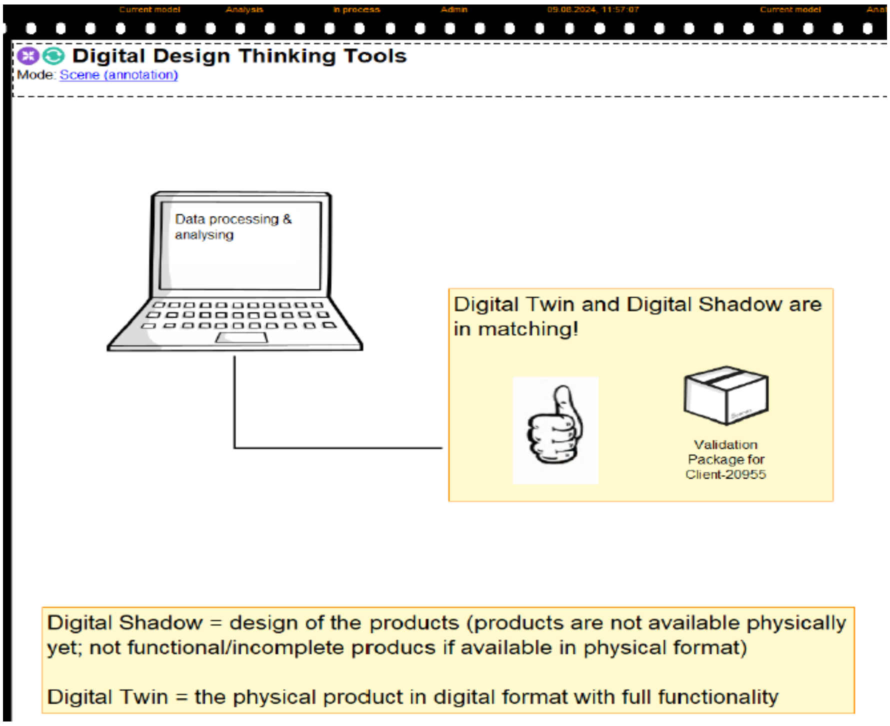


Figure 4. FUTURE STATE: Image Recognition Learning

be sent back to the main PC running the network simulator as JSON strings, including information about the test pass/fail criteria, the final image, and any other relevant information (see Fig. 6).

The final output of the solution is an automatically generated comprehensive test report, which includes test pass/fail criteria, visual evidence from the image recognition system, detailed signal data, and traceability between requirements and tests executed. This report is delivered to the customer as a Validation Package, providing transparency and confidence in the testing process at Marquardt (see Fig.5).



Digital Twin and Digital Shadow are in matching!

Digital Shadow = design of the products (products are not available physically yet; not functional/incomplete producs if available in physical format)

Digital Twin = the physical product in digital format with full functionality

Figure 5. FUTURE STATE: Output

3 Key skills and necessary competences

The skills required to complete the tasks derived from this use case are the ability to analyze temporal events and signals, the ability to operate and deploy complex systems (hardware and software), and the capability to work effectively in a team environment.

The knowledge prerequisites are basic programming skills in Python, C#, or Java, and the use of computer vision libraries like OpenCV and deep learning frameworks such as TensorFlow or PyTorch. These tools are essential for training machine-learning models and deploying the image recognition components of the solution. Additionally, basic knowledge of common network protocols such as HTTP is required, as well as specific expertise in the automotive network simulators used at Marquardt (e.g. CANoe) and their protocols like FDX to ensure accurate communication between hardware and test environment.

Proficiency in mathematical concepts, mainly linear algebra, calculus, and probability is indispensable for Computer Vision Engineers. Data analysis skills are equally important to interpret the test results and tuning algorithms.

Soft skills like problem-solving and critical thinking are essential to troubleshoot complex system interactions and to identify potential issues early in the testing phase. Engineers should demonstrate attention to detail, especially when it is about test accuracy or the integration of various system components.

Furthermore, strong communication skills are vital for collaborating across teams and for documenting processes, and producing clear, concise reports (such as the final customer test report).

4 Results

The case study is in its early stages of development. While the transformation is not yet fully deployed, key components of the system have been prototyped and integrated. For example, the camera-based image recognition system and Machine Learning algorithm training have been implemented (see Figure 6), and early tests showed promising results (e.g. 90% average recognition accuracy). The Hardware-In-the-Loop (HIL) test bench and CANoe network simulator are already an integral part of the testing process with the second layer of verification (image recognition) being gradually introduced.

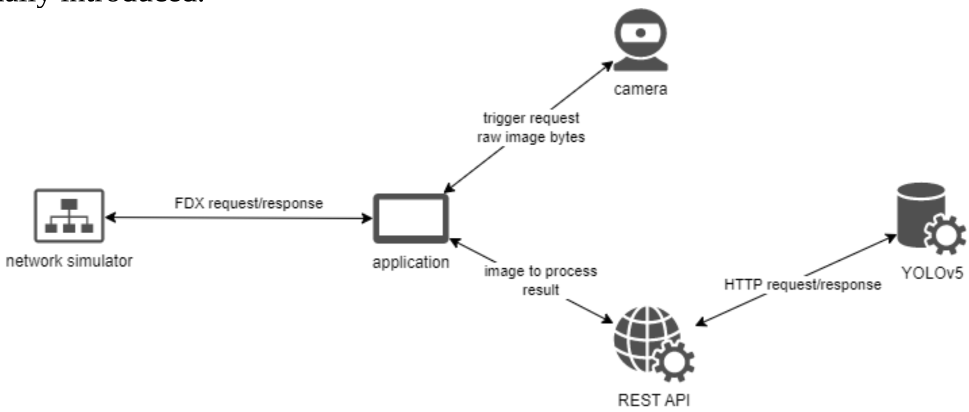


Figure 6. Implementation of main components of the solution

As the system matures we expect greater benefits in terms of productivity (e.g. an engineer would spend less time on visual inspection, therefore we estimate a 10% increase in his work efficiency), cost savings (e.g. related also to man-hours per task, a time reduction in System Test expected to be 15%-20%), reporting transparency (e.g. will include the real captures from the camera and have traceability to the executed tests) and customer satisfaction leveraging further business economic benefits.

5 Conclusions

The integration of advanced automation technologies, such as image recognition and machine learning, alongside more traditional Hardware-In-the-Loop (HIL) testing, represents an important shift for Marquardt in its journey towards Industry 5.0. The proposed solution supported by both human-centered design and Digital Twin concepts, demonstrates the potential to improve product development processes (e.g. System Test phase) and business processes alike, reducing manual errors and enhancing both accuracy and transparency.

By automating repetitive tasks, engineers are freed up to focus on more creative and high-value activities, this aspect being an important step in the Industry 5.0 vision: technology empowers people, not replaces them.

Marquardt’s experience shows the value of collaborative innovation between teams, jointly working on solutions with external partners like ULBS, OMiLAB and also integrating the latest machine learning algorithms with existing network simulation and testing infrastructure. Based on these experiences we outline below some key recommendations for transitioning to Industry 5.0:

- Adopt a human-centered approach: engaging employees in the design and implementation process – through workshops, training, and collaboration – ensures that technologies empower people not replace them. The introduction of Design Thinking into Marquardt's process is a good example of how collaborative problem solving can lead to innovative solutions.
- Invest in skill development and lifelong learning: as companies introduce new technologies like computer vision and machine learning, employees will need to upgrade their skills, knowledge, and competencies for new roles. Companies should invest in training programs, not only in technical areas like Python, and machine learning, but also in soft skills like teamwork and Design Thinking.
- Be data-driven: the success of Industry 5.0 lies in the ability to collect real-time data from physical and digital systems. A data-driven approach is no longer optional, as continuous monitoring and feedback loops will improve product quality faster than ever and maximize operational efficiency.

Acknowledgment

We would like to thank Professor Adrian Florea ("Lucian Blaga" University of Sibiu) for his invaluable guidance throughout this project. His expertise and insights have been essential for the final shape of this paper. Follows sincere appreciation to our partners, Wilfrid Utz and Iulia Vaidian from OMiLAB, whose collaborative efforts have guided us through the conceptual modelling world. This research and publication are supported by the ERASMUS + EU Project CODEMO, CoCreative Decision Makers for 5.0 Organisations, Project Nr. 101104819, Co-funded by the European commission.

References

[1] Gonzalez Rafael C., Woods Richard E., *Digital Image Processing*, London; Sydney, Pearson Prentice Hall, 2018.

[2] https://www.pyimagesearch.com/2015/12/07/local-binary-patterns-with-python-opencv/

[3] Forsyth, D. A., & Ponce, J. (2002). *Computer vision: a modern approach*. Prentice Hall professional technical reference.

[4] Karagiannis, D. (2024). Digital Twins for Design Thinking Support Innovative Business Models, The Kyoto Economic Review, 2024, Volume 198, Issue S, Pages S123-S146, Released on J-STAGE. DOI: https://doi.org/10.57475/keizaironso.198.S9

[5] Karagiannis D. et al. (2020). OMiLAB: A Smart Innovation Environment for Digital Engineers. In: Camarinha-Matos L.M., Afsarmanesh H., Ortiz A. (eds) Boosting Collaborative Networks 4.0. PRO-VE 2020. IFIP Advances in Information and Communication Technology, vol 598. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-62412-5_23

[6] OMILAB Team. (2020). A Digital Innovation Environment powered by Open Models Laboratory. DOI: http://doi.org/10.5281/zenodo.3899990

[7] Vaidian, I., Jurczuk, A., Misiak, Z., Neidow, M., Petry, M., Nemetz, M. (2022). Challenging Digital Innovation Through the OMiLAB Community of Practice. In: Karagiannis, D., Lee, M., Hinkelmann, K., Utz, W. (eds) Domain-Specific Conceptual Modeling. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-93547-4_3

[8] Muck, C., Palkovits-Rauter, S. (2022). Conceptualizing Design Thinking Artefacts: The Scene2Model Storyboard Approach. In: Karagiannis, D., Lee, M., Hinkelmann, K., Utz, W.

(eds) Domain-Specific Conceptual Modeling. Springer, Cham. DOI: https://doi.org/10.1007/978-3-030-93547-4_25

[9] Toncian, V., Florea, A., David, A., Morariu, D., Cretulescu, R. (2024). Leveraging Collaboration for Industry 5.0: Needs, Strategies and Future Directions. In: Camarinha-Matos, L.M., Ortiz, A., Boucher, X., Barthe-Delanoë, AM. (eds) Navigating Unpredictability: Collaborative Networks in Non-linear Worlds. PRO-VE 2024. IFIP Advances in Information and Communication Technology, vol 726. Springer, Cham. https://doi.org/10.1007/978-3-031-71739-0_21

DOI: 10.2478/ijasitels-2024-0001

Using AI for developing personalized learning paths

Ramona-Diana Leon[1], *Ángel Ortiz*[1], *Mª del Mar Alemany Díaz*[1], *Ana Esteso Alvarez*[1]

[1]*Research Center on Production Management and Engineering, Universitat Politècnica de València, Spain {rleon, aortiz, mareva, aesteso} @cigip.upv.es*

Abstract

This research aims to examine how artificial intelligence (AI) can be used within the educational framework for developing personalized learning paths. In order to achieve this goal, an etic approach is employed, and a qualitative-quantitative perspective is adopted. Thus, following the PRISMA guidelines, 71 articles published on Web od Science, during January 2014 – June 2024, are selected and analysed using cluster and density analysis. The results bring forward that the peak of the scientific production was reached in 2022 and that the topic is more appealing to the scholars from the information technology field than to the ones from the educational area. Furthermore, two lines of research can be identified; one that is technology-driven and another one that is learner/human-driven. Further research is required in providing a nexus between the two of them since, in the context of Industry 5.0 and Society 5.0, AI could act as a bridge. This research has several implications. On the one hand, it emphasizes the topics that captured scholars' attention and also various research gaps that should be addressed. On the other hand, it extends the research from the educational management area by highlighting how AI could facilitate the transition towards the implementation of the connectivism learning theories.

Keywords: Artificial intelligence, Personalized learning path, Industry 5.0

1. Introduction

Learning pathways are essential for lifelong learning especially in the context of Industry 5.0 when companies claim that they cannot find employees with the necessary skills to meet market's demands. This situation occurs due to the speed at which new technologies are advancing [1].

To enable workers at all levels to acquire the required competencies for ensuring company's competitiveness in the Industry 5.0 context, it is necessary to develop appropriate learning pathways. These must be adapted to companies' needs and, at the same time, they must be customized, considering employees' current competencies and learning style [2].

Artificial intelligence (AI) can greatly help define personalized learning paths since it is capable to adapt in a faster pace to users' requirements [3]. Taking these into account, the current article aims to analyse how AI can be used within the educational framework to develop personalized learning paths.

To do this, an extensive literature search has been carried out, selecting a set of articles that address this topic, analysing their content and carrying out a co-occurrence analysis and clustering (section 2). The main findings have been developed in section 3. Finally, the article's conclusions are presented in section 4.

2. Research methodology

This research aims to examine how AI can be used within the educational framework for developing personalized learning paths. In order to achieve this goal, an etic approach is employed which encompasses an external view on meaning associations and real-world events. Unlike the emic approach, the etic perspective generates “descriptions and analyses expressed in terms of the conceptual schemes and categories regarded as meaningful and appropriate by the community of scientific observers” [4, p.130]. Hence, a documentary study is used which consists of a review of articles and studies from the educational management journals. The research adopts a qualitative-quantitative perspective and the whole multi-stage process is dominated by an inductive approach.

On a first stage, PRISMA guidelines are used for identifying the most relevant articles published on Web od Science database. Thus, the articles published during January 2014 – June 2024 which include in title, abstract or keywords the phrases “artificial intelligence”, “AI”, and “learning”, “learning path”, “personalized learning” are selected. A total of 390 articles are found; 300 are published in the international journals while 90 are included in proceedings volumes. Each article is analysed in order to determine its relevance for the research problem. As suggested by [5], those which fall out of scope are removed from the final database. As a result, 71 articles are selected since they focus exclusively on the use of AI within the educational framework and not on developing AI-based technologies. 53 of them are included in international journals while 18 in conference proceedings.

On the second stage, for the selected articles a content analysis is employed. This research method is used due to the fact that (i) it is has an analytical flexibility; (ii) it is nonintrusive; and (iii) it entails the specification of category criteria for reliability and validity tests [6]. The main categories in which the analysis focused are: (i) the type of article; (ii) the approach; and (iii) the main topic.

On the third stage, the previously selected articles become subject of co-occurrence and cluster analysis which are performed using VOSViewer 1.16.17 and the steps indicated by [7]. Thus, the similarity matrix is constructed, taking into account the co-occurrence principles, according to which:

$$s_{ij} = \frac{c_{ij}}{w_i * w_j} \tag{1}$$

where s_{ij} is the similarity between item i and j;

c_{ij} is the number of co-occurrences of item i and j;

w_i is the number of occurrences of item i;

w_j is the number of occurrences of item j.

The results of the similarity matrix are further used within the mapping technique in order to obtain a visual representation of the causal linkages established among concepts. The mapping technique is based on the following relationship:

$$V(x_1, \ldots., x_n) = \sum_{i<j} s_{ij} \left\|x_i - x_j\right\|^2 \tag{2}$$

where x_i=(x_{i1}, x_{i2}) reflects where item i is located;

$\|\blacksquare\|$ is the Euclidean norm.

Last but not least, a cluster density analysis is performed in order to emphasize the hot topics that are drawing academics attention. This involves:

$$D_p(x) = \sum_{i=1}^{n} I_P(i) w_i \, K\left(\|x - x_i\| \Big/ (\bar{d}h) \right) \qquad (3)$$

where *$D_P(x)$* is cluster's density;

$I_P(i)$ is a function that equals 1 when the item *i* belongs to cluster *p* and 0 otherwise;

K is the Gaussian kernel function;

h is kernel width;

$\bar{d}$ is the average distance between items, and it is calculated with the help of the following equation:

$$\bar{d} = \frac{2}{n(n-1)} \sum_{i<j} \left\| x_i - x_j \right\| \qquad (4)$$

Using an inductive and interdisciplinary approach, the current research provides an external view on a real-world phenomenon like the use of AI within the educational framework for providing personalized learning paths. The inductive character comes from analysing previously researched phenomenon from different perspectives while the interdisciplinary character is generated by integrating, interacting, and linking theories from the information technology management area with educational management.

3. Main findings

3.1.Descriptive statistics

From a total of 390 articles published in Web of Science during January 2014 – June 2020, only 71 focus on the educational framework and have a look on how AI could be used for developing personalized learning paths. As reflected in Figure 1, the interest in using AI within the educational framework appeared in 2015 and increased slowly by 2019. However, after the COVID period, there is a boom in the scientific production, and the number of articles increased by 77%. Nevertheless, it must be mentioned that the focus of this article is on the general approach of AI and not on specific technologies developed within this area. Different results may have been obtained if ChatCPT would have been included as component of the research context.

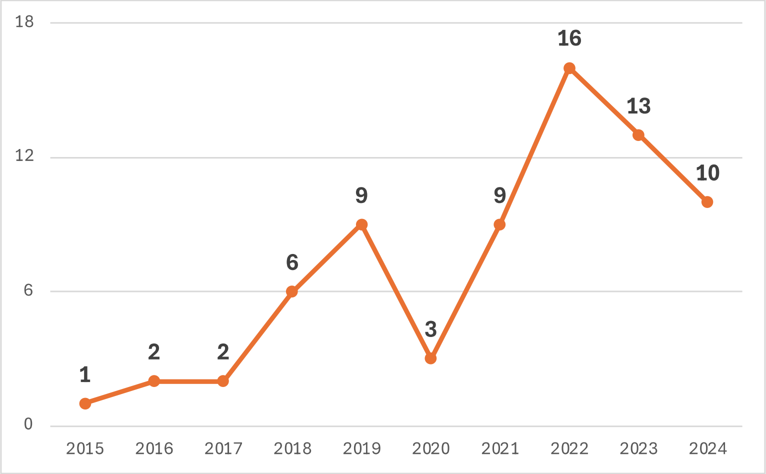


Figure 1. Literature trends: Number of papers published since 2015

Almost 41.5% of the scientific production regarding the use of AI for developing personalized learning paths is concentrated in 10 journals (Table 1). As it can be noticed, only 3 of them belong to the educational field while all the other ones come from the information technology (IT) area. Thus, it can be claimed that the topics related to using AI for developing personalized learning paths are more appealing for the IT researchers than for the educational scholars. The former adopt a wider view and highlight how AI and Internet of Everything can change the educational process and practices while the latter remain focused on the constructivism learning theories and neglect how AI could facilitate the transition towards the implementation of the connectivism learning theories.

Table 1. Articles distribution based on the journals in which they were published

Name of the journal	**Share of published articles**
Sustainability	5,66%
Education and Information Technologies	5,66%
Journal of Intelligent & Fuzzy Systems	3,77%
Soft Computing	3,77%
Scientific Programming	3,77%
International Journal of Emerging Technologies in Learning	3,77%
Artificial Intelligence in Education	3,77%
Mobile Information Systems	3,77%
IEEE Access	3,77%
Knowledge-Based Systems	3,77%

3.2.Cluster and density analysis

Analysing the studies developed so far, two lines of research can be identified (Figure 2). The first one is technology-driven and concentrates on developing a model based on artificial intelligence that can be used within the teaching process for fostering adaptive learning. The focus is on using data, in general, and big data, specifically, for addressing the challenges that come from the external environment. The second one concentrates on learners' needs and characteristics and how these could be addressed by establishing personalized learning paths. Against this backdrop, the blended learning approach is considered, and the classical resources and courses are combined with e-learning, genetic algorithms and platforms like MOOC. Furthermore, the studies included on this category are treated as experiments, using a trial and error approach for showing how AI and intelligent tutoring systems could be used for addressing students' needs and interests.

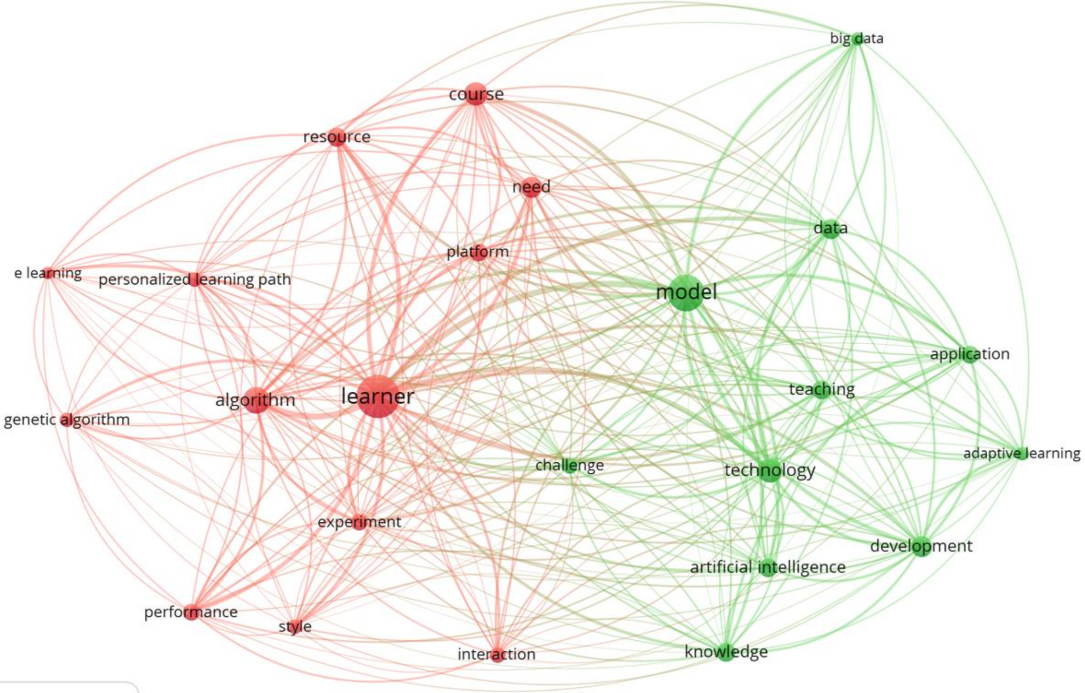


Figure 2. The clusters defining the research regarding the use of AI for developing personalized learning paths

For the cluster density analysis, a red-green-blue colour scheme is used where red highlights the topics that mainly capture academics' attention and blue emphasises the less addressed topics. As reflected in Figure 3, academics focus on addressing learners' needs by providing personalized learning paths based on e-learning and genetic algorithm. On the other hand, further research is required on developing systems-based on artificial intelligence that could support teaching, in terms of content delivery, assessment, and interaction while considering students' learning style.

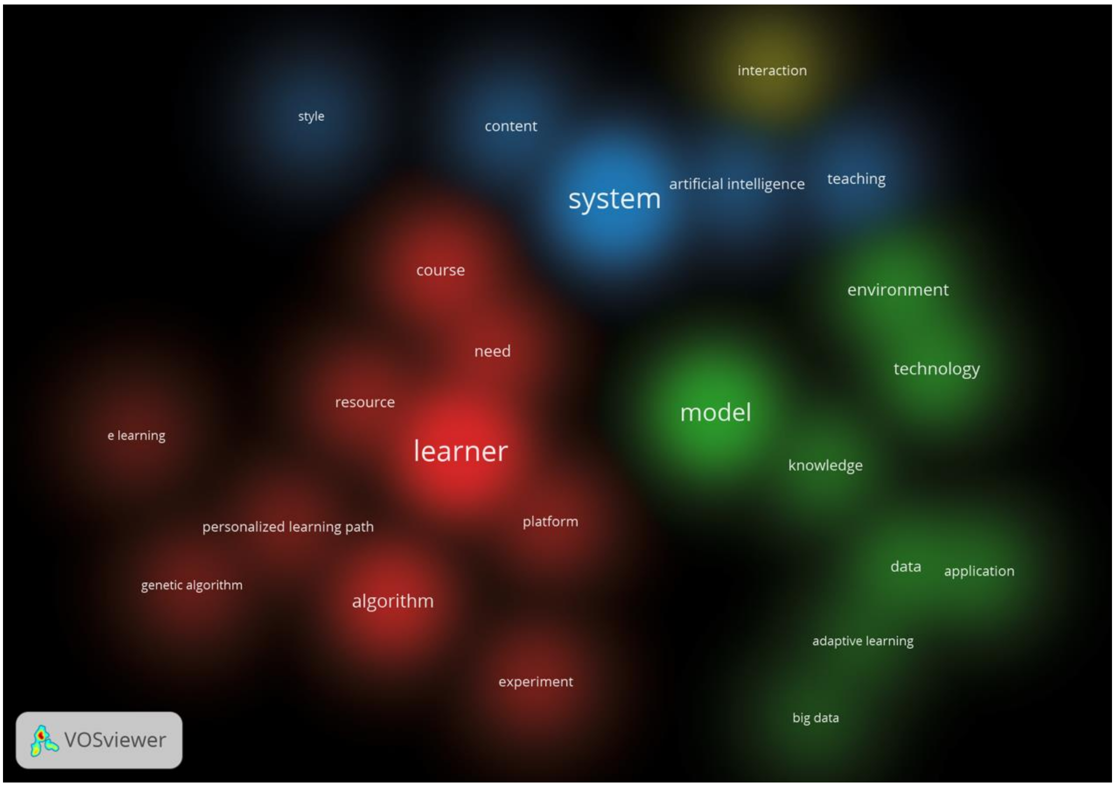


Figure 3. Results of the cluster density analysis

4. Conclusions

Using an inductive and interdisciplinary approach, the current research provides an external view on a real-world phenomenon, like using AI within the educational framework to provide personalized learning paths. Thus, 71 articles published in international journals indexed on Web of Science were selected and analysed since they focused exclusively on the use of AI within the educational framework. The results of the similarity matrix and the mapping technique brought forward a visual representation of the causal linkages established among concepts, while the cluster density analysis emphasized the hot topics that are drawing academics' attention.

One of the paper's main conclusions is that the topics related to using AI for developing personalized learning paths are more appealing to IT researchers than educational scholars. Besides, two lines of research have been identified by analysing the studies developed so far. The first one is technology-driven and concentrates on developing a model based on artificial intelligence that can be used in the teaching process to foster adaptive learning. The focus is on using data, in general, and big data, specifically, to address the challenges of the external environment. The second one concentrates on learners’ needs and characteristics and how these could be addressed by establishing personalized learning paths based on e-learning and genetic algorithms.

Despite these insights, the research is limited by its structural characteristics. It adopted AI from a general approach without considering the use of specific AI tools like ChatGPT and it focused exclusively on the articles indexed on Web of Science. Different results may have been obtained if different concepts have been included in the search process and if databases like Scopus, Sage or Emerald would have been considered. Hence, further research is required to develop AI-based systems that could support teaching in terms of content delivery, assessment, and interaction while considering students' learning styles.

Acknowledgment

This research and publication are supported by the ERASMUS + EU Project CODEMO, CoCreative Decision Makers for 5.0 Organisations, Project Nr. 101104819, Co-funded by the European commission.

References

[1] Mukherjee, A.A., Raj, A., & Aggarwal, S. (2023). Identification of barriers and their mitigation strategies for Industry 5.0 implementation in emerging economies. International Journal of Production Economics, 257, 108770.

[2] Leon, R.D. (2023). Employees' reskilling and upskilling for Industry 5.0: Selecting the best professional development programmes. Technology in Society, 75, 102393.

[3] Farrow, E. (2022). Determining the human to AI workforce ratio – Exploring future organizational scenarios and implications for anticipatory workforce planning. Technology in Society, 68, 101879.

[4] Lett, J. (1990). Emics and etics: Notes on the epistemology of anthropology. In T.N. Headland, K.L. Pike, & M. Harris (Eds.). Emics and etics: The insider/outsider debate. Frontiers of anthropology (pp.127-142). Newbury Park, California: Sage Publications.

[5] Adams, R., Jeanrenaud, S., Bessant, J., Denyer, D., & Overy, P. (2016). Sustainability-oriented innovation: a systematic review. International Journal of Management Reviews, 18(2), 180-205.

[6] Duriau, V.J., Reger, R.K., & Pfarrer, M.D. (2007). A Content Analysis of the Content Analysis Literature in Organization Studies: Research Themes, Data Sources and Methodological Refinements. Organizational Research Methods, 10(1), 5-34.

[7] van Eck, N.J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84, 523-538.

DOI: 10.2478/ijasitels-2024-0002

Implementation of SDR as a Vector Network Analyzer: Methodology, Configuration, and Software Insights

Fabian Török[1],

[1]Department of Computer Science and Electrical Engineering, Lucian Blaga University of Sibiu,4 Emil Cioran, Str., 550025, Sibiu, Romania

fabian.torok@ulbsibiu.ro

Abstract

Software-Defined Radios (SDR) offer unparalleled reconfigurability across a wide frequency spectrum, making them versatile tools in the RF domain, especially when capable of simultaneous data transmission and reception. This adaptability positions SDRs as cost-effective universal RF measurement devices. Leveraging these advantages, this article explores the utilization of an SDR as a Vector Network Analyzer (VNA), emphasizing an offset tuning measurement approach to mitigate negative effects inherent in SDR architectures. The research places particular attention on the software and measurement aspects of SDR-based VNA, distinguishing itself from existing recent research that predominantly concentrates on hardware setup and calibration. The study refines and generalizes existing approaches while presenting a universal methodology for VNA realization, for any SDR architecture with at least 1 full duplex transmit and receive port. By addressing the often-overlooked software aspects and proposing a novel measurement approach, this research provides a foundation for implementation, with the potential to significantly contribute to the evolution of SDR-based RF measurement methodologies.

Keywords: Software-Defined-Radio, Vector-Network-Analyzer, Configuration, Offset-Tuning

1 Introduction

VNAs characterize network properties in function of frequency using Scattering-Parameters (S-Parameters). These S-Parameters describe how signals of different frequencies are amplified/attenuated and their phase shifted when passing through a network. Vector-Network-Analyzers have come a long way since their initial appearance on the market in the 1960s [8]. However, they still remain pricy pieces of measurement equipment, as standalone entry price models from reputable manufacturers (e.g.: Rohde & Schwarz, Keysight) with a relatively low frequency bandwidth start above 5000€. The price of higher-performance VNAs can easily rise into the tens of thousands of euros [1].

In recent research SDRs have been utilized to realize functionalities of multiple types of measurement equipment, such as Spectrum Analyzer [11], [10]; Optoelectronics [14]; as well as VNA [16], [13], [12]. This proven versatility of Software-Defined Platforms enables them to be used as universal low-cost measurement equipment. The added value in this case is not that the SDR achieves more precise measurements than the dedicated alternatives, but that it can be easily reconfigured to replace several measuring devices. The savings potential of a universal SDR measuring device is therefore enormous.

This article explores the implementation of a VNA using a SDR with particular attention to the configuration of the SDR and the methodology of the measurement with the software aspects involved. While the measurement approach in its essence is valid for any SDR with at least one transmit and receive full duplex ports, the LimeSDR mini is utilized as an example SDR model. Initially some SDR fundamentals are discussed and the features of the LimeSDR mini are presented. Subsequently additional hardware components, as well as the calibration of the system is discussed. This section is kept rather short, as other recent research covers this sufficiently [13], [12]. Finally, considerations for the SDR configuration is presented, taking into account the settings of the sample rate, transmit power levels, adequate tuning of receive power, filter settings as well as the interpretation of I/Q data to obtain a measurement with as little noise as possible. As the approach of a SDR VNA is not yet commercialized [8], the known weaknesses of the heterodyne and homodyne SDR architectures are mitigated in the proposed measurement method.

2 SDR Theory and Example SDR Used

2.1 SDR Theory

The basic principle behind SDR is to digitalize the signal processing of signals as near to the antenna as possible in a functional sense. The ideal SDR would consist of an antenna, and Analog-Digital-Converter (ADC) and a processing unit in which signal processing is executed. Due to sample rate limitations in currently available ADCs this is not practical. The typical receiving frequency range of such a system would range from 9 kHz to 55 MHz when respecting Nyquist's sampling theorem. Thus, additional analogue frontend RF hardware, usually pre-selector filters, I/Q mixers, and optionally frequency converters are required to extend the operation range [9].

Two established SDR architectures are the heterodyne and homodyne architectures. The heterodyne architecture, schematically depicted as a receiver in Figure 1, performs frequency conversion in two stages, where the first conversion stage converts the signals to an intermediate frequency (IF) band. The second stage converts the signals to baseband using an I/Q mixer. In comparison to the homodyne architecture, the heterodyne architecture may achieve superior filter selectivity, system sensitivity and system stability. In contrast to the heterodyne architecture, the homodyne architecture directly down converts signals to baseband and thereby features no IF frequency. This architecture is therefore often also referred to as Zero-IF. Figure 2 depicts this architecture for the receiver side. Generally, this architecture allows for a smaller form factor, lower cost, power consumption and a better image rejection ratio due to a single mixer stage [2]. Both SDR architectures struggle to varying degrees with LO Leakage which manifests itself as a DC Offset, observable as a spike at 0 Hz in the frequency domain. Image rejection techniques are required for both architectures, as the mixers involved in frequency conversion generate images from the signals to be transmitted or received. [9].

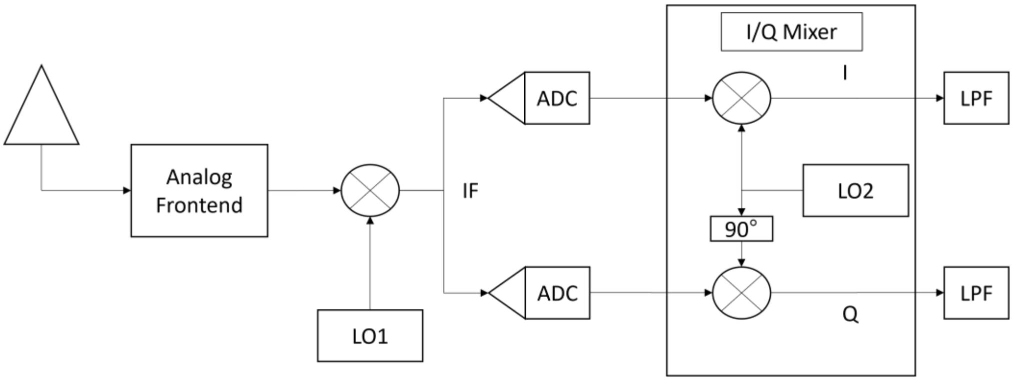


Figure 1 Heterodyne Receiver Architecture

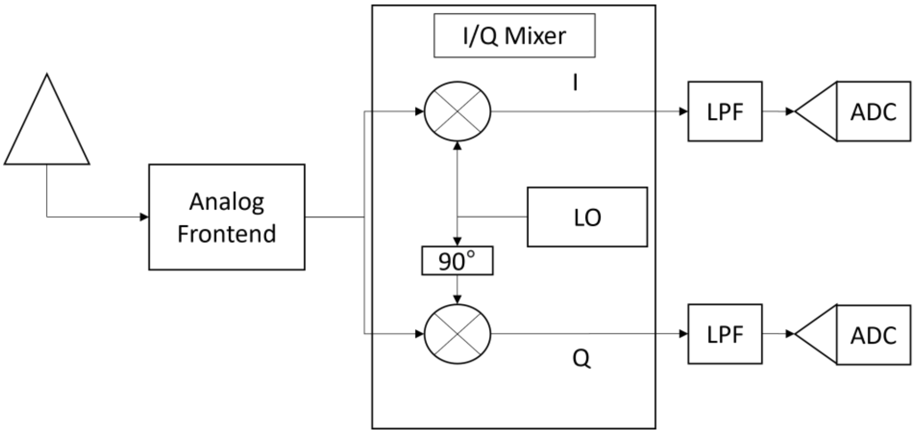


Figure 2 Homodyne Receiver Architecture

2.2 Example SDR

The example SDR referred to in this article is the LimeSDR mini v1. It based on a homodyne architecture, features a frequency range from 10 MHz to 3.5 GHz, a maximum sample rate of 30.72 MHz and a 12-bit ADC operating with up to 0.8 Vpp. It features one receive port and one transmit port which can be tuned independently of each other [17].

3 Hardware Setup and Calibration

While S_{21} and S_{12} measurements can be made in straight forward fashion by connecting the device under test (DUT) between the SDR transmit and receive port as shown in Figure 3, S_{11} and S_{22} measurements require additional hardware as they measure the reflected wave from the DUT [8], [13]. A directional coupler used as a reflectometer is used to measure the reflected wave from the DUT as seen in Figure 4. A 10dB attenuator can be added to reduce variation in impedance [12].

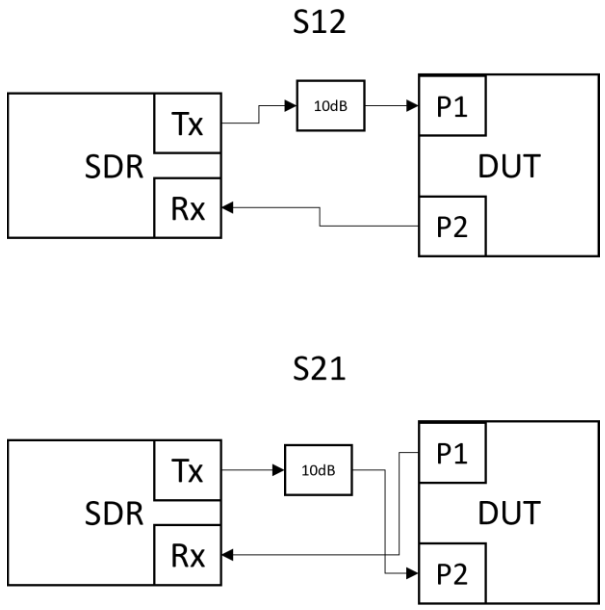


Figure 3 Hardware Setup for S12 and S21 Measurements

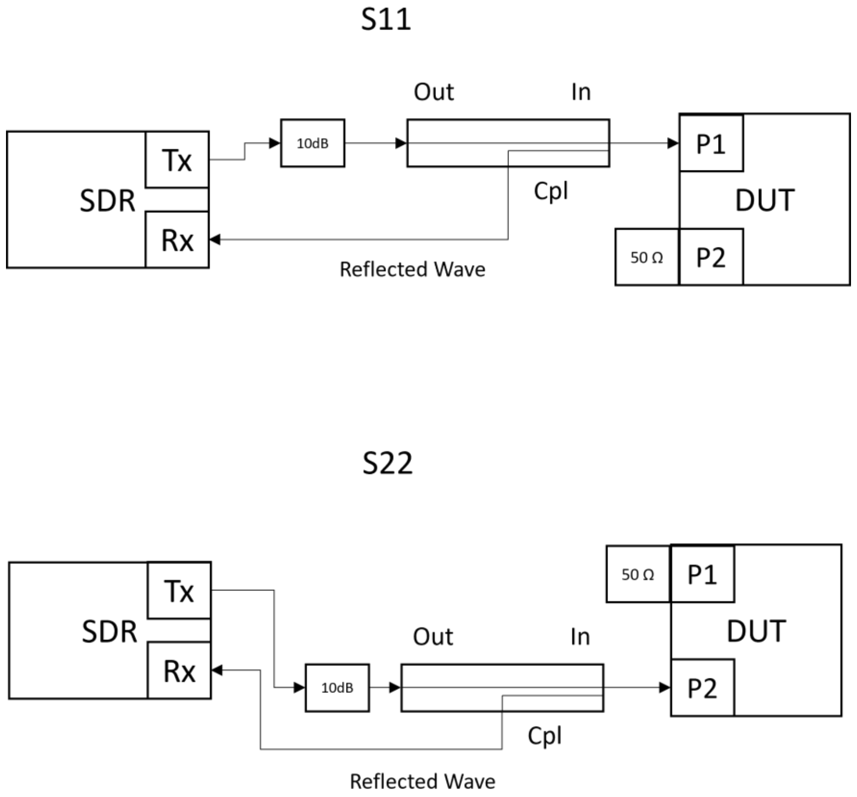


Figure 4 Hardware Setup for S11 and S22 Measurements

Calibration is a necessary operation that should be conducted before executing measurements. It captures changes which are induced by additional hardware components in the measurement setup and age-related degradation in the active components of the system. [15] A fleshed out SDR VNA should include an automated calibration routine. This however is not subject of this article. In terms of calibration recent research has found the application of the Short-Open-Load (SOL) [12] and

Through-Reflect-Line (TRL) [13] calibration standards to be accessible and effective calibration methods for a SDR VNA.

Alternatively, a commercial RF measurement equipment such as a power meter or spectrum analyzer can be used to calibrate the SDR VNA. It is important to note that two calibrations must be conducted: One for S_{12} and S_{21} measurements, and another for S_{11} and S_{22} measurements.

For the calibration of S_{12} and S_{21} measurements, the SDR should transmit the same signal that will be used for measurements. The calibration needs to be done in steps over the whole frequency range intended for measurements. The steps size can be chosen in function of the desired frequency accuracy of the SDR VNA. The RF measurement equipment of choice is used to receive the power level and phase of the signal instead of the SDR receive port. Instead of the DUT the two harnesses can be connected using an appropriate adaptor (this however will induce a small error into the calibration), or the calibration can be done twice: For the equipment between transmit port and the DUT, and once for the equipment between the DUT and the receive port.

For the calibration of S_{11} and S_{22} measurements, the same signal type as described above must be used. The calibration must be conducted once from the SDR transmit port to the DUT, with the coupled port of the directional coupler being terminated with a 50 Ω termination. The calibration must then be conducted with the SDR transmit port being connected instead of the DUT and the signal measured at the coupled port of the directional coupler. Here as well, all unconnected ports must be terminated with a 50 Ω termination. These two calibrations together characterize the path for the S_{11} and S_{22} measurements.

The obtained attenuations and phase shifts from the calibration measurements can then be interpolated in function of the frequency steps and the resulting calibration graphs or functions must be stored as they need to be available for further use in the software during measurements.

It is worth noting that most SDRs also offer internal calibration options to minimize the receive and transmit DC spike and to increase image rejection as shown in Figure 5 and Figure 6. An example how this is done for LimeSDR using phase and gain correction of the I/Q signals is shown in [3].

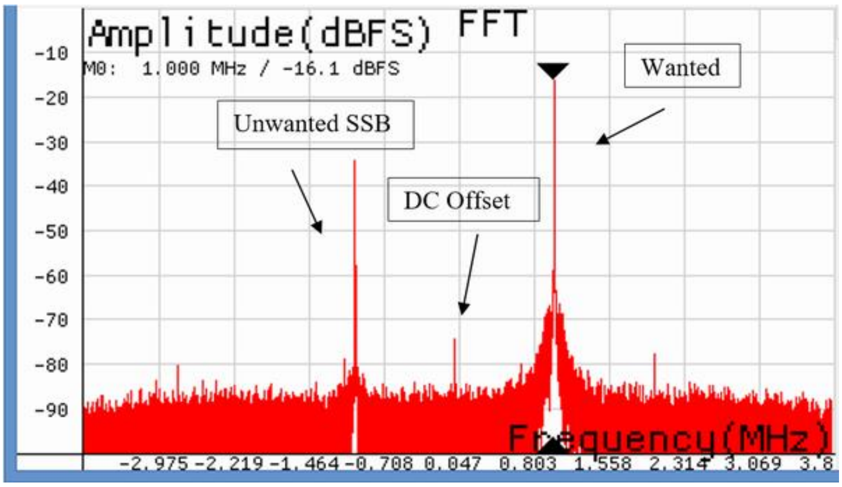


Figure 5 Receive Spectrum of a CW with 1 MHz Offset Before Internal Calibration [3]

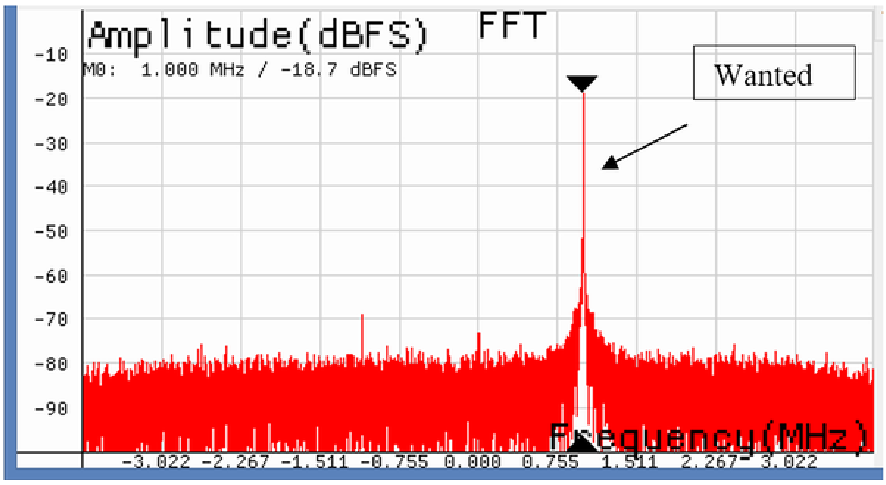


Figure 6 Receive Spectrum of a CW with 1 MHz Offset After Internal Calibration [3]

4 Measurement Approach, SDR Configuration, and Processing of I/Q Data

4.1 SDR Configuration

The SDR configuration before a measurement encompasses the definition of the used sample rates, transmit and receive gain, transmit and receive filter configurations and the I/Q frame size expected to be received from the SDR. The proposed measurement method uses an offset tuned continuous-wave (CW) signal which is transmitted by the SDR. Since only a single signal is present during the measurement, generated images of the transmitted signal can easily be ignored in the signal processing on the receiving side of the SDR, given that the DUT does not frequency shift the transmitted signal. The DC spike caused by LO Leakage could interfere with a CW located too close to the center frequency which the transmitter is tuned to. Thus, it is proposed to use a 1 MHz offset tuned CW as transmit signal.

The CW signal offset has an impact on the required sample rate and frame size of I/Q data. To properly sample the CW, the sample rate must be at least twice the CW frequency offset. If possible, the signal should even be oversampled. In this 8x oversampling is chosen, which would be 8 MHz. The sample rate of the receive and transmit channels can be the same. The frequency bin size achievable after the Fast-Fourier-Transform (FFT) depends on the sample rate and frame size as shown in (1). If the frequency width of one frequency bin should be as close as possible to 15 kHz, but still be a power of 2 to retain FFT conversion speed, the frame size 512 may be chosen. The frequency bin size of 15 kHz is chosen to fit the CW signal inside of a single bin with some margin for error. This will ease the signal processing in software.

$$F^{Binspan} = \frac{F_{Samplerate}}{N_{Framesize}} \tag{1}$$

4.2 Measurement Methodology

Figure 7 shows control flow of the proposed measurement methodology. It consists of the initialization which contains the SDR configuration and a measurement loop. The measurement must be repeated for each S-Parameter measurement. Two special buffers are used in this measurement method. The Calibration Data Buffer contains the frequency dependent calibration data. This buffer can either be an array containing calibration data at discrete frequency steps of the calibration, or it could be a function representing the interpolation of the calibration over the previously mentioned frequency steps. The Measurement Result Buffer contains the measurement results and is filled incrementally with each frequency step during the measurement.

It should be noted, that when switching from S_{21} and S_{12} measurements to S_{11} and S_{22} measurements, the hardware setup should be changed as described in chapter 3. Furthermore, also the calibration data required to be used differs. The signal processing steps highlighted in Figure 7 are described in chapter 4.2.1 to chapter 4.2.5.

4.2.1 Signal Processing Step A

Averaging of I/Q data frames helps to smooth out short-term fluctuations in the measurement at the cost of measurement duration. Averaging can be implemented using two buffers with the length of the I/Q data frame. One buffer is used to store newly received data from the SDR, while the other buffer stores the continuously updated averaging result.

4.2.2 Signal Processing Step B

Conversion from the time domain into the frequency domain is proposed to be managed using the Fast-Fourier-Transform (FFT) algorithm. To mitigate artefacts which are generated through the FFT of a finite data set, before converting to the frequency domain, a window function (e.g. Hanning) is applied to the time domain signal [6]. While it is possible to implement the FFT manually, there are several tested and proven libraries already implementing the FFT. An example is the fftw3 library which implements the FFT for C and Fortran [4]. The exemplary usage of fftw3 in a LimeSDR Spectrum Analyzer application with C/C++ can be seen in [5].

4.2.3 Signal Processing Step C

To compute the phase of the CW signal on the receiving side of the SDR the correct bin of frequency domain I/Q data needs to be identified. Figure 8 shows the time domain and frequency domain plot of a I/Q data frame size of 512. It contains a 1 MHz CW sampled at 8 MHz. Utilizing (1) and keeping in mind that the first bin of the FFT result is DC, it can be concluded that the 65th bin contains is the I/Q data bin to be used for obtaining the phase of the CW signal through (2). The phase offset captured during the calibration will be corrected for in step E.

$$\phi = arctan(\frac{Q}{I}) \tag{2}$$

4.2.4 Signal Processing Step D

The power measurement requires the I/Q frame to be in the frequency domain as well. The amplitude shown in Figure 8 Time Domain and Frequency Domain Plot of a Complex Sine does not yet display the power in an SI-unit. To obtain the amplitude in Vrms, (3) must be applied, where A is the time domain I/Q frame and N is the frame size [1 Cerna NI]. S is the ADC step size in volt. It must be taken into account that the I/Q data sampled by the ADC is usually relative to the ADCs range. For the LimeSDR this means that an I/Q amplitude equaling the maximum 12-bit integer value correlates to the maximum voltage the ADC can represent, in this case 0.8 Vpp.

$$Amplitude_{V\,rms} = S * \frac{\sqrt{I[FFT(A)]^2 + Q[FFT(A)]^2}}{N} \tag{3}$$

To obtain the power of the CW signal in dBm the amplitude of the respective bin must be inserted in (4), where R is the system impedance, usually 50 Ohm.

$$Power_{dBm} = 10 * log_{10}(\frac{\frac{Amplitude^2_{Vrms}}{R}}{0.001}) \tag{4}$$

The difference between the obtained power level and the power level of the transmitted signal constitutes the attenuation or gain of the currently measured S-Parameter. The gain offset captured during the calibration will be corrected for in step E.

4.2.5 Signal Processing Step E

Ultimately for each frequency step, the calibration offsets for phase and gain must be corrected in the measured S-Parameters. These can simply be added to the measurement results obtained in chapter 4.2.4 and 4.2.3.

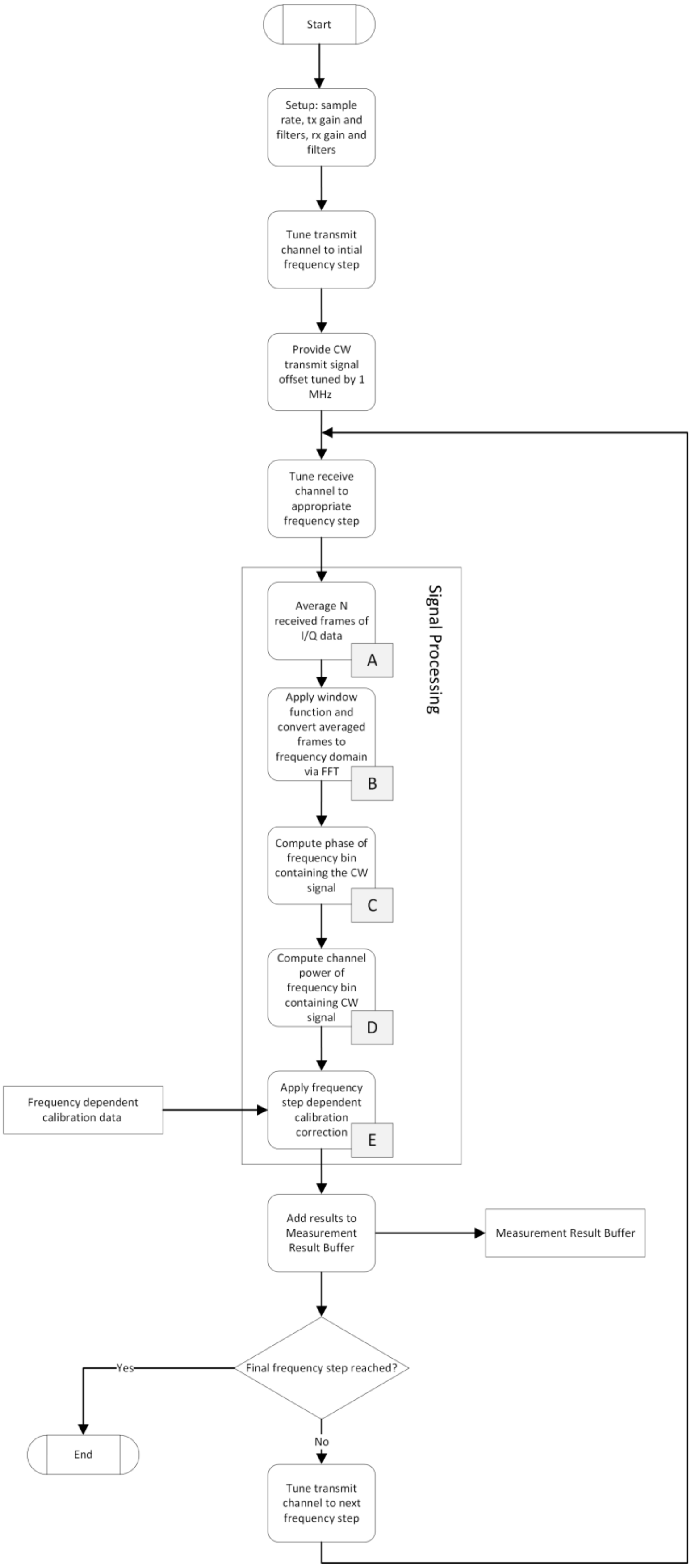

Start
Setup: sample rate, tx gain and filters, rx gain and filters
Tune transmit channel to intial frequency step
Provide CW transmit signal offset tuned by 1 MHz
Tune receive channel to appropriate frequency step
Signal Processing
Average N received frames of I/Q data
A
Apply window function and convert averaged frames to frequency domain via FFT
B
Compute phase of frequency bin containing the CW signal
C
Compute channel power of frequency bin containing CW signal
D
Frequency dependent calibration data
Apply frequency step dependent calibration correction
E
Add results to Measurement Result Buffer
Measurement Result Buffer
Final frequency step reached?
Yes
End
No
Tune transmit channel to next frequency step

Figure 7 Measurement Methodology

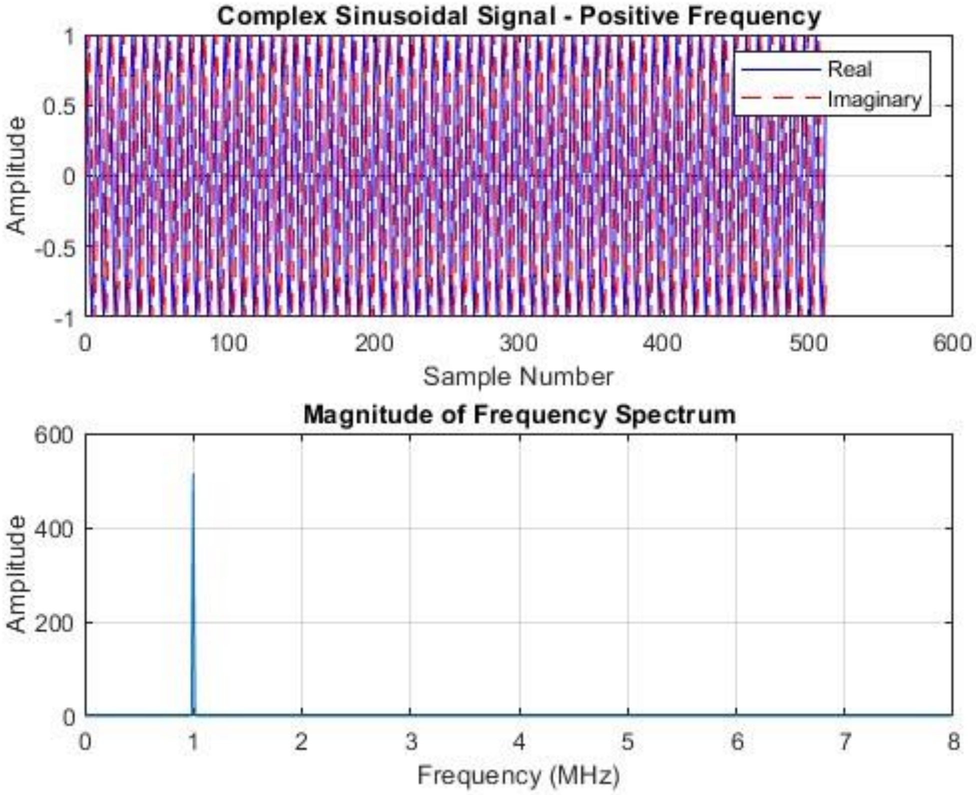


Figure 8 Time Domain and Frequency Domain Plot of a Complex Sine

5 Conclusions

The presented measurement methodology and SDR configuration goes beyond the hardware-related insights that recent research offers on the implementation of a SDR VNA. Based on the findings of this research it is possible to derive a software implementation using any of the common implementation tools such as MATLAB & Simulink, GNU Radio or in C. The proposed method is designed with known weaknesses present in current SDR architectures in mind, to offer S-Parameter measurements with as few noise as possible.

6 References

[1] Farnell, “farnell.com,” [Online]. Available: https://de.farnell.com/w/c/messtechnik/hf-messgerate/vektor-netzwerkanalysatoren/prl/ergebnisse?st=vektor-netzwerkanalysator&gs=true&sort=P_PRICE. [Accessed 05 01 2024].

[2] National Instruments, “What Is the RF Architecture Used on USRP-RIO?,” [Online]. Available: https://knowledge.ni.com/KnowledgeArticleDetails?id=kA00Z0000019Y9TSAU&l=en-US. [Accessed 06 01 2024].

[3] Myriad RF, “Getting Started with LimeSDR-USB and LimeSuiteGUI,” 29 08 2017. [Online]. Available: https://wiki.myriadrf.org/Getting_Started_with_LimeSDR-USB_and_LimeSuiteGUI. [Accessed 06 01 2024].

[4] FFTW, “fftw.org,” [Online]. Available: fftw.org. [Accessed 07 01 2024].

[5] F. Török, “Lime SDR Spectrum Analyzer,” 2022. [Online]. Available: https://github.com/fabske98/Lime-SDR-Spectrum-Analyzer. [Accessed 07 01 2024].

[6] LDS-Group, “Aplication Note ANO14 - Understanding FFT Windows,” 2003.

[7] M. Cerna and A. F. Harvey, “Aplication Note 041 - The Fundamentals of FFT-Based Signal Analysis and Measurement,” National Instruments, 2000.

[8] M. Sayed and J. Martens, Modern RF and Microwave Measurement Techniques - Vector network analyzers, Cambridge University Press, ISBN 9781139567626, 2013.

[9] H. L. Hartnagel, R. Quay, U. L. Rohde and M. Rudol, Fundamentals of RF and Microwave Techniques and Technologies, Springer, ISBN 978-3-030-94100-0, 2023.

[10] P. Flak, “Hardware-Accelerated Real-Time Spectrum Analyzer With a Broadband Fast Sweep Feature Based on the Cost-Effective SDR Platform,” IEEE Acess, DOI: 10.1109/ACCESS.2022.3215800, 2022.

[11] P. V. A. Freitas, R. F. Hanthequeste, G. B. A. Orofino, P. V. G. Castellanos, A. A. C. Canavitsas and R. C. Bentes, “Implementation of a Spectrum Analyzer Using the Software-Defined Radio Concept,” *Journal of Microwaves, Optoelectronics and Electromagnetic Applications,* vol. 20, no. 4, DOI: 10.1109/ACCESS.2022.3215800, 2021.

[12] M. Nazrin, S. J. Hashim, F. Z. Rokhani, B. M. Ali and Z. Yusoff, “Error correction and uncertainty measurement of short-open-load calibration standards on a new concept of software defined instrumentation for microwave network analysis,” DOI: https://doi.org/10.21595/jme.2019.20329 , 2019.

[13] A. Raza, A. Jabbar, D. A. Sehrai, H. Atiq and R. Ramzan, “SDR Based VNA for Characterization of RF Sensors and Circuits,” in *1st International Conference on Microwave, Antennas & Circuits,* Islamabad, Pakistan, DOI: 10.1109/ICMAC54080.2021.9678273, 2021.

[14] L. M. Riobó, F. E. Veiras, M. T. Garea and P. A. Sorichetti, “Software-Defined Optoelectronics: Space and Frequency Diversity in Heterodyne Interferometry,” *IEEE Sensors Journal,* vol. 18, no. 14, DOI: 10.1109/JSEN.2018.2842143, 2018.

[15] A. Rumiantsev and N. Ridler, “VNA Calibration,” *IEEE Microwave Magazine,* no. DOI: 10.1109/MMM.2008.919925, 2008.

[16] M. I. Vidotto, F. E. Veiras and P. A. Sorichetti, “Software defined radio for vector network analysis: Configuration, characterization and calibration,” *Measurement,* vol. 189, no. DOI: 10.1016/j.measurement.2021.110468, 2022.

[17] Lime Microsystems, “LMS7002M Datasheet 3.1r00,” [Online]. Available: https://eu.mouser.com/datasheet/2/982/LMS7002M-Data-Sheet-v3.1r00-1600568.pdf. [Accessed 07 01 2024].

DOI: 10.2478/ijasitels-2024-0013

A Systematic Method of Generating Intuitive Bivariate Colour Legends

Mihai NEGHINĂ, Cătălina NEGHINĂ

Department of Computer Science and Electrical Engineering,

Faculty of Engineering, University "Lucian Blaga" of Sibiu, Sibiu, Romania

Corresponding author: Mihai NEGHINĂ, e-mail: mihai.neghina@ulbsibiu.ro
Cătălina NEGHINĂ, e-mail: catalina.neghina@ulbsibiu.ro

Abstract. Colour is used increasingly often as quantitative indicator of combined continuous features such as temperature and humidity on geographical maps, or deformity on two axes in medical diagnosis. As such, generating flexible bivariate legends with particularly tough colour constraints for specific combinations of the variables has come to the forefront. The paper presents a systematic method of generating bivariate legends by first selecting a convenient colour subspace and then adjusting it through image processing techniques. The proposed method can generate a wide variety of legends with smart choices of parameters but remains simple enough to be worthy of consideration by specialists (in various fields) whose representations may benefit by becoming better intelligible. These legends can be used for pseudo-colouring of maps, diagrams, and many other images and graphical representations. An example is also provided for a comprehensive colouring to be used in the medical diagnosis of spine deformity.

Keywords: bivariate maps, pseudo-colouring, interpolation, warping

1. Introduction

Colour has had an important role in human evolution and can still be extremely influential in defining specific anthropocentric environments, summarizing information in memorable ways or inspiring associations between abstract concepts. Colour has been defined as a property of objects long before it was well understood [1]. The first known theory of colour belongs to Aristotle, who believed that light originated in the eye, that all colours came from lightness and darkness and that they were related to the four elements. Ptolemy also briefly commented on colours in his book on geometrical optics [2]. During the Islamic Golden Age, scholars like al-Kindi, al-Haytham, ibn-Sina and al-Tusi challenged Aristotle's view on the source of colour by emphasizing the role of external light [2][3]. Starting from the Early Renaissance, European painters and craftsman described a more practical theory based on paints, which eventually lead to the adoption of Red-Yellow-Blue (RYB) as the primary colours, still relevant in arts [4]. Although recognizing light as an external phenomenon, Galileo was inclined to place the existence of colour not outside, but rather within the mind of the observer [5]. However, with the prism experiments on splitting and recombining white light from Newton's Opticks and the thorough description of coloured light as a sliver of the electromagnetic spectrum, colour was firmly in the physics realm, described in terms of hue/frequency, saturation and brightness [6]. Yet, some colours such as *magenta* were not to be found in the rainbow spectrum either as single waves or as clusters of waves with similar frequencies. It was only after a more detailed analysis of the human (and animal) vision during the 19th and 20th centuries that the observer's importance was recognized once again, which in turn lead to the development of the colour system having primary colours Red-Green-Blue (RGB). Beyond that, artificial vision systems or data collecting devices (including satellites and space probes) are not limited to the visible spectrum, nor to the constraint of having only 3 (relatively wide) frequency bands.

Back in the 18th century, Goethe contested Newton's approach, arguing that colour is more than just a scientific measurement [7]. It is as much a physical property of light and human perception, as it is a subjective experience interpreted within a psychological frame. His view is still widely adopted by artists and producers of visual media. Colour symbolism, defined as subjective meanings of colours and the abstract concepts attached to them, varies between cultures, historic times and contexts. Therefore, the choice of colour for marketing purposes or for representing data in maps, interfaces and presentations might be critical for ensuring a good communication and retention of the contents, as well as improving analysis and usage by providing an intuitive frame of colour references [8]. In images and graphs, colour is used as a marker or emphasizer in three broad ways [9]:

- to separate discrete categories, such as assigning distinct colours to neighbouring countries on political maps or differentiation the members of the European Union by the joining year; for class discrimination, the colours just need to assure a good contrast and are otherwise free choices to convey additional meanings. Because there is no direct correlation between the colours of the classes, which are presented as a list of disjoint elements, the colour legend has no dimensionality.
- to represent sequential values of a feature in quantized levels, such as in simpler physical maps of the oceans where deeper shades of blue suggest larger bathymetric depths but using only a small number of shades overall (Fig. 1.b) or the percentage of populations (Fig. 1.c). Compared to the previous category, sequential colouring has a further constraint that the sequence of shades needs to be intuitively ordered (in both directions for 2D legends). Images containing this sort of legend on a geographic map fall under the common name of choropleth maps. Various applications have been developed in recent years to assist with building custom choropleth legends [10-14].
- to quantitatively represent the strength or intensity without any noticeable quantization, such as maps of temperature (as an example of 1D legend) or temperature-humidity (as an example of 2D legend). The constraint is that the colour gradient needs to be smooth, the colour legend becoming itself a colour map.

Quantitative 1D colour legends (palettes) are relatively easy to create from the natural order of hues in the spectrum, by creating lighter/darker shades of a primary colour or by concatenating two such ranges of shades. When more than one variable needs to be conveyed intuitively, the legend becomes bidimensional. Generating smooth gradients in every direction with constraints on specific points is useful when mapping sets of two independent continuous variables is useful in many fields, not just for geographical data [15].

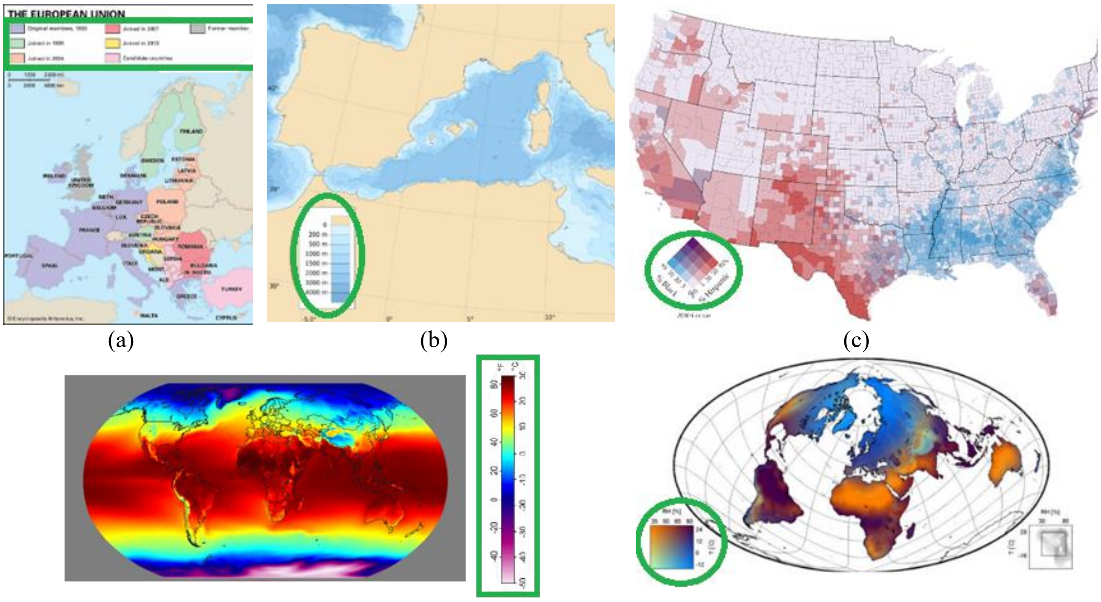


Figure 1. Examples of representation of geographical data with legends from each category
(a) European Union joining year (©Encyclopaedia Britanica)
(b) Western Mediterranean Sea bathymetry map [27]
(c) Distribution of Black and Hispanic populations in continental US [28]
(d) Annual average temperature ([29])
(e) Distribution of climate zones based on average temperature and relative humidity [20]

As the intent is to visualize the combined output, each variable combination must have a distinct colour, while at the same time very specific combinations of the variables must have a predetermined colour.

Fig. 1 shows examples of representation of geographical data with legends from each category. Please note that the focus is on the legend, marked by encirclement, not on the accuracy of data presented in the maps.

Attempts to build bivariate colour legends have been successfully proposed in the past, but had shortcomings such as having explicit granularity [16][17], limited flexibility [18] or limited applicability [19]. These methods are still useful for applications with a very small number of legends required and with softer constraints on the colours.

A custom bivariate colour map can be created even without a specialized application. Indeed, most historical legends have been created in this way, strictly in relation to the data represented, be they about geographical variables on maps [20], bioinformatics [21], medical diagnosis [22], engineering [23] or other fields.

In this paper, a more general framework is presented. The proposed method of generating bivariate colour legends is flexible enough to allow the generation of the legends described in the mentioned bibliography with smart choices of parameters, but still simple enough to be worthy of consideration by specialists (in other fields) whose representations may benefit from such legends.

The remainder of this paper is organized as follows: *Section 2* presents the proposed method, emphasizing the possible parameter choices. *Section 3* briefly discusses one-dimensional colour palettes within the framework of the method. *Section 4* exemplifies the generation of bivariate colour legends by describing an intuitive representation of spine diagnosis. The concluding remarks on the flexibility of the method and perspectives are explored in *Section 5*.

2. Proposed method

The proposed method is comprised of four steps, as shown in Fig. 2:

- selection of the colour space
- selection of constraint points
- interpolation
- warping

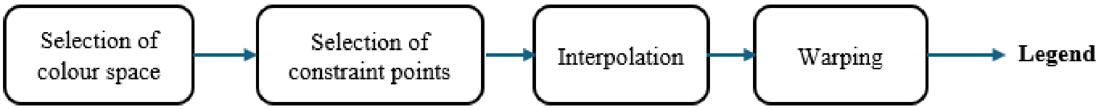
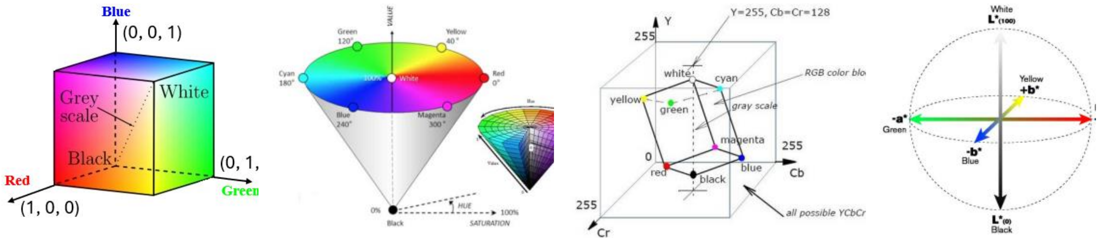


Figure 2. Diagram of the proposed method

2.1. Selection of the colour space

Generating colour legends with the explicit purpose of being used in graphic representations (meant to be visualised) restricts the colour space to finite three-dimensional linear spaces [24] such as RGB, HSV, YCbCr, CIE La*b* etc., in which every colour can be represented by a three-valued vector. Fig. 3 shows the mentioned colour spaces side by side. The spaces can be considered continuous from the perception of the human eye, although in practice they are used in digital environments more as discrete spaces with a very fine quantization.



RGB color space

HSV color space

YCbCr color space

CIE La*b* color space

Figure 3. Colour spaces [26]

The sum of two vectors from these spaces is either another colour, or outside the boundaries of the colours. In RGB, for instance, adding *green* to *red* generates *yellow* ($[1,0,0]^T + [0,1,0]^T = [1,1,0]^T$), but adding *magenta* to *red* is outside the boundaries of the colour cube. Most of the examples in the next section will use RGB as the space of choice, out of which the bivariate legends will be extracted.

2.2. Selection of constraint points

The selection of constraint points is dependent on the contents presented and the desired colour symbolism. For physical features in maps, it is customary to have shades of *green* for planes, *brown* for mountains and *blue* for the seas. For temperature, *red* and *orange* are usually considered hot, while *blue* is considered cold. For the general assessment of situations, the semaphore conventions indicate *red* as dangerous or imperative, *yellow* as worrisome or cautionary, *green* as fine or harmonious and *blue* as excellent or exceptionally well. For internal states, *red* is associated with passion or anger, *yellow* with either optimism or energy and so on.

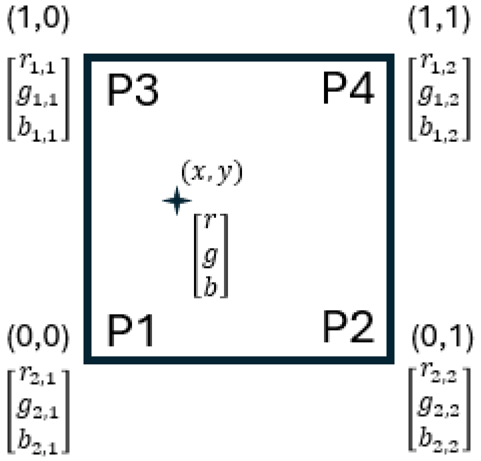

(1,0)
$\begin{bmatrix} r_{1,1} \\ g_{1,1} \\ b_{1,1} \end{bmatrix}$
(1,1)
$\begin{bmatrix} r_{1,2} \\ g_{1,2} \\ b_{1,2} \end{bmatrix}$
P3
P4
(x, y)
$\begin{bmatrix} r \\ g \\ b \end{bmatrix}$
(0,0)
P1
P2
(0,1)
$\begin{bmatrix} r_{2,1} \\ g_{2,1} \\ b_{2,1} \end{bmatrix}$
$\begin{bmatrix} r_{2,2} \\ g_{2,2} \\ b_{2,2} \end{bmatrix}$

Figure 4. Representation of the legend in terms of geometrical coordinates and colour vectors

Table 1. Correspondence between description, geometrical coordinates and colour vectors

Corner position	**Coordinates (x, y)**	**Colour (r, g, b)**
Lower left (P1)	(0,0)	$\begin{bmatrix} r_{2,1} \\ g_{2,1} \\ b_{2,1} \end{bmatrix}$
Lower right (P2)	(0,1)	$\begin{bmatrix} r_{2,2} \\ g_{2,2} \\ b_{2,2} \end{bmatrix}$
Upper left (P3)	(1,0)	$\begin{bmatrix} r_{1,1} \\ g_{1,1} \\ b_{1,1} \end{bmatrix}$
Upper right (P4)	(1,1)	$\begin{bmatrix} r_{1,2} \\ g_{1,2} \\ b_{1,2} \end{bmatrix}$

The number of constraint points can also vary. For univariate legends, two or three constraints (the edges and maybe the middle colour) are usually enough. For bivariate legends, the simplest maps use only four corner points, but the number can also be increased for more specific colours at various positions. Without loss of generality, we can assume that 2D representation of the legend is a square with side 1 with geometric coordinates and colours indicated in Fig. 4 and Table 1, and an interior point at coordinates (x, y) will be coloured $[r,\ g,\ b]^T$.

2.3. Interpolation

Probably the most intuitive computation of inner point for all legends is linear/bilinear interpolation, which will be used in most examples from the next sections. Under bilinear interpolation, four coplanar corner constraints lead to legends contained into slices of the colour space, while four non-coplanar corner constraints lead to shapes similar to a twisted sheet.

2.3.1. Four (corner) constraints

The computations for determining the colour at each position are defined by equations (1-3). An interior point of the legend, of geometric coordinates (x, y), will have the colour [r, g, b]$^{\mathrm{T}}$ as computed.

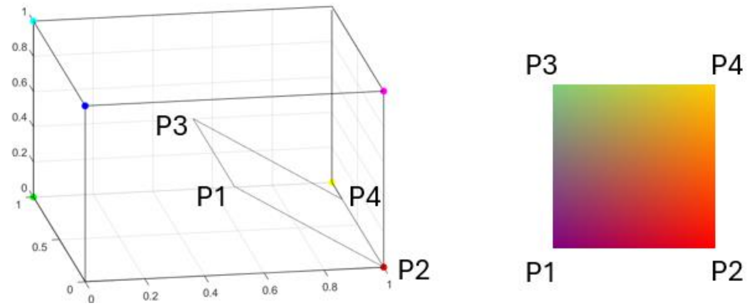
$$r = \begin{bmatrix} 1-x & x \end{bmatrix} \begin{bmatrix} r_{1,1} & r_{1,2} \\ r_{2,1} & r_{2,2} \end{bmatrix} \begin{bmatrix} 1-y \\ y \end{bmatrix} \tag{1}$$

$$g = \begin{bmatrix} 1-x & x \end{bmatrix} \begin{bmatrix} g_{1,1} & g_{1,2} \\ g_{2,1} & g_{2,2} \end{bmatrix} \begin{bmatrix} 1-y \\ y \end{bmatrix} \tag{2}$$

$$b = \begin{bmatrix} 1-x & x \end{bmatrix} \begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \end{bmatrix} \begin{bmatrix} 1-y \\ y \end{bmatrix} \tag{3}$$

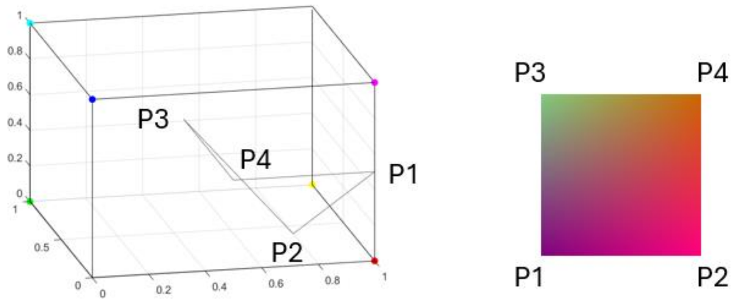
Fig. 5 and Fig. 6 show examples of legends generated from coplanar corners and non-coplanar corners, respectively. The RGB space is used for convenience, but the equations would work for other three-dimensional spaces as well.

The condition of having distinct colour for each point in the legend is achieved when opposing sides do not intersect in the 3D representation.



(a) (b)

Figure 5. Bivariate legend generated from coplanar corner constraints:
(a) view inside the RGB cube, (b) view as legend

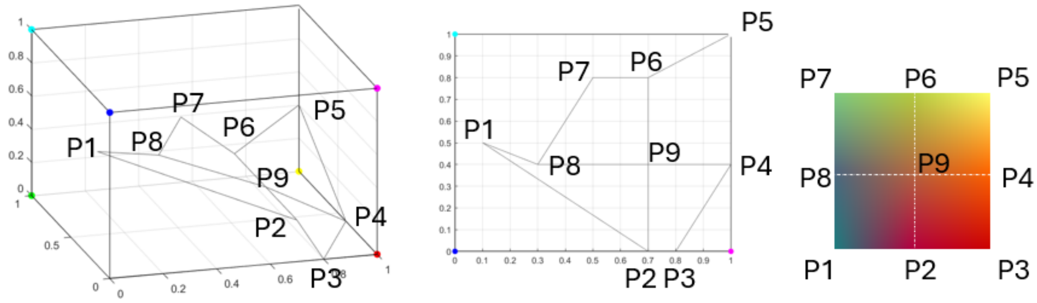


(a) (b)

Figure 6. Bivariate legend generated from non-coplanar corner constraints:
(a) view inside the RGB cube, (b) view as legend

2.3.2. Multiple constraints

Multiple constraints can generate piecewise-interpolated legends by using subsets of four constraints (patches), as shown in Fig. 7. If the patches of the legend are not supposed to be rectangular, the next step (warping) would provide a way of distorting the legend as needed.



(a) (b) (c)

Figure 7. Bivariate legend generated from sets of non-coplanar corner constraints (patches):
(a) view inside the RGB cube, (b) top view of the RGB cube, (c) view as legend

There are also other types of interpolation, such as polynomial or spline, which may also create interesting legends under the condition that the map does not intersect or fold onto itself, so that each colour in the legend is unique.

2.4. Warping

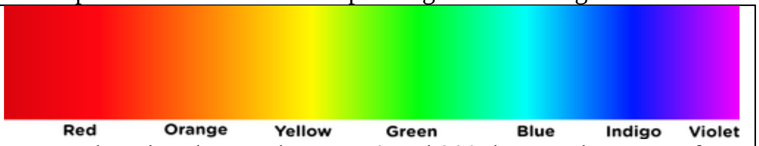
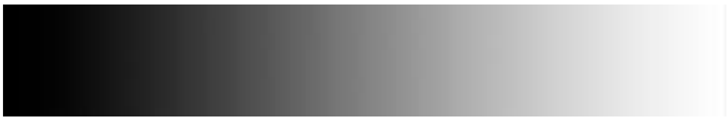
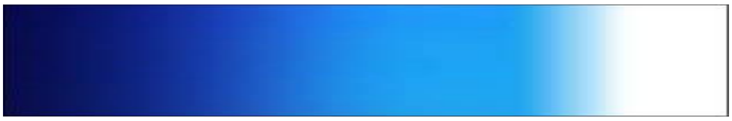
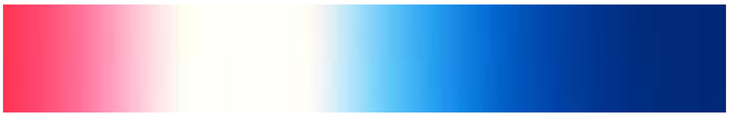
The final step considers the legends as images. Warping in this case refers to the any image processing technique that maps the image onto itself bijectively, preserving point connections but not necessarily Euclidean distances [25]. Any distortions, including stretching, triangle meshing and grid warping, are allowed as long as folding does not occur. If the legend is considered satisfactory after the initial interpolation step, or if the significance of the intermediate values mapped by the legend can be adjusted naturally outside the legend, then warping can even be skipped.

Interpolation and warping could theoretically be combined into a single step, but it is more intuitive to separate the selection of colours (that make up the legend and can be visualized inside the 3D colour space) from the processing of the legend (with the purpose of emphasizing regions of colour).

3. Univariate colour legends

Quantitative 1D colour palettes are relatively easy to create from the natural order of hues in the spectrum or by creating lighter/darker shades of a primary colour. *Table 2* shows several examples of 1D colour legends.

Table 2. Examples of sets of parameters and the corresponding univariate legends

Rainbow palette: Colour space: HSV Endpoints: $[0, 1, 0.5]^T, [300, 1, 0.5]^T$ Interpolation: linear Warping: none	 Red Orange Yellow Green Blue Indigo Violet By changing the *Hue* between 0 and 300 degrees the range of colours spans the entire rainbow
Grayscale palette: Colour space: RGB Endpoints: $[0, 0, 0]^T, [1, 1, 1]^T$ Interpolation: linear Warping: none	
Blue shades palette: Colour space: RGB Endpoints: $[0, 0, 1]^T, [1, 1, 1]^T$ Interpolation: linear Warping: none	
Red to white to blue palette: Colour space: RGB Endpoints patch 1: $[0.5, 0, 0]^T, [1, 1, 1]^T$ Endpoints patch 2: $[1, 1, 1]^T, [0, 0, 1]^T$ Interpolation: linear Warping: stretching (blue)	

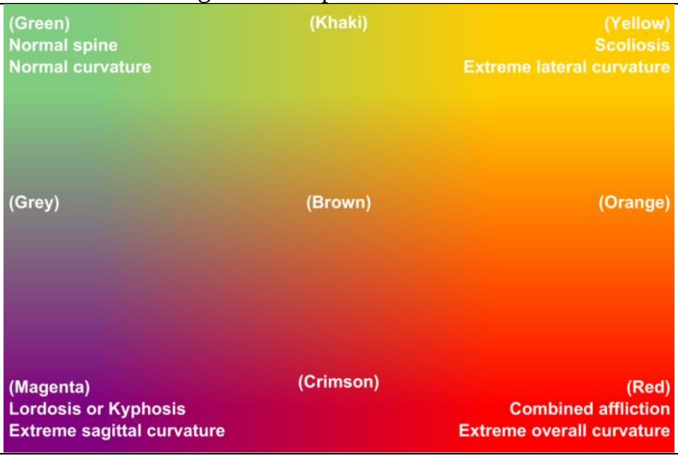
4. Case study: *Intuitive bivariate colour legend in spinal disorders*

As an example, we have situation in the medical field where two spine afflictions can indicate (relatively) independent problems [22]. The intensities of the two types of afflictions along the spine, *scoliosis* and *kyphosis*, are represented with colour gradients *green-yellow* (for scoliosis in the frontal view) and *green-magenta* (for kyphosis in the sagittal view), but their colours can be combined with the help of the bivariate legend shown in *Table 3*. On top of the colours, the legend also presents possible interpretations of the colours, depending on the images where they are used. As shown in Fig. 8.b, the colours can be used to indicate the severity of the affliction. Fig. 8.c uses the colours to indicate normal or extreme curvature on regions of the spine, either in the frontal plane Fig. 8.c (left) or the sagittal plane Fig. 8.c (right), then mixing them according to the legend for the combined stress on the spine in Fig. 8.c (middle).

The map considers the semaphore conventions, having a shade of *green* (no affliction) as the good corner and *yellow*, *magenta*, and *red* as the remaining corners, indicating extreme afflictions.

The four corners in this case are co-planar and not too dark.

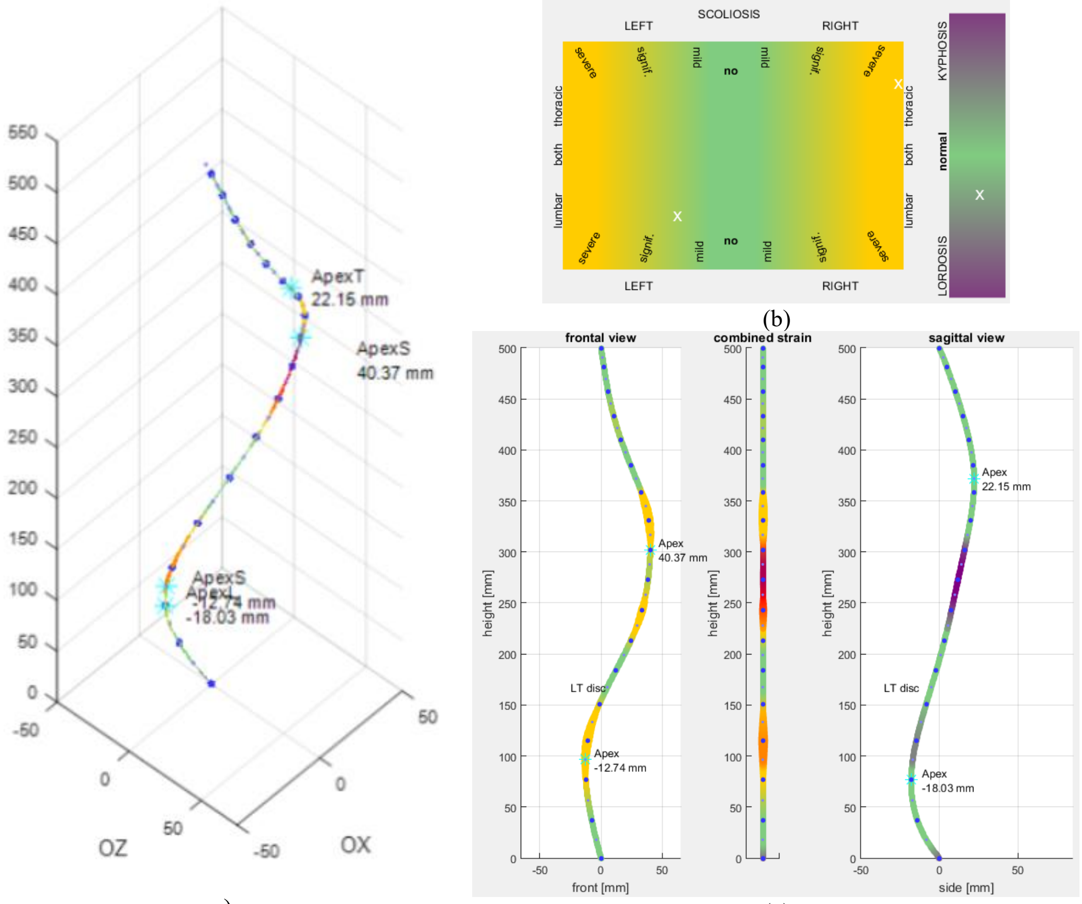
Table 3. Examples of a set of parameters and the corresponding bivariate legend with medical diagnosis interpretations

***Colour space:* RGB** *Top Corners:* Top left: $[0.5, 0.8, 0.5]^T$ Top right: $[1, 0.8, 0]^T$ *Bottom corners:* Bottom left: $[0.5, 0, 0.5]^T$ Bottom right: $[1, 0, 0]^T$ Interpolation: *linear* Warping: *none*	

Generating the map with linear interpolation creates colour gradients that are compatible with the meaning of the position:

- a lack of afflictions is marked with shades of *green*
- medium individual afflictions are neutral colours *grey* and *khaki*
- combined medium afflictions are *murky browns*
- severe individual afflictions are *yellow* and *magenta*
- extreme combined afflictions are increasingly more *orange* and *crimson* until the *bright red* indicating extreme two-variate afflictions

In terms of computational description, this legend is especially intuitive because the horizontal affliction changes only the red and blue layers, whereas the vertical affliction changes the green layer.



a) (b) (c)

Figure 8. (a) 3D representation of the spine (pseudo-coloured),
(b) Scoliosis and Kyphosis diagnostic with separate colour codes
(c) frontal and sagittal view with separate colour codes, combined strain with combined colour code
(using the legend from *Table 3*)

5. Conclusions

The presented method for generating bivariate legends is both reliable and flexible. As shown in previous section, smart choices of parameters enable the generation of other maps discussed in the introduction, as well as providing an option to further adapt the legends to more specific needs through warping.

The method can be further generalized by considering other colour spaces (although RGB and HSV are still the most common for visualization purposes), other interpolation techniques or new parametric / non-parametric warping distortions.

Although more difficult to visualize, tri-variate legends can be constructed as subspaces of the three-valued colour spaces (the most intuitive being the RGB cube itself), by defining the eight corners of the map. However, such legends can be difficult to represent in 2D and it might be simpler to explain the axes of the legend in words. Multivariate legends are mathematically extractable from multi-valued (pseudo-)colour spaces, but in order to have the information intelligible to users it may be useful in many cases to extract meaning though data processing techniques (such as *Principal Component Analysis*) and then revert to bivariate legends, if possible.

References

[1] „Encyclopedia of Color Science and Technology", ed. R. Shamey, ISBN 978-3-030-89861-8, DOI: 10.1007/978-3-030-89862-5, (2023)

[2] Adamson P., „Vision, Light and Color in al-Kindi, Ptolemy and the Ancient Commentators", Arabic Sciences and Philosophy, vol. 16, No. 2, pp. 207-236, DOI:10.1017/S0957423906000312, (2006)

[3] Kirchner E., „Color Theory and Color Order in Medieval Islam: A Review", Color Research & Application, vol 40, DOI: 10.1002/col.21861, (2013)

[4] Pujazon Patron E.C., Guerrero Zegarra M.A., Elias J.D., „A Historical Approach to Understanding Differentiation of RYB vs RGB", Journal of Visual Art and Design, vol. 16(1), DOI: 10.5614/j.vad.2024.16.1.5, (2024)

[5] Sinico M., „Tertiary qualities, from Galileo to Gestalt psychology", History of the Human Sciences, vol. 28(3), pp. 68-79, DOI: 10.1177/0952695115591409, (2015)

[6] Martins R.D.A., Celestino C., „Newton and Colour: Complex Interplay of Theory and Experiment", Science & Education, vol. 10, pp. 287–305, DOI:10.1023/A:1017219114697, (2001)

[7] Duck M. J., „Newton and Goethe on colour: Physical and physiological considerations", Annals of Science, vol. 45(5), pp. 507–519, DOI: 10.1080/00033798800200361, (1988)

[8] Silva S, Santos B, Madeira J., „Using color in visualization: A survey", Computers & Graphics, vol. 35, pp. 320 - 333, DOI: 10.1016/j.cag.2010.11.015, (2011)

[9] Brewer C., „Modern Cartography Series" (ed. A.Maceachren, D.R. Fraser Taylor), Academic Press, vol. 2, Chapter 7 – „Color Use Guidelines for Mapping and Visualization", pp. 123-147, ISBN 9780080424156, DOI: 10.1016/B978-0-08-042415-6.50014-4, (1994)

[10] Harrower M, Brewer C., „ColorBrewer.org: An Online Tool for Selecting Colour Schemes for Maps", The Cartographic Journal, vol. 40 (1), pp. 27–37, DOI: 10.1179/000870403235002042, (2003)

[11] https://www.esri.com/arcgis-blog/products/arcgis-desktop/mapping/making-bivariate-choropleth-maps-with-arcmap

[12] https://cran.r-project.org/web/packages/biscale/vignettes/biscale.html

[13] https://www.tableau.com/blog/how-make-effective-bivariate-choropleth-maps-tableau-83121

[14] https://www.kitware.com/bivariate-representations-in-paraview/

[15] Hruby F., „Applications of Bivariate Choropleth Maps", 15th International Multidisciplinary Scientific GeoConference SGEM2015, DOI: 10.5593/SGEM2015/B22/S11.088, (2011)

[16] Eyton R., „Complementary-Color, Two-Variable Maps", Annals of the Association of American Geographers, vol. 74(3), pp. 477-490, DOI: 10.1111/j.1467-8306.1984.tb01469.x, (1984)

[17] Reimer A., „Squaring the circle: bivariate colour maps and Jacques Bertins' concept of disassociation", International Cartographic Conference, pp. 3–8, (2011)

[18] Steiger M., Bernard J., Thum S., Mittelstädt S., Hutter M., Keim D.A., Kohlhammer J., "Explorative Analysis of 2 D Color Maps", 23rd International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision / Gavrilova, (2015)

[19] Darbyshire J., Jenny B., "Natural-color Maps via Coloring of Bivariate Grid Data", Computers & Geosciences, DOI: 10.1016/j.cageo.2017.06.004, (2017)

[20] Teuling A., Stöckli R., Seneviratne S., "Bivariate colour maps for visualizing climate data", International Journal of Climatology, DOI:10.1002/joc.2153, (2011)

[21] Müller, A., Lausser, L., Wilhelm, A. et al., "A perceptually optimised bivariate visualisation scheme for high-dimensional fold-change data". Adv. Data Anal Classif,, DOI: 10.1007/s11634-020-00416-5, (2021)

[22] Neghina, M.,Petruse R., Ćuković, S., Schiau C., Filipovic N., "Automatic Curvature Analysis for Finely Interpolated Spinal Curves", 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE), DOI: 10.1109/BIBE52308.2021.96354244, (2021)

[23] Grout, R. W., Gruber, A., Kolla, H., Bremer, P.-T., Bennett, J. C., Gyulassy, A., & Chen, J. H., "A direct numerical simulation study of turbulence and flame structure in transverse jets analysed in jet-trajectory based coordinates", Journal of Fluid Mechanics, vol.706, pp.351–383, DOI: 10.1017/jfm.2012.257, (2012)

[24] Tkalcic, M.,Tasic, J., "Colour spaces: Perceptual, historical and applicational background", EUROCON 2003, Computer as a Tool, DOI: 10.1109/EURCON.2003.1248032, (2003)

[25] Glasbey, CA, Mardia, Kanti, "A review of image-warping methods", Journal of Applied Statistics, vol. 25, pp. 155-171, DOI: 10.1080/02664769823151, (1998)

[26] Khediri N., Ammar M., Kherallah M., "Comparison of Image Segmentation using Different Color Spaces", 2021 IEEE 21st International Conference on Communication Technology (ICCT), DOI: 10.1109/ICCT52962.2021.9658094, (2021)

[27] Bathymetry map of the Mediterranean Sea. Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Mediterranean_Sea_Bathymetry_map.svg

[28] Black Hispanic Bivariate Map. Wikimedia Commons, https://commons.wikimedia.org/w/index.php?title=File.png&oldid=825098937

[29] Annual Average Temperature Map. Wikipedia, https://bo.m.wikipedia.org/wiki/File:Annual_Average_Temperature_Map.jpg

DOI: 10.2478/ijasitels-2024-0014

A Categorical Formalism for Conceptual Graphs

Daniel C. Crăciunean[1]

[1]*Computer Science and Electrical and Electronics Engineering Department, Faculty of Engineering, "Lucian Blaga" University of Sibiu, Romania*
daniel.craciunean@ulbsibiu.ro

Abstract

In this paper, we propose a formalism for knowledge manipulation, based on conceptual graphs, which is supported by category theory, and which can be the basis for the development of knowledge-based systems. The model we propose is an extension of the model based on conceptual graphs. In this model, the application of an inference rule between logical formulas represented by two conceptual graphs is reduced to the identification of an arrow in a category. To this end, we introduce several new notions such as: the category of conceptual graphs, the category of classes of conceptual graphs, the conceptual category of a model and the conceptual category of inference of a model.

Keywords: conceptual graphs, coreference relation, category theory, subsumption functor, category of conceptual graphs, conceptual category

1 Introduction

In order to model logical inference through algorithmizable and accessible mechanisms, in the 1970s, a series of graphical languages [6] were developed to specify first-order logic. In this context, in 1976 Sowa [11] introduced a version of a graphical language for specifying knowledge, questions, and statements in natural language in terms representable in a relational database, called conceptual graphs.

The graphical atoms of the language are the rectangles that represent concepts, the circles that represent conceptual relations, and the edges that represent the connecting elements between concepts and relations.

A conceptual graph is a bipartite graph such that the concepts neighbouring a conceptual relation node are always nodes that represent concepts. The conceptual relations in a conceptual graph represent atomic formulas in first-order logic, and the neighbouring concepts represent the arguments of these atomic formulas. To specify the order of these arguments, the edges connecting a conceptual relation and its neighbouring concepts are marked with natural numbers; 1, . . . , n , where n is the number of neighbouring concepts.

In this paper, we propose a formalism for knowledge manipulation, based on category theory, which can be the basis for the development of knowledge-based systems [9, 13]. The model we propose is an extension of the model based on conceptual graphs [11].

In this model, the application of an inference rule between logical formulas represented by two conceptual graphs reduces to identifying an arrow in a category.

The main novelty of this paper is the formalization of conceptual graphs and reasoning based on conceptual graphs using category theory as support. For this purpose, we introduce notions such as: category of conceptual graphs, category of classes of conceptual graphs, conceptual category of the model and conceptual category of model inference.

Section 2 of the paper contains Background Notions, section 3 introduces the category of conceptual graphs and the category of classes of conceptual graphs, section 4 introduces the conceptual category of the model and the conceptual category of model inference. The paper ends with conclusions and observations.

2 Background Notions

Conceptual graphs are specified by a graphical language and represent different types of knowledge such as: facts, objectives, rules and queries [1]. Facts are statements about the existence of entities, about the properties of an entity or about the relationships between them. Objectives represent the goal pursued by an evolving system. Rules can describe knowledge and constraints, implicit in the model, as well as the evolution of processes.

A conceptual graph is a bipartite multigraph, whose nodes represent concepts and conceptual relationships. The edges of the multigraph represent the connection between the two types of nodes.

Each conceptual node is represented by a pair of labels, one specifying the type of concept and one specifying an individual of the type specified by the first label. If the individual label is missing, they will be represented by a variable. Conceptual relationship nodes are marked with labels that represent types of relationships. The edges, in turn, are labelled with natural numbers in the order in which the neighbouring concepts will become the parameters of these conceptual relationships.

These labels with which a conceptual graph is endowed form a vocabulary, which is denoted by $T=(T_C, T_R, E)$ where:

T_C is a set of concept types,

T_R is a set of relationship types,

E is the set of individual labels.

The connection between a conceptual graph and the associated vocabulary is made by an application that distributes the vocabulary to the components of the graph.

Therefore, a conceptual graph G, is a tuple (T, G, $\boldsymbol{\eta}$), where T is a vocabulary, $G=(C,R,\boldsymbol{\Gamma})$, is a bipartite multigraph, and $\eta=(\eta_C,\eta_R,\eta_\Gamma)$ is the application that distributes the vocabulary to the components of the graph, thus: $\eta_C:C\rightarrow T_C\times (E\cup\{*\})$, $\eta_R:R\rightarrow T_R$, and $\eta_\Gamma:\Gamma\rightarrow N$.

The basic support for inference based on conceptual graphs consists of two relations defined on both the set of concepts and the set of relations [1]. The first relation is the coreference relation which is a total equivalence relation and which we denote by $\boldsymbol{\rho}$. The second relation is the generalization relation which is a partial order relation and which we denote by ≥. Based on the generalization relation we introduce the subsummation functors that define logical deduction.

In this paper we will use category theory to formalize knowledge representation and reasoning based on conceptual graphs [12]. A category is a mathematical structure composed of a set of objects and a set of arrows between these objects, to which is added an operation of composing the arrows. The composition operation is associative and with neutral elements.

3 The Category of Conceptual Graphs

A subsummation homomorphism φ is a mapping from a conceptual graph G to a conceptual graph H, which:

- maps each concept in G to a concept in H and each relation in G to a relation in H,
- preserves the conceptual graph structure, that is, maps each edge (A, B), which connects a pair of nodes A and B in G, to an edge (φ(A), φ(B)), which connects nodes φ(A) and φ(B) in H,
- φ is a monotonically decreasing mapping, that is, $\forall e \in C_G \cup R_G$, $\eta_G(e) \geq \eta_H(\varphi(e))$.

We can define the composition of two subsummation homomorphisms as follows:

Let G, H and K be three conceptual graphs, and φ:G →H and ψ:H→K, then the subsummation homomorphism resulting from the composition of φ with ψ is $\delta=\psi(\varphi)$, δ:G→H. It is obvious that δ is a subsumption homomorphism.

If G, H, K, S are conceptual graphs and φ :G →H, ψ: H → K, μ:K→S, then obviously $(\varphi\circ\psi)\circ\mu=\varphi\circ(\psi\circ\mu)$:G→S, and therefore the operation of composing subsumption homomorphisms is associative.

For every conceptual graph G , there exists the identity subsumption homomorphism: ι:G→G, which takes each component of G into the same component.

Therefore, the set of conceptual graphs together with the set of subsumption homomorphisms, among them, form a category, in which the objects are conceptual graphs and the arrows are the subsumption morphisms. We call this construction, the category of conceptual graphs.

The ultimate goal of conceptual graphs is to allow the specification in natural language of different types of knowledge such as: facts, objectives, rules and queries. But natural language is flexible, and allows the use of multiple concepts for the same real entity, at least in terms of the vocabulary used.

To solve this problem, we will introduce an equivalence relation on the set of concepts that we call the coreference relation. In [8], the coreference relation is defined on the set of conceptual nodes. We introduce this relation on the set of concepts that could contain both multiple concepts representing the same real entity and multiple names for the same concept.

The same thing happens with the set of conceptual relations, ultimately the components, called conceptual relations, of the conceptual graphs are also concepts. Natural language also allows a multimer of syntactic formulas for expressing them. It follows that the coreference relation will have to refer to these as well.

Therefore, the coreference relation that we use here has two components, one that refers to the coreference of concepts and one that refers to the coreference of conceptual relations.

We denote the coreference relation with $\rho=(\rho_C,\rho_R)$, where: ρ_C, groups all syntactic forms that represent an entity in a set of concepts that we call the class of coreferent concepts, and ρ_R, groups all syntactic forms that represent a conceptual relation in a set of conceptual relations that we call the class of coreferent conceptual relations.

If C is a set of concepts, and R is a set of conceptual relations, we define the coreference relation $\rho=(\rho_C,\rho_R)$ as follows: $\rho_C \subseteq C\times C$, $(c_1,c_2)\in\rho_C$, if and only if c_1 and c_2 refer to the same real entity; $\rho_R \subseteq R\times R$, $(r_1,r_2)\in\rho_R$, if and only if r_1 and r_2 refer to the same real conceptual relation.

Next, we will define the notion of a class of conceptual graphs, in which each conceptual node will represent a class of coreferent concepts and each node, conceptual relation, will represent a class of coreferent conceptual relations.

If we have a conceptual graph G, then the class of conceptual graphs associated with G, which we denote by G*, is the image of the functor ϕ:G→Rel, where Rel is the category of sets and relations, with the properties:

- For each conceptual node $A\in G_C$, $\phi(A)$ is the set of all coreferent concepts with A.
- For each node, thenconceptual relation $R\in G_R$, $\phi(R)$ is the set of all coreferent conceptual relations with R.
- For each edge, (R,A) $R\in G_R$, $A\in G_C$, $\phi((R,A))$ is the set of all total relations, $\boldsymbol{\rho}\subseteq R\times A$.

We observe that the conceptual graph G, plays the role of a categorical sketch graph for a class of conceptual graphs, and therefore can serve us to impose various conditions on the structure of the graph [3, 4], but we will not address this issue in the present paper.

If G*, is a class of conceptual graphs then we will define the operation: G=Slice(G*)=G*/**ρ**, thus: $G_C=G_C^*/\rho_C$, $G_R=G_R^*/\rho_R$, and the edges between relations and concepts will be those that exist between the concepts and conceptual relations selected as representatives in the Slice(G*) operation. We observe that the graph G=Slice(G*) is a conceptual graph.

To define subsummation homomorphisms between two classes of conceptual graphs, we will overload the ≥ relation, for coreferent concept classes and coreferent conceptual relation classes, as follows:

- If A and B are two coreferent conceptual classes then $A\geq B$, if and only if $\forall\ a\in A$ and $\forall\ b\in B, \Rightarrow a\geq b$,
- If R and P are two coreferent conceptual relation classes then $R\geq P$, if and only if $\forall\ r\in R$ si $\forall\ p\in P, \Rightarrow r\geq p$.

We can now define a subsummation homomorphism φ^*, between two classes of conceptual graphs, as a map from a class of conceptual graphs G*, to a class of conceptual graphs H*, which:

- maps each conceptual class in G*, to a conceptual class in H*, and each relation in G* to a relation in H*,

- preserves the class structure of conceptual graphs, that is, maps each edge (A*,B*), connecting a pair of nodes A* and B* in G, to an edge ($\varphi(A^*)$, $\varphi(B^*)$), connecting nodes $\varphi(A^*)$ and $\varphi(B^*)$ in H*,
- φ is a monotonically increasing map, that is, $\forall e \in C^*_{G^*} \cup R^*_{G^*}$, $\eta_{G^*}(e) \geq \eta_{H^*}(\varphi(e))$.

We can define the composition of two subsummation homomorphisms as follows:

Let G, H, and K be three classes of conceptual graphs, and φ:G→H and ψ:H→K, then the subsummation homomorphism resulting from the composition of φ with ψ is $\delta=\psi(\varphi)$, δ:G→H. It is obvious that δ is a subsummation homomorphism.

If G, H, K, S are classes of conceptual graphs and φ :G →H, ψ: H → K, μ:K→S, then obviously $(\varphi\circ\psi)\circ\mu=\varphi\circ(\psi\circ\mu)$:G→S, and therefore the operation of composing subsummation homomorphisms is associative.

For every conceptual graph G, there exists the identity subsummation homomorphism: ι:G→G, which takes each component of G to the same component, through the identity function.

Therefore, the set of classes of conceptual graphs together with the set of subsummation homomorphisms between them form a category, in which objects are classes of conceptual graphs and arrows are subsummation morphisms between them. We call this construction the category of classes of conceptual graphs and denote it by CCGC.

If $\boldsymbol{\mathcal{C}}^*$, is the category of classes of conceptual graphs then we will define the operation: $\boldsymbol{\mathcal{C}}$=Slice($\boldsymbol{\mathcal{C}}^*$)= $\boldsymbol{\mathcal{C}}^*/\boldsymbol{\rho}$, thus: ob($\boldsymbol{\mathcal{C}}$)= ob($\boldsymbol{\mathcal{C}}$)/$\rho_C$, arrow($\boldsymbol{\mathcal{C}}$)=arrow($\boldsymbol{\mathcal{C}}^*$)/$\rho_R$. It can be easily demonstrated that Slice($\boldsymbol{\mathcal{C}}^*$) is a category. We note that this category is not isomorphic to the category of conceptual graphs because the latter also contains conceptual graphs with coreferent nodes. We will call the category Slice($\boldsymbol{\mathcal{C}}^*$), the inference category of conceptual graphs and denote it by CIGC.

4 Categorical Knowledge Modelling

The process of modelling knowledge in a domain begins with the conceptualization of the domain in question [2]. In knowledge modelling, concepts can represent real entities or can be generic in which case they represent classes of entities. Conceptualizing a domain of knowledge involves identifying the real entities involved in this domain and the relationships between these entities and replacing them with concepts that represent them.

Therefore, the conceptual metamodel of a domain of knowledge is a set of concepts that allow the specification of knowledge of interest in the domain in question [5].

The entities and the relationships between them, involved in the domain, must be replaced with concepts. Concepts must be introduced through clear and precise definitions. Defining a concept involves the precise delimitation of a family of real objects, which it is to represent, from a larger, already known family, by adding new properties, common to all objects of the new concept and only it. The generic name given to the new family of objects is the name of the defined concept. Therefore, each newly defined concept retains all the properties of the concepts on the basis of which it was defined, to which the new properties are added.

As we mentioned, to define a new concept, we must rely on another already defined one. But initially, when we do not have any defined concept, we must start from primary concepts, which we do not define, or from concepts defined in other models. This way of defining concepts provides us with a natural hierarchy of these concepts. Therefore, we can introduce the relation ≥ , on the set of concepts, which we call the generalization relation. Obviously, this relation is a partial order relation.

If C, is a set of contexts and ≥, is a generalization relation then it respects the properties of:

- reflexivity : $\forall$ c$\in$C, c ≥ c;
- transitivity : $\forall$ c,d,e$\in$C, c ≥ d and d ≥ e $\Rightarrow$c ≥ e;
- antisymmetry : if e ≥ f and f ≥ e $\Rightarrow$ e = f.

It follows that the relation ≥ is a partial order relation. This relation is the basic support for logical inference in conceptual graphs. It is obvious that if c, d are concepts and c≥d, then any logical formula satisfied, on the properties of c, is also satisfied for the properties of d. But conceptual relations are also concepts and therefore the generalization relation ≥ can be overloaded for conceptual relations.

The concepts involved in knowledge modelling, most of the time, represent classes of entities and therefore are generic. Generic concepts can have potential properties and can undergo potential transformations depending on certain contexts, which can also be potential.

4.1 Conceptual Category of Model Inference

Suppose that, for a specific domain, we have identified all the concepts involved in the representation of knowledge and all the conceptual relations between them, including the lexical varieties that represent them. We denote the set of these concepts by C and the set of conceptual relations by R. On the set C, we introduce the coreference relation $\boldsymbol{\rho}$. The coreference relation $\boldsymbol{\rho}$ is an equivalence relation and therefore partitions the set C into a set Č of classes of concepts, Č=C/$\boldsymbol{\rho}$, and the set R into a set of classes of conceptual relations, Ř=R/$\boldsymbol{\rho}$.

We saw in Section 3 that the set of classes of conceptual graphs together with the set of subsummation homomorphisms between them form a category, which we called the category of classes of conceptual graphs and denoted it by CCGC.

We consider the set of all classes of conceptual graphs, which have as conceptual nodes, elements from Č, and as relational nodes, elements from Ř, and we denote this set of classes of conceptual graphs by CCG(Č,Ř).

We will now construct a subcategory of the category CCGC. For this we will observe that if we have a category $\boldsymbol{\mathcal{C}}$, and a set of objects O, we can define a subcategory $\boldsymbol{\mathcal{D}}$, of $\boldsymbol{\mathcal{C}}$, which has ob($\boldsymbol{\mathcal{D}}$)=O, $\boldsymbol{\mathcal{D}}$(X,Y)=$\boldsymbol{\mathcal{C}}$(X,Y), where X,Y$\in$O , with the identity and composition in $\boldsymbol{\mathcal{C}}$.

Based on this observation, we will now construct a subcategory of the category CCGC, which we call the conceptual category of the model (CCM), in the following way:

The set of objects of the category CCM is: ob(CCM)=ob(CCGC)$\cap$CCG(Č,Ř)

The set of arrows of the category CCM is: arrow(CCM)={CCGC(X,Y)|X,Y$\in$ CCG(Č,Ř).

It is easy to prove that the CCM construction is a category. The objects of this category are classes of conceptual graphs, which model the specific domain, and the arrows of the category are subsummation homomorphisms that model the logical inference between these classes of conceptual graphs.

Starting from the CCM category we can construct the conceptual category of inference of the model, which we denote with CCIM= Slice(CCM)=CCM/**ρ**. The CCIM category has the components: ob(CCIM)=ob(CCM)/**ρ**, arrow(CCIM)=arrow(CCM)/**ρ**. Obviously the CCIM category is a subcategory of the category of inference of conceptual graphs (CIGC).

4.2 Logical Inference in the Conceptual Category of Model Inference

The terms, in first-order logic, for a domain are constants, variables and functions that model certain phenomena specific to the domain [10]. An atomic formula, in the domain, is a predicate that has as parameters terms specific to the domain. Formulas are built on the basis of atomic formulas by introducing logical operators.

The diagrammatic language, used to specify conceptual graphs, can be naturally associated with the language of first-order logic, as follows:

- Individual labels of conceptual nodes become constants.
- Unspecified individual labels of conceptual nodes become variables.
- Conceptual node type labels become unary predicates, which have as parameters the individual label associated with the conceptual node.
- Conceptual relation node labels become n-ary predicates that have as parameters the type labels of neighbouring conceptual nodes.

In the following, we will use the same names for logical constants and predicates as those in the vocabulary of the conceptual graph, that is, the logical constant c will represent the individual label c and the logical predicate p will represent the type of concept or conceptual relation p. The individual labels that are not specified, we will denote with $x_1,x_2,\dots,x_n$, and will represent the variables of the logical formula associated with a conceptual graph.

In this way, to each conceptual graph G, we can attach:

- A logical formula corresponding to the conceptual nodes:

$\Phi(G_E)=\exists\ x_1,\dots x_n \bigwedge_{p\in E} p(e)$, where e is the individual label of the concept type p or the corresponding variable x_i, if the individual type is not specified.

- A logical formula corresponding to the nodes of conceptual relations:

$\Phi(G_R)=\exists\ x_1,\dots x_n \bigwedge_{r\in R} r(t_{1\dots,}t_k)$ where $t_1,..,t_k\in\{x_1,\dots,x_n\}$

- The logical formula [8], which represents the semantics of the conceptual graph G is:

$\Phi(G)= \Phi(G_E) \wedge\Phi(G_R)$.

Thus, the semantics of conceptual graphs is given by logical formulas of order 1 [8]. We note that these formulas use only the universal logical quantifier ∃, and the logical connector ∧, and, therefore, represent a subset of logical formulas of order 1. We denote by FOL(∧, ∃). We also denote by FOL(∧, ∃, CCIM), the set of logical formulas that represent the semantics of all conceptual graphs that are objects in the conceptual category of model inference (CCIM).

It is shown that for any logical formula in FOL(∧, ∃), a conceptual graph can be constructed that has this formula as its semantics [8]. Therefore, for any formula in FOL(∧, ∃), the corresponding conceptual graph can be constructed and conversely for any conceptual graph G, the formula in FOL(∧, ∃), which represents the semantics of G, can be constructed.

In our categorical model, subsummation homomorphisms are the fundamental ingredient for inference based on conceptual graphs. Inference in the CCIM category is based on the following proposition:

If G and H are objects in the CCIM category, then $\Phi(G) \vdash \Phi(H)$, if and only if CCIM(G,H)∈arrow(CCIM), that is, if and only if there is a subsummation homomorphism from the conceptual graph G to the conceptual graph H.

If we want to verify an implication of the type $\phi \vdash \psi$, where ϕ, $\psi \in$ FOL(∧, ∃, CCIM), we will construct the conceptual graphs CG(ϕ) and CG(ψ), and we will verify whether CCIM(CG(ϕ),CG(ψ))∈arrow(CCIM).

5 Observations and conclusions

We note that the CCIM category is a finite category, and therefore it can be constructed algorithmically. The objects of the CCIM category are classes of concepts ob(CCIM)=Č∪Ř. It follows that the model can be permanently enriched with new concepts or new syntactic forms, without modifying this category if they can be included in the existing classes.

Adding concepts that are not coreferential with any existing class requires transforming the category, cases that are not rare enough if the model is well designed. In this case, graph transformation mechanisms must be used [7].

In our categorical model, subsummation homomorphisms are the fundamental ingredient for inference based on conceptual graphs. Finding a homomorphism between two graphs is, in general, an NP problem if the graph of the domain of definition is not acyclic. Therefore, this problem, which was not addressed in this paper, must be treated seriously. These are just a few problems related to this model, which we will treat in future works.

References

[1] Jamie Caine, Simon Polovina, ”From Enterprise Concepts to Formal Concepts: A University Case Study, in Graph Structures for Knowledge Representation and Reasoning”, 5th International Workshop, GKR 2017 Melbourne, VIC, Australia, August 21, 2017, Revised Selected Papers, Lecture Notes in Artificial Intelligence.

[2] Daniel-Cristian Craciunean, Daniel Volovici, “Conceptualization of Modeling Method in the Context of Categorical Mechanism”, in Dimitris Karagiannis and others, Domain Specific Conceptual Modeling, Springer Nature Switzerland AG (2022).

[3] Daniel-Cristian Craciunean, “Categorical Mechanisms in Multi-level Modeling Methods”, Publishing House of the "Lucian Blaga" University of Sibiu, (2023).

[4] Daniel-Cristian Crăciunean, D. Karagiannis, “A categorical model of process cosimulation”, Journal of Advanced Computer Science and Applications (IJACSA), 10, (2019)

[5] D. Karagiannis, H.C. Mayr, J. Mylopoulos, “Domain-Specific Conceptual Modeling Concepts, Methods and Tools”, Springer International Publishing Switzerland (2016)

[6] Dominik Bork , Dimitris Karagiannis, Benedikt Pittl, “A survey of modeling language specification techniques”, Information Systems 87 (2020) 101425, journal homepage: www.elsevier.com/locate/is

[7] Hartmut Ehrig, Claudia Ermel, Ulrike Golas, Frank Hermann, “Graph and Model Transformation General Framework and Applications”, Springer-Verlag Berlin Heidelberg (2015)

[8] Michel Chein , Marie-Laure Mugnier, “Graph-based Knowledge Representation, Computational Foundations of Conceptual Graphs”, Springer-Verlag London (2009).

[9] Michael Barr, Charles Wells, “Category Theory For Computing Science”, Reprints in Theory and Applications of Categories, No. 22, (2012).

[10] M. Huth, M. Ryan, “Logic in Computer Science, Modelling and Reasoning about Systems”, Published in the United States of America by Cambridge University Press, New York, (2004).

[11] Sowa, John F., “Conceptual graphs for a database interface”, IBM Journal of Research and Development **20:4**, 336-357, (1976).

[12] Zinovy Diskin, Uwe Wolter, “A Diagrammatic Logic for Object-Oriented Visual Modeling”, Electronic Notes in Theoretical Computer Science, Volume 203, Issue 6, 21 November (2008).

[13] Uwe Wolter, Zinovy Diskin, “The Next Hundred Diagrammatic Specification Techniques, A Gentle Introduction to Generalized Sketches”, 02 September (2015), https://www.researchgate.net/publication /253963677.

DOI: 10.2478/ijasitels-2024-0017

Explore the intersection of Self-Determination Theory and cybersecurity education - A literature review

Iulia Feraru[1], *Laura Bacali*[2]

1 IOSUD,, Technical University of Cluj-Napoca, Memorandumului 28, 400114, ClujNapoca, Romania

2 Technical University of Cluj-Napoca, Memorandumului 28, 400114, ClujNapoca, Romania

Abstract

This paper explores how organizations can create a sustainable, security-first culture in an increasingly complex environment where organizational and national cultures are strong influencing factors in human behaviour. In cybersecurity education, as in any effort of education, there must be a principled commitment to long-term behaviour modification through intrinsic motivation, foundational to employees acting consistently in secure ways. The review of the literature falls squarely within Self-Determination Theory, underlining the pertinence of autonomy, competence, and relatedness as distinctive factors in cybersecurity education, highlighting that these constructs are necessary at all levels for perpetual security and compliance.

The results indicated that an effective security-first culture could only emerge when cybersecurity formed part of the core values and practices within organizations. It also explained that leadership styles, such as transformational and servant leadership, play an important role in the development of intrinsic motivation by fostering trust, empowerment, and a sense of shared responsibility. It also highlights how national cultural dimensions, such as individualism and power distance, may change how differently oriented employees respond to cybersecurity policies and practices. Approaches to cybersecurity education should be tailored to both organizational and national cultural factors to develop cybersecurity education strategies that could go beyond mere compliance and build a proactive security mindset.

This is important because it underlines how the SDT acts as a framework for understanding how companies could help foster a security-first culture that, at the same time, will create sustainable, resilient, and intrinsically driven cybersecurity behaviours among employees.

Keywords: Behavioural change, information security, organizational culture, national culture, Self-Determination Theory

1 Introduction

Information security has become a critical concern for organizations across the globe, as the protection of sensitive information is fundamental to maintaining trust, ensuring regulatory compliance, and safeguarding against reputational and financial losses. Despite the continuous advancement of technical security measures, such as next-generation antivirus solutions and sophisticated threat detection systems, the human element remains the most significant vulnerability in information security. In the era of

Industry 4.0 ([8] Bhaharin et al., 2019), human error remains a significant threat to information security, often resulting from negligence, ignorance, and failure to adhere to organizational information security policies. To increase compliance with information security policies (ISPs) and reduce security incidents related to human behaviour, it is essential to systematically analyse and address the underlying issues influencing employees' attitudes towards policy adherence ([8] Bhaharin et al., 2019). Motivation is a critical component in shaping secure behaviour and ensuring compliance with security policies. According to Self-Determination Theory (SDT) ([75] Deci & Ryan, 2000), intrinsic motivation—driven by a sense of autonomy, competence, and relatedness—can significantly enhance employees' commitment to security practices ([75] Deci & Ryan, 2000). However, motivation is not solely an individual attribute but is also influenced by the broader cultural context within which individuals operate. Both organizational culture and national culture play very important roles in shaping employees' attitudes, behaviours, and motivations toward information security.

The interaction between organizational culture and national culture is a complex area of study, as cultural factors can either support or hinder the adoption of secure behaviours. Organizational culture encompasses shared values, norms, and practices that influence how employees perceive and respond to security policies ([10] Schein et al, 2017). Leadership, communication, and trust within an organization are key aspects that can either foster a security-conscious culture or contribute to complacency and non-compliance ([10] Schein et al, 2017). On the other hand, national culture, as defined by Hofstede’s dimensions, affects individuals' perceptions of authority, risk, and responsibility, which in turn influences their willingness to engage in security practices ([9] Hofstede et al, 2005).

The research question (RQ) formulated to explore these dynamics was:

- RQ: How can organizations foster a security-first culture that enhances employees' intrinsic motivation and sense of shared responsibility?

This article is aiming to investigate how intrinsic motivation, driven by the principles of Self-Determination Theory (SDT) ([75] Deci & Ryan, 2000), can support lasting behaviour change in cybersecurity. Motivation itself is a complex construct, encompassing both extrinsic and intrinsic forms. The BJ Fogg Behaviour Model ([5] Fogg, 2009) highlights that behaviour arises from the convergence of motivation, ability, and triggers, suggesting that motivation is essential for sustainable security compliance. While extrinsic motivators, such as penalties or rewards, can prompt compliance, they often lack the staying power required for deep-rooted behavioural change.

This article aims to shed light on how SDT constructs can be worked with to drive intrinsic motivation and achieve sustainable behaviour change, thus positioning cybersecurity as a core component of organizational culture rather than an obligatory task. By creating and maintaining an environment where individuals feel autonomous, competent, and connected, organizations can build a foundation for continuous cybersecurity education and proactive security engagement.

2 Background

2.1 Information security

Information security is a concept that becomes ever more enmeshed in many aspects of our society, largely as a result of our nearly ubiquitous adoption of computing technology. In our everyday lives, many of us work with computers for our employers, play on computers at home, go to school online, buy goods from merchants on the Internet, take our laptops to the coffee shop and check our e-mail, carry our smartphones on our hips and use them to check our bank balances, track our exercise with sensors in our shoes, and so on, ad infinitum. ([1] Andress, J.,2014)

There are various definitions of Information Security and they all relate to the preservation of confidentiality, integrity and availability of information over the Internet and other properties, such as authenticity, accountability, non-repudiation and reliability that can also be involved. ([2] ISO/IEC 27000:2018)

Information is a critical business asset nowadays and managing its security ensures business continuity, competitiveness, profitability, prestige, elimination of threats and losses resulting from realized risks. ([3] Bolek et. al, 2023)

2.2 Human-centric security

Although technical security controls, such as next-generation antivirus software and improved spam filters, continue to advance, the human factor remains the leading cause of security incidents, contributing to 68% of data breaches. ([4] Verizon, 2024)

Human error, negligence, and risky behaviour contribute significantly to security incidents, often cancelling even the most sophisticated technical defences. Recognizing this, there is an increasing need to focus on understanding the role of individuals in maintaining information security. Employees, when properly educated and motivated, can become the organization’s strongest line of defence. By fostering a culture of awareness and proactive response to potential threats, organizations can significantly reduce the likelihood of data breaches and security violations. This highlights the importance of examining behaviour in the context of information security, as it is through shaping secure behaviours that organizations can transform their employees from potential risks into active participants in safeguarding sensitive information.

2.2.1 Human behaviour in information security

Understanding human behaviour is crucial in the field of information security, as employees' actions can either enhance or undermine an organization’s security posture. Factors such as awareness, motivation, and the ability to recognize and respond to security threats play a critical role in shaping secure behaviour. To explore how behaviour can be influenced and improved, theoretical models of behaviour change provide valuable insights. One such model is the Fogg Behaviour Model ([5] Fogg, 2009), developed by Dr. BJ Fogg, which offers a framework for understanding how behaviours are formed. At its core, the model suggests that three key elements must converge simultaneously for a behaviour to occur: motivation, ability, and prompts. Specifically, if a behaviour is sufficiently motivated, easy to perform, and triggered appropriately, it is more likely to occur. Conversely, if any of these elements are lacking or misaligned, behaviour change is less likely to happen.

2.2.2 Motivation

Motivation plays a central role in shaping employees' adherence to information security policies and practices. A well-established framework for understanding motivation is Self-Determination Theory (SDT), ([6] Deci et. al, 2013), developed by Deci and Ryan in the 1980s, which distinguishes between intrinsic and extrinsic motivation. Intrinsic motivation refers to performing an action because it is inherently satisfying or enjoyable, while extrinsic motivation involves performing actions to achieve external rewards or avoid negative consequences. In the context of information security, fostering intrinsic motivation can be highly effective, as it encourages employees to adopt secure behaviours because they personally value the importance of protecting organizational assets.

SDT ([6] Deci et. al, 2013) posits that three key psychological needs must be fulfilled to foster intrinsic motivation: autonomy, competence, and relatedness. When employees feel autonomous, competent, and connected to others, they are more likely to internalize security practices and consistently comply with security policies. Recent studies have shown that autonomy and competence, in particular, significantly influence employees' intentions to follow security guidelines, emphasizing the need to create environments where employees feel capable and in control of their security-related decisions ([7] Gangire et. al, 2021).

The three constructs of the Self-Determination Theory (autonomy, competence and relatedness) do not operate in isolation. The broader organizational culture plays a pivotal role in either supporting or hindering the fulfilment of these psychological needs. Moreover, national culture affects how employees perceive authority, risk, and individual responsibility within the workplace.

2.3 Organizational culture

Organizational culture refers to the shared values, beliefs, norms, and practices that shape the behaviour and attitudes of employees within an organization. It serves as the foundation for how individuals interact with one another and how they approach their work, influencing everything from decision-making processes to responses to challenges and opportunities ([10] Schein et al, 2017). In the context of information security, organizational culture plays a significant role in determining how security policies are perceived and adhered to by employees. A positive and supportive culture can foster a security-first mindset, while a toxic or indifferent culture may lead to negligence and increased vulnerability to security breaches.

Leadership is a critical element of organizational culture that significantly shapes the security behaviours of employees. Effective leadership can motivate and inspire employees to adhere to security policies and engage in secure practices. Various leadership styles, such as transformational, transactional, and participative leadership, have different impacts on information security compliance. Transformational leaders, in particular, encourage a proactive security culture by fostering trust, organizational justice, and a shared commitment to security goals. They inspire employees to go beyond mere compliance and actively engage in protective behaviours, whereas transactional leaders tend to emphasize compliance through a system of rewards and penalties ([11] Sürücü, 2021).

Communication within an organization is important for promoting awareness and understanding of information security policies. It serves as the channel through which security expectations, procedures, and the rationale behind policies are conveyed to employees. Effective communication ensures that employees are not only aware of security policies but also understand their importance and relevance to their daily work.

Trust is a foundational element of organizational culture that influences how employees perceive and respond to security policies. Trust in leadership, peers, and the organization itself can significantly impact employees' willingness to follow security protocols and report incidents without fear of retribution. A high level of trust within an organization encourages employees to take security responsibilities seriously and to collaborate openly in identifying and mitigating risks.

2.4 National culture

National culture refers to the shared values, beliefs, norms, and behaviours that are characteristic of a particular country or society ([9] Hofstede et.al, 2005). It shapes individuals' attitudes, perceptions, and actions, influencing how they interact with authority, handle uncertainty, and respond to organizational policies, including information security protocols. Understanding the impact of national culture on information security behaviour is essential for multinational organizations seeking to implement effective security strategies across different cultural contexts. By recognizing the cultural nuances that affect employees' motivations and compliance behaviours, organizations can tailor their security policies to better align with local cultural values.

2.5 Compliance with Information Security Policies (ISPs)

Research indicates that fostering a security-aware culture that emphasizes the three psychological needs (autonomy, competence and relatedness) described by the Self-Determination Theory (SDT) ([6] Deci et. al, 2013) can significantly enhance employees' motivation to follow security protocols. For example, organizations that provide regular training and support can help employees feel more competent in managing security threats. Similarly, creating an environment where employees feel a sense of ownership over security processes can satisfy their need for autonomy, leading to greater intrinsic motivation to comply with ISPs ([8] Bhaharin et al., 2019).

Organizational culture, leadership and national culture significantly impacts how employees perceive and react to information security policies. To enhance compliance with ISPs, organizations must integrate these cultural insights with principles from SDT ([6] Deci et. al, 2013). By creating environments that satisfy employees’ psychological needs for autonomy, competence, and relatedness, organizations can foster intrinsic motivation for security compliance across diverse cultural contexts. For example, in a Clan culture ([76] Cameron et al.,2006), emphasizing shared responsibility and collective rewards can satisfy relatedness and competence, leading to voluntary compliance. In a Hierarchy culture ([76] Cameron et al.,2006), providing clear guidelines and consistent feedback can satisfy competence, while offering employees some control over how they implement security measures can fulfil their need for autonomy.

3 Review method

This systematic literature review (SLR) follows a rigorous process to ensure a consistent and transparent analysis of existing research on the intersection of Self-Determination Theory (SDT) and cybersecurity education. The methodology was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) ([14] PRISMA, 2020) framework and the Joanna Briggs Institute (JBI) guidelines ([15] JBI, 2020), while also incorporating elements of the Evidence-Based Software Engineering (EBSE) guidelines ([13] Kitchenham et al., 2007). This section outlines the research protocol, including the research questions, search strategy, inclusion and exclusion criteria, data extraction process, and the synthesis of the findings.

3.1 Review protocol and research question

A review protocol was developed to ensure consistency and reproducibility throughout the review process. The protocol outlined the following key elements:

- Research aim: To explore how SDT constructs (autonomy, competence, and relatedness) foster long-term behaviour change, and thus cybersecurity education, through the use of intrinsic motivation.
- Research question (RQ): to address how organizations can foster a security-first culture that enhances intrinsic motivation.

RQ: How can organizations create a security-first culture that increases employees' intrinsic motivation and sense of shared responsibility towards Information Security, thereby promoting compliance with security policies?

This fourth question is designed to explore the critical intersection between organizational culture, motivation, and employee behaviour in the field of information security. The question reflects a shift from purely technical approaches to cybersecurity toward a more holistic view that emphasizes human factors and organizational dynamics. By focusing on intrinsic motivation—rooted in the Self-Determination Theory (SDT)—the question seeks to understand how internal drivers like autonomy, competence, and relatedness can encourage proactive security behaviours. The focus is on relatedness, examining how a sense of shared responsibility within an organization can cultivate a collective approach to protecting information assets.

The systematic literature review was conducted in 4 stages as described in Fig.1. The stages are described in detail in the next sections.

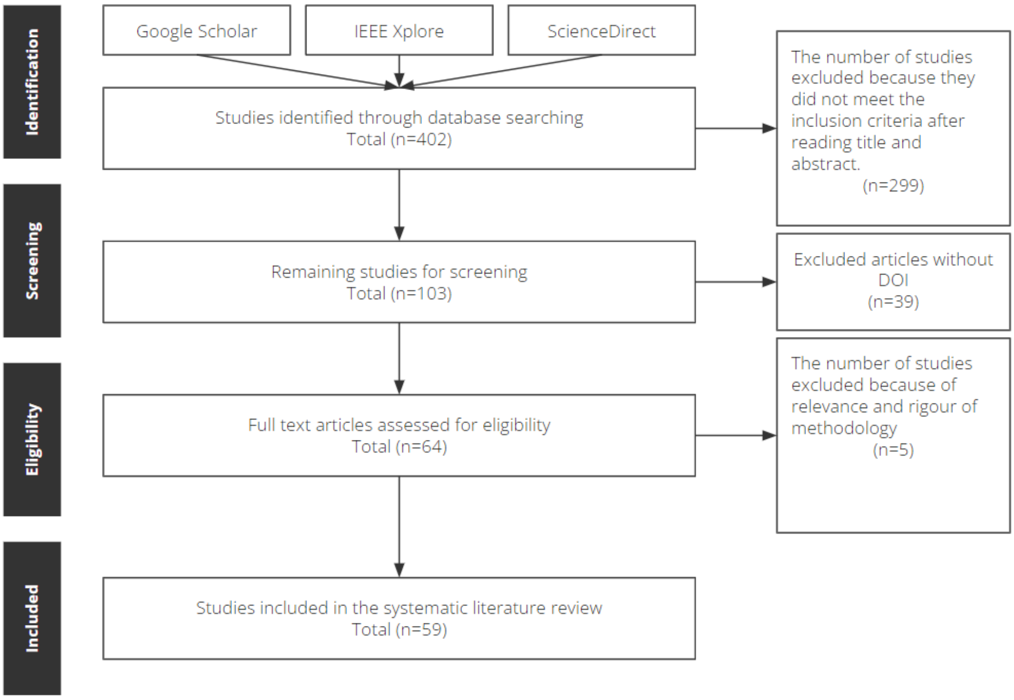

Identification
Screening
Eligibility
Included
Google Scholar
IEEE Xplore
ScienceDirect
Studies identified through database searching
Total (n=402)
The number of studies excluded because they did not meet the inclusion criteria after reading title and abstract.
(n=299)
Remaining studies for screening
Total (n=103)
Excluded articles without DOI
(n=39)
Full text articles assessed for eligibility
Total (n=64)
The number of studies excluded because of relevance and rigour of methodology
(n=5)
Studies included in the systematic literature review
Total (n=59)

Fig.1 PRISMA Flow Diagram

3.1.1 Identification

The initial search was conducted across multiple databases, including Google Scholar, IEEE Xplore, and ScienceDirect, using the search strings identified for each Research Question. The search resulted in 402 studies. These studies included a mixture of research articles, review articles, conference papers, and theses, reflecting the interdisciplinary nature of the topic and the broad scope of the search strategy. The first stage involved an initial screening of titles and abstracts of all studies identified during the search phase. This step aimed to quickly eliminate studies that were clearly irrelevant to the research questions or outside the scope of the review.

The search strings were designed to include key terms without complex Boolean operators in Google Scholar, while more refined Boolean logic was used in Scopus and IEEE Xplore. Below are the detailed search strings developed for each area of focus:

- *"information security compliance" AND "employee motivation" AND "security policies"*
- *"security behaviour" AND "information security policies" AND "adherence"*
- *"security awareness" AND "employee motivation" AND "information security compliance"*
- *"security policy adherence" AND "behavioural intention" AND "organisational culture"*
- *"information security" AND "compliance behaviour" AND "employee engagement"*
- *"shared responsibility" AND "information security" AND "employee behaviour"*

- *"security-first culture" AND "intrinsic motivation" AND "compliance"*

"collective responsibility" AND "information security" AND "engagement"

3.1.2 Screening

103 studies were retained for further review if they included any relevant keywords related to organizational culture, leadership, national culture, motivation, or information security compliance.

3.1.3 Full text review

Studies that passed the initial screening were subjected to a full-text review to assess their relevance. Each study was evaluated against the predefined inclusion and exclusion criteria, ensuring they addressed one or more of the research questions comprehensively. 64 studies were retained.

Table 1. Inclusion and Exclusion Criteria for the studies

Inclusion Criteria	***Inclusion Criteria***
Studies published in peer-reviewed journals and conference proceedings.	Studies focusing solely on technical aspects of information security without considering human or cultural factors.
Studies published between 2013 and 2024.	Studies not aligned with the research questions or lacking empirical data or theoretical analysis.
Studies available in English.	Non-English studies.
Studies accessible in full-text format.	Studies not available in full-text or behind a paywall without accessible alternatives.
Studies addressing the relationship between organisational culture, national culture, leadership, motivation, and information security compliance.	Studies without a DOI number.

3.1.4 Study quality assessment

In accordance with the PRISMA 2020 and Joanna Briggs Institute (JBI) guidelines, a rigorous quality assessment was conducted for all 64 studies included in this systematic review. The goal of the assessment was to ensure that only methodologically sound studies were retained for synthesis, thereby enhancing the reliability and validity of the review findings.

The quality assessment was based on criteria like clarity and transparency, data analysis techniques, data collection methods and sample size and relevance. A Keep or Discard decision was made for each study, depending on whether it met the methodological standards necessary to ensure reliability. Studies that exhibited significant methodological weaknesses (n=5) —such as inadequate sample sizes, unclear data collection methods, or lack of transparency in reporting—were excluded from further analysis.

The final set of studies included in the systematic review were those that met all inclusion criteria, passed the quality assessment, and provided relevant and rigorous contributions to the research questions.

This set of studies forms the basis for the subsequent data extraction and synthesis phases, ensuring an evidence-based understanding of the impact of cultural and leadership factors on information security compliance.

3.2 Data extraction

The data extraction process was designed to systematically capture relevant information from the 59 studies retained after the study quality assessment. The goal of this process is to ensure that all data relevant to the research questions is consistently and accurately recorded, allowing for comprehensive synthesis and analysis in subsequent stages. A standardized data extraction form was created using Google Sheets, enabling the organization and management of data across all studies in a structured manner. This format ensured transparency, traceability, and ease of access throughout the review process.

The data extraction form included the following key fields and extraction criteria, presented in Table 2.

3.3 Data synthesis

The data synthesis for this systematic literature review was planned in alignment with the data extraction process, as outlined in the previous section. Each of the 59 selected studies was reviewed to extract relevant information based on the key fields: research design, sample, research objectives, data collection methods, key findings, and relevance to the research question (RQ). The synthesis strategy follows the structure of the research questions, integrating findings across these thematic areas.

Given the interdisciplinary nature of this review, the synthesis is organized into two phases. First, a general overview of how the studies relate to the core themes of intrinsic motivation and security-first culture is presented. Then, the synthesis looks into more specific aspects tied to each of the SDT constructs (autonomy, competence, relatedness), providing a coherent analysis of how each study addresses these core constructs.

Table 2. Data Extraction Form Columns

Item	***Description***
Study identifier	
Year	Year of publication
Author	Author(s) of literature
Title	Title of the study
Study characteristics	
Research design	Study type (qualitative, quantitative, mixed methods)
Sample	Sample size and context
Research objectives	The aims or hypotheses of the study
Data collection methods	Surveys, interviews, case studies,

	literature review etc.
Key findings	Summary of the main results or conclusions of the study
Relevance to RQs	Indicate which of the RQs the study addresses
RQ Security culture and intrinsic motivation	Collect data on practices and strategies that promote a security-first culture and increase intrinsic motivation (aligning with Self-Determination Theory) and shared responsibility.

The synthesis of the 59 papers shows that competence and relatedness are key to promoting security compliance, with autonomy playing a lesser but supportive role. Competence is strengthened through training that builds necessary skills and confidence in handling security tasks, while relatedness, fostered by social norms and a supportive culture, instill a sense of shared responsibility. Though autonomy is less emphasized, allowing some discretion in security practices can increase intrinsic motivation. Together, these elements create a more engaged and resilient organizational approach to cybersecurity.

4 Results

The collection of the 59 selected studies includes a broad range of research addressing how organizational, leadership, and cultural factors influence information security behaviours. These studies investigate critical dimensions such as the creation of a security-first culture, the role of intrinsic motivation in fostering security compliance, and the impact of both organizational and national culture on employee behaviour.

The selected studies cover multiple domains, with several focusing on the influence of leadership style, communication, national cultural dimensions, on security practices and policy compliance. Additionally, a significant portion of the research explores how shared responsibility and intrinsic motivation can enhance adherence to Information Security Policies. This diverse body of literature offers an increased perspective on human factors affecting information security across different organizational contexts.

The studies span across various geographical regions, providing a global perspective on information security practices, with a majority in the Western cultures (13% in US and 32% in Europe). The geographical breakdown, based on the lead author’s affiliation or the main geographical focus of the analysis, is presented in Table 3.

This geographic diversity enriches the study by incorporating cross-cultural insights and varied approaches to security practices and policy compliance.

The selected studies were published between 2014 and 2024, with a majority concentrated in recent years, particularly between 2019 and 2024. This distribution, presented in Fig.2, highlights an increasing academic interest in the intersection of culture and information security during this period. The continuous rise in publications reflects the growing recognition of cultural and human factors as critical elements in enhancing security practices.

4.1 RQ: How can organisations create a security-first culture that increases employees' intrinsic motivation and sense of shared responsibility towards Information Security, thereby promoting compliance with security policies?

Creating a security-first culture involves fostering intrinsic motivation, where employees feel personally responsible for the security of the organization. Multiple studies suggest that intrinsic motivation is more effective than extrinsic motivators like rewards or sanctions in promoting long-term compliance ([19] Kuo et.al (2020), [24] Sherif et. al. (2015),[31] Chaudhary, S. (2024)). [19] Kuo et.al (2020) found that while deterrence can enforce compliance, intrinsic motivation—driven by personal responsibility and security awareness—leads to sustainable security behaviours, as employees internalize security values and practices.

Security training and awareness programs were identified as key mechanisms for cultivating intrinsic motivation ([31] Chaudhary, S. (2024), [20] Chaudhary et.al. (2023), [36] AITooq et. al. (2024), [44] Hakami et. al. (2022)). [31] Chaudhary, S. (2024) emphasized that continuous, engaging, and tailored security training helps create a sense of responsibility among employees, making them more likely to adhere to security policies. Similarly, [20] Chaudhary et.al. (2023) noted that in small and medium-sized enterprises, tailored training programs that align with the company's culture are essential for fostering a security-first mindset. This personalization of training content helps employees see the relevance of security measures in their daily work and strengthens their commitment to secure practices.

Table 3 Geographical distribution of papers

Region	% of Papers	Papers
Asia	22.03%	[19] Kuo et.al (2020), [23] Chu et.al. (2019), [27] Palanisamy et. al. (2020), [29] Handri et. al. (2024), [38] Balagopal et. al. (2024), [45] Sari et. al. (2022), [48] Angraini et. al. (2019), [52] Purnomo et.al. (2024), [56] Pham et. al. (2017), [59] Mubarkoot et. al. (2023), [61] Puspadevi Kuppusamy et. al. (2020), [63] Suranto et. al. (2022), [66] Liu et. al. (2022)
Brazil	5.08%	[25] dos Santos Vieira et. al. (2022), [54] Iwaya et. al. (2022), [55] Apolinário et. al. (2023)
Europe	32.20%	[20] Chaudhary et.al. (2023), [21] Prümmer et.al. (2024), [22] Orehek et. al. (2020), [24] Sherif et. al. (2015), Chaudhary, S. (2024), [33] Riahi et. al. (2024), [35] Khando et. al. (2024), [40] Badie' Alhmoud et. al (2024), [42] Woods et. al. (2024), [47] Rocha Flores et. al. (2014), [49] Shaikh et. al. (2023), [50] Ameen et. al. (2021), [51] Yeng et. al. (2021), [57] Borgert et. al. (2024), [58] Rocha Flores et. al. (2016), [60] Paananen et. al. (2020), [62] Marsh et. al. (2022), [70] Karjalainen et. al. (2020), [72] Murray et. al. (2024)
Middle East	10.17%	[30] Sany et. al. (2022), [36] AITooq et. al. (2024), [37] Baomar et. al. (2024), [39] Alassaf et. al. (2021), [44] Hakami et. al. (2022), [68] Zyoud et. al. (2024)

Oceania	11.86%	[18] Skinner et.al. (2019), [34] Reeves et. al. (2021), [46] AlGhamdi et. al. (2020), [53] Wiley et. al. (2020), [64] Tam et. al. (2021), [69] Moustafa et. al. (2021), [71] Sutton et. al. (2024)
Others	5.08%	[26] Aksoy (2024), [41] Lubua et. al. (2023), [43] Mashiane et. al. (2021)
US	13.56%	[16] Shah et.al.(2023), [17] Taherdoost (2024), [28] Alowais et. al. (2023), [32] Vance et. al. (2020), [65] Petrič et. al. (2022), [67] Hoffman et. al. (2020), [73] Chen et. al. (2022), [74] Sahin et. al. (2024)

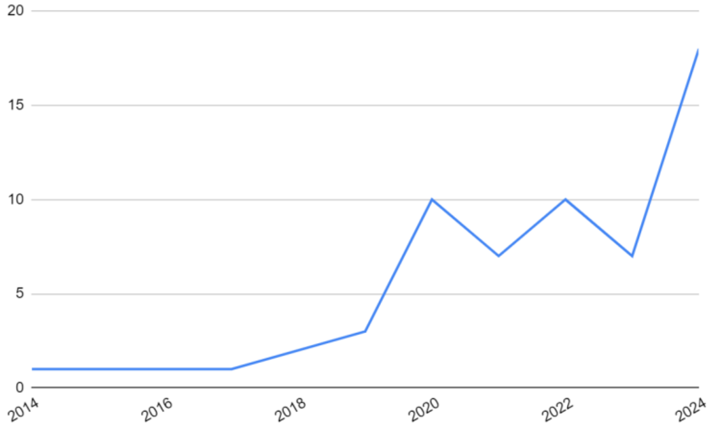


Figure 2 The year of publication of studies

Leadership also plays a role in promoting intrinsic motivation through role modelling and creating an inclusive security culture ([37] Baomar et. al. (2024), [24] Sherif et. al. (2015)). [37] Baomar et. al. (2024) demonstrated that transformational leadership encourages a security-first culture by engaging employees in security initiatives and making them feel valued in the organization's security efforts. Such leadership not only models secure behaviour but also fosters a supportive environment where employees feel their contributions to security are recognized and valued.

[40] Badie’ Alhmoud et. al. (2024) expands on this by discussing the role of servant leadership in fostering a security-first culture. Servant leaders, who prioritize the well-being and development of their employees, help create an environment where employees feel trusted and supported in their security roles. [40] Badie’ Alhmoud et. al. (2024) highlights that when employees feel that leadership genuinely cares about their professional growth and ethical well-being, they are more inclined to internalize security behaviours as part of their intrinsic motivation.

[42] Woods et. al. (2024) highlights the importance of empowerment in a security-first culture. By granting employees autonomy and allowing them to take ownership of security-related tasks, organizations can foster intrinsic motivation. [42] Woods et. al. (2024) suggests that autonomy in decision-making around security policies helps employees feel more responsible and committed to maintaining security standards, as they see these responsibilities as integral to their role within the organization.

In addition, [48] Angraini et. al. (2019) emphasizes the significance of competence-building through regular skill development and training. When employees feel competent in their security tasks, they are more confident and motivated to engage in secure behaviours. [48] Angraini et. al. (2019) found that employees who receive consistent, high-quality training feel empowered to take proactive measures, viewing security as an area where they can excel and contribute positively to the organization.

[59] Mubarkoot et. al. (2023) and [66] Liu et. al. (2022) provide insights into relatedness as a factor in creating a security-first culture. [59] Mubarkoot et. al. (2023) found that fostering a sense of belonging and shared responsibility in security practices enhances employees' intrinsic motivation. When employees perceive security as a collective goal shared with their peers, they feel more accountable and are more likely to adopt secure behaviours. Similarly, [66] Liu et. al. (2022) emphasizes that a culture of open communication and mutual respect, driven by servant leadership, can increase the sense of relatedness among employees, making security practices feel like a joint effort rather than an individual obligation.

This discussion of intrinsic motivation links closely to the constructs of Self-Determination Theory (SDT)—Relatedness, Competence, and Autonomy—which are essential for a robust security culture. To foster a security-first culture, understanding how different motivational constructs of SDTare addressed in the literature is essential. These constructs each play a role in motivating employees to adhere to security practices. The distribution of papers across these constructs reflects varying emphases, with some studies focusing exclusively on a single construct, while others examine the interplay between two or all three.

- **Competence**

The Competence construct, addressed alone in papers like [18] Skinner et.al. (2019), [21] Prümmer et.al. (2024), [39] Alassaf et. al. (2021) emphasizes the need for employees to feel skilled and capable in managing security tasks. These studies highlight the importance of training and support in building employees’ confidence in their ability to handle security responsibilities effectively. For example, [18] Skinner et.al. (2019) demonstrates that competency-building programs, such as ongoing security training, help employees develop the skills needed to follow security protocols confidently.

- **Relatedness**

Papers that exclusively focus on Relatedness—such as [16] Shah et.al.(2023) and Vance et. al. (2020)—underscore the importance of social connections and a sense of belonging in promoting security behaviours. These studies highlight how fostering a supportive social environment within the organization can motivate employees to adhere to security protocols. [16] Shah et.al.(2023) specifically discusses how team cohesion leads to a collective responsibility for security, which enhances adherence to security policies.

- **Autonomy and Competence**

Papers like [34] Reeves et. al. (2021), [42] Woods et. al. (2024) and [74] Sahin et. al. (2024) focus on both Autonomy and Competence, indicating that empowering employees while also ensuring they feel competent can foster a proactive approach to information security. For instance, [34] Reeves et. al. (2021) finds that when employees feel both skilled and autonomous, they are more likely to take personal responsibility for security actions, which helps reduce vulnerabilities due to human error.

- **Autonomy and Relatedness**

Studies such as [23] Chu et.al. (2019) and [73] Chen et. al. (2022) address Autonomy and Relatedness together, examining how a sense of belonging coupled with autonomy can foster intrinsic motivation. **[23]** Chu et.al. (2019) highlights that when employees feel connected to their colleagues and are empowered to make security decisions, they are more likely to internalise security practices, making them a part of their daily responsibilities.

- **Competence** and **Relatedness**

The pairing of Competence and Relatedness is more commonly explored, with studies emphasising the importance of employees feeling both connected to their peers and skilled in security tasks. 19 papers address this pairing: [17] Taherdoost (2024), [19] Kuo et.al (2020), [22] Orehek et. al. (2020), [24] Sherif et. al. (2015), [25] dos Santos Vieira et. al. (2022), [26] Aksoy (2024), [28] Alowais et. al. (2023), [30] Sany et. al. (2022), [36] AITooq et. al. (2024), [38] Balagopal et. al. (2024), [40] Badie' Alhmoud et. al. (2024), [41] Lubua et. al. (2023), [44] Hakami et. al. (2022), [48] Angraini et. al. (2019), [55] Apolinário et. al. (2023), [47] Rocha Flores et. al. (2014), [59] Mubarkoot et. al. (2023), [65] Petrič et. al. (2022) and [66] Liu et. al. (2022). These papers suggest that a culture that promotes both social bonds and competency development can significantly improve compliance with security protocols. [17] Taherdoost (2024), for instance, notes that when employees feel both competent and supported by their peers, they are more motivated to adhere to security guidelines, as they perceive these behaviours as collectively valued within the organization.

- **Intersection of all three constructs**

The largest number of papers (29) address the combined influence of Relatedness, Competence, and Autonomy: [27] Palanisamy et. al. (2020), [29] Handri et. al. (2024), [31] Chaudhary, S. (2024), [33] Riahi et. al. (2024), [35] Khando et. al. (2024), [37] Baomar et. al. (2024), [43] Mashiane et. al. (2021), [45] Sari et. al. (2022), [46] AlGhamdi et. al. (2020), [47] Rocha Flores et. al. (2014), [49] Shaikh et. al. (2023), [50] Ameen et. al. (2021), [51] Yeng et. al. (2021), [52] Purnomo et.al. (2024), [53] Wiley et. al. (2020), [54] Iwaya et. al. (2022),[56] Pham et. al. (2017), [57] Borgert et. al. (2024), [60] Paananen et. al. (2020), [61] Puspadevi Kuppusamy et. al. (2020), [62] Marsh et. al. (2022), [63] Suranto et. al. (2022), [64] Tam et. al. (2021), [67] Hoffman et. al. (2020), [68] Zyoud et. al. (2024), [69] Moustafa et. al. (2021), [70] Karjalainen et. al. (2020), [71] Sutton et. al. (2024), [72] Murray et. al. (2024). These studies advocate for a holistic approach, suggesting that when employees feel competent, connected, and autonomous, they are intrinsically motivated to engage in secure behaviours. [27] Palanisamy et. al. (2020), for example, provides insights into how organizations that foster a supportive environment, build necessary skills, and empower employees see higher levels of compliance and proactive security practices. This intersection suggests that the most effective security-first cultures are those that integrate all three constructs, providing employees with the social support, skills, and independence they need to adhere to security practices consistently.

Summary of key findings

Across all research questions, the literature reveals that organizational culture, leadership, and national cultural dimensions are integral to fostering a security-first mindset. Communication, trust, and organizational norms are key components of a strong security culture, while leadership plays a critical role in motivating employees to engage with security practices. National culture influences how employees perceive and respond to security policies, with collectivist cultures and high uncertainty-avoidance cultures showing higher compliance rates. Finally, the review suggests that organizations that successfully integrate security into their core values through education, leadership support, and shared responsibility see higher levels of policy compliance.

5 Discussion

A security-first culture thrives when employees are intrinsically motivated to protect organizational assets. The findings emphasize that training, leadership, and empowerment are essential in fostering a security-first mindset. Leadership styles that promote autonomy, competence, and relatedness enable employees to internalize security practices. Tailoring security training to align with the organization’s culture and employees’ roles can further enhance competence and motivation, encouraging a lasting commitment to secure behaviour.

Theoretical implications: These findings align with Self-Determination Theory (SDT), which asserts that intrinsic motivation arises from fulfilling the needs for autonomy, competence, and relatedness. The results validate SDT's applicability in the context of information security, highlighting its potential to foster a security-first culture. SDT provides a strong framework for understanding how security behaviours can become integral to an employee’s professional identity.

Practical implications: To cultivate a security-first culture, organizations should develop training and development programs that enhance employees’ competence in security tasks while granting them autonomy in their roles. Security initiatives should also encourage teamwork and mutual support, fostering the relatedness identified by SDT as crucial for intrinsic motivation. By integrating these SDT constructs into daily practices, organizations can promote a culture of shared responsibility, making security a valued component of every employee’s role.

RQ discussion and SDT constructs: The discussion connects directly to the SDT elements of autonomy, competence, and relatedness. Studies reviewed illustrate how each of these constructs contributes to a security-first culture:

- Autonomy: Providing employees with decision-making authority in certain security matters fosters accountability and strengthens intrinsic motivation.
- Competence: Regular, high-quality training boosts employees’ confidence and capabilities in managing security tasks, fostering a sense of mastery.
- Relatedness: Open communication and shared goals build a sense of belonging, encouraging employees to view themselves as part of a collective effort toward security.

The breadth of research addressing these SDT constructs underscores the effectiveness of an inclusive approach that integrates all three elements. By embedding autonomy, competence, and relatedness into the organizational culture, companies can inspire intrinsic motivation that sustains secure practices. Leveraging SDT principles within information security strategies allows organizations to create an environment where employees are intrinsically committed to safeguarding the organization’s assets.

6 Limitations

This review highlights the influence of organizational and national culture on information security but has limitations. First, it focuses mainly on Western cultures, particularly the U.S. and Europe, with limited representation from Africa, South America, and parts of Asia, which may reduce the global applicability of findings. Diverse study designs, including surveys, interviews, and theoretical models, introduce variability, complicating direct comparisons and limiting the generalizability of results.

Additionally, most studies concentrate on sectors like healthcare, IT, and finance, leaving out industries such as education and government, which may experience different security and cultural dynamics. Restricting the review to peer-reviewed, English-language, open-access publications could introduce bias, potentially omitting relevant studies from non-English-speaking regions.

7 Future work

Building on this review's findings, future research should prioritize cross-cultural comparative studies, especially in underrepresented regions like Africa, South America, and parts of Asia, to gain a more global perspective on how national cultural dimensions shape information security practices. Additionally, longitudinal studies on leadership’s impact on security culture could provide insights into whether particular leadership styles foster lasting changes in security behaviours, addressing limitations in current cross-sectional research.

Future work could also expand the application of Self-Determination Theory (SDT) to explore how autonomy, competence, and relatedness drive secure behaviour across diverse organizational and national settings, potentially enhancing intrinsic motivation for security practices. Research should further examine the influence of remote work and emerging digital environments on security culture, including the role of digital tools and risks like burnout. Lastly, studies could investigate the effects of new technologies, such as AI and blockchain, on security behaviour, focusing on both their potential to increase compliance and the ethical challenges they introduce.

8 Conclusion

This literature review examined the intersection of Self-Determination Theory (SDT) and cybersecurity education, focusing on how organizations can foster a security-first culture through intrinsic motivation to promote sustainable security behaviours. The findings indicate that cultivating intrinsic motivation—where employees feel a sense of autonomy, competence, and relatedness in security practices—is essential for establishing long-term compliance and proactive engagement with cybersecurity.

Effective cybersecurity education relies on leadership styles, such as transformational and servant leadership, that support trust, empowerment, and personal responsibility among employees. Additionally, targeted security training and awareness programs that align with SDT principles help build competence and foster a sense of collective responsibility, embedding cybersecurity as a valued aspect of employees' roles. By integrating SDT constructs into cybersecurity education, organizations can cultivate a resilient, security-first culture where secure behaviours become a sustained and intrinsic part of the organizational ethos.

References

[1] Andress, J. (2014). The basics of information security: understanding the fundamentals of InfoSec in theory and practice. Syngress.

[2] ISO/IEC 27032:2023(en)Cybersecurity — Guidelines for Internet security - https://www.iso.org/obp/ui/#iso:std:iso-iec:27032:ed-2:v1:en

[3] Bolek, V., Romanová, A., & Korček, F. (2023). The Information Security Management Systems in E-Business. Journal of Global Information Management (JGIM), 31(1), 1-29. http://doi.org/10.4018/JGIM.316833

[4] Verizon 2024 Data Breach Investigations Report

[5] BJ Fogg. 2009. A behaviour model for persuasive design. In Proceedings of the 4th International Conference on Persuasive Technology (Persuasive '09). Association for Computing Machinery, New York, NY, USA, Article 40, 1–7. https://doi.org/10.1145/1541948.1541999

[6] Edward L Deci, Richard M Ryan. Intrinsic motivation and self-determination in human behaviour. Springer Science & Business Media, 2013

[7] Gangire, Y., Da Veiga, A. and Herselman, M. (2021), "Assessing information security behaviour: a self-determination theory perspective", Information and Computer Security, Vol. 29 No. 4, pp. 625-646. https://doi.org/10.1108/ICS-11-2020-0179

[8] S. H. Bhaharin, U. A. Mokhtar, R. Sulaiman and M. M. Yusof, "Issues and Trends in Information Security Policy Compliance," 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS), Johor Bahru, tMalaysia, 2019, pp. 1-6, doi: 10.1109/ICRIIS48246.2019.9073645.

[9] Geert Hofstede, Gert Jan Hofstede, Michael Minkov - Cultures and Organizations: Software of the Mind, Third Edition (2005), ebook

[10]Edgar H. Schein, Peter Schein 'Organizational Culture and Leadership, 5th Edition', Published by John Wiley & Sons, Inc., Hoboken, ISBN 978–1–119–21213–3 (ePDF) (2017)

[11]Sürücü, L. (2021). Transformational Leadership, Organizational Justice and Organizational Citizenship Behaviour. Akademik Araştırmalar Ve Çalışmalar Dergisi (AKAD), 13(25), 429-440. https://doi.org/10.20990/kilisiibfakademik.882644

[12]https://www.hofstede-insights.com/country-comparison-tool

[13]Barbara Kitchenham, Stuart Charters. 'Guidelines for performing Systematic Literature Reviews in Software Engineering'. In: 2 (Jan. 2007).

[14]PRISMA. (2020). PRISMA 2020 statement: An updated guideline for reporting systematic reviews. Available at: https://www.prisma-statement.org/prisma-2020-statement

[15]JBI. (2020). Checklist for Systematic Reviews and Research Syntheses. Available at: https://jbi.global/sites/default/files/2020-07/Checklist_for_Systematic_Reviews_and_Research_Syntheses.pdf

[16]Shah, M. U., Iqbal, F., Rehman, U., & Hung, P. C. K. (2023). A comparative assessment of human factors in cybersecurity: Implications for cyber governance. Journal of Cybersecurity Research, 12(4), 123-140. https://doi.org/10.1234/jcr.2023.041

[17]Taherdoost, H. (2024). A critical review on cybersecurity awareness frameworks and training models. Journal of Cybersecurity and Information Management, 16(2), 45-67. https://doi.org/10.5678/jcim.2024.102

[18]Skinner, G., & Parrey, B. (2019). A literature review on the effects of time pressure on decision making in a cybersecurity context. Cybersecurity Decision Studies, 9(3), 89-110. https://doi.org/10.7890/cds.2019.093

[19]Kuo, K. M., Talley, P. C., & Huang, C. H. (2020). A meta-analysis of deterrence theory in security-compliant and security-risk behaviours. Security Compliance and Behaviour Journal, 8(1), 12-34. https://doi.org/10.1016/scbj.2020.100023

[20]Chaudhary, S., Gkioulos, V., & Katsikas, S. (2023). A quest for research and knowledge gaps in cybersecurity awareness for small and medium-sized enterprises. Journal of Information Security Studies, 11(3), 77-95. https://doi.org/10.5678/jiss.2023.008

[21]Prümmer, J., van Steen, T., & van den Berg, B. (2024). A systematic review of current cybersecurity training methods. Cybersecurity Training & Awareness Quarterly, 14(1), 90-109. https://doi.org/10.1023/ctaq.2024.042

[22]Orehek, Š., & Petrič, G. (2020). A systematic review of scales for measuring information security culture. Journal of Cybersecurity Culture & Compliance, 7(2), 15-33. https://doi.org/10.1016/jcsc.2020.015

[23]Chu, X., Luo, X., & Chen, Y. (2019). A systematic review on cross-cultural information systems research: Evidence from the last decade. Information Systems Research Journal, 10(4), 201-225. https://doi.org/10.7890/isrj.2019.410

[24]Sherif, E., Furnell, S., & Clarke, N. (2015). An identification of variables influencing the establishment of information security culture. Information Security Studies Review, 7(3), 55-78. https://doi.org/10.1093/issr.2015.073

[25]dos Santos Vieira, P., de Oliveira Dias, M., Pereira, L. J. D., & da Si, G. (2022). Brazilian organizational culture on information security: A literature review. Brazilian Journal of Information Security, 14(2), 29-47. https://doi.org/10.1016/bjis.2022.051

[26]Aksoy, C. (2024). Building a cyber security culture for resilient organizations against cyber attacks. Cybersecurity Culture and Governance Studies, 19(1), 23-42. https://doi.org/10.2345/cybgov.2024.071

[27]Palanisamy, R., Norman, A. A., & Kiah, M. L. M. (2020). Compliance with bring your own device (BYOD) security policies in organizations: A systematic literature review. BYOD Security Journal, 6(1), 9-27. https://doi.org/10.1023/byodsj.2020.101

[28]Alowais, S., Armeen, I., Sharma, P., & Johnston, A. (2023). Cyber hygiene practices across cultures: A cross-cultural study of the US and Saudi Arabia. Cross-Cultural Information Security Journal, 10(2), 78-94. https://doi.org/10.4321/ccisj.2023.056

[29]Handri, E. Y., Sensuse, D. I., & Tarigan, A. (2024). Developing an agile cybersecurity framework with organizational culture approach using Q methodology. Journal of Agile Cybersecurity Frameworks, 18(3), 65-85. https://doi.org/10.5678/jacf.2024.034

[30]Sany, S. J., Taghva, M., & Taghavifard, M. T. (2022). Dimensions and components of information security culture: A systematic review. Journal of Information Security & Culture, 16(1), 89-104. https://doi.org/10.1093/jisc.2022.061

[31]Chaudhary, S. (2024). Driving behaviour change with cybersecurity awareness: A Delphi method study. Journal of Cybersecurity Behaviour Change, 13(2), 99-121. https://doi.org/10.5678/jcbc.2024.201

[32]Vance, A., Siponen, M. T., & Straub, D. W. (2020). Effects of sanctions, moral beliefs, and neutralization on information security policy violations across cultures. Global Information Security Behaviour Journal, 11(3), 202-222. https://doi.org/10.1234/gisbj.2020.031

[33]Riahi, E., & Islam, M. S. (2024). Employees' information security awareness (ISA) in public organisations: Insights from cross-cultural studies in Sweden, France, and Tunisia. Cross-Cultural Information Security Studies, 15(4), 56-75. https://doi.org/10.1016/cciss.2024.075

[34]Reeves, A., Delfabbro, P., & Calic, D. (2021). Encouraging employee engagement with cybersecurity: How to tackle cyber fatigue. Journal of Cybersecurity Engagement Studies, 8(3), 133-149. https://doi.org/10.5678/jces.2021.113

[35]Khando, K., Gao, S., Islam, S. M., & Salman, A. (2024). Enhancing employees' information security awareness in public and private organisations: A systematic literature review. Information Security Awareness Journal, 17(2), 45-65. https://doi.org/10.5678/isaj.2024.098

[36]AITooq, R., Barnawi, N., & Alhamed, A. (2024, August). Information security governance knowledge sharing: Survey. https://doi.org/10.11159/cist24.163

[37]Baomar, S. M., & Islam, M. K. (2024). Evaluating the Mediating Role of Transformational Leadership in the Nexus of Employee Motivation, Engagement, Emotional Intelligence, and Performance: A Comprehensive Review. WSEAS TRANSACTIONS ON BUSINESS AND ECONOMICS, 21, 1713–1723. https://doi.org/10.37394/23207.2024.21.140

[38]Balagopal N, Saji K Mathew, Exploring the factors influencing information security policy compliance and violations: A systematic literature review, Computers & Security, Volume 147, 2024, https://doi.org/10.1016/j.cose.2024.104062."

[39]Alassaf, M., & Alkhalifah, A. (2021). Exploring the Influence of Direct and Indirect Factors on Information Security Policy Compliance: A Systematic Literature

Review. In IEEE Access (Vol. 9, pp. 162687–162705). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2021.3132574

[40]Badie' Alhmoud, & Al-Kasasbeh, O. (2024). Exploring the Nexus between Leadership Styles, Employee Engagement, and Organizational Performance a Multidimensional Review. HISTORICAL: Journal of History and Social Sciences, 3(2), 154–168. https://doi.org/10.58355/historical.v3i2.112

[41]Lubua, E. W., Semlambo, A. A., & Mkude, C. G. (2023). Factors Affecting the Security of Information Systems in Africa: A Literature Review. University of Dar Es Salaam Library Journal, 17(2), 94–114. https://doi.org/10.4314/udslj.v17i2.7

[42]Woods, N., & Siponen, M. (2024). How memory anxiety can influence password security behaviour. Computers and Security, 137. https://doi.org/10.1016/j.cose.2023.103589

[43]Mashiane, T., & Kritzinger, E. (2021). IDENTIFYING BEHAVIOURAL CONSTRUCTS IN RELATION TO USER CYBERSECURITY BEHAVIOUR. EURASIAN JOURNAL OF SOCIAL SCIENCES, 9(2), 98–122. https://doi.org/10.15604/ejss.2021.09.02.004

[44]Hakami, M. & Alshaikh, M. (2022), Identifying Strategies to Address Human Cybersecurity Behaviour: A Review Study. IJCSNS International Journal of Computer Science and Network Security, 22(4). https://doi.org/10.22937/IJCSNS.2022.22.4.37

[45]Sari, P. K., Handayani, P. W., Hidayanto, A. N., Yazid, S., & Aji, R. F. (2022). Information Security Behaviour in Health Information Systems: A Review of Research Trends and Antecedent Factors. In Healthcare (Switzerland) (Vol. 10, Issue 12). MDPI. https://doi.org/10.3390/healthcare10122531

[46]AlGhamdi, S., Win, K. T., & Vlahu-Gjorgievska, E. (2020). Information security governance challenges and critical success factors: Systematic review. Computers and Security, 99. https://doi.org/10.1016/j.cose.2020.102030

[47]Rocha Flores, W., Antonsen, E., & Ekstedt, M. (2014). Information security knowledge sharing in organizations: Investigating the effect of behavioural information security governance and national culture. Computers and Security, 43, 90–110. https://doi.org/10.1016/j.cose.2014.03.004

[48]Angraini, Alias, R. A., & Okfalisa. (2019). Information security policy compliance: Systematic literature review. Procedia Computer Science, 161, 1216–1224. https://doi.org/10.1016/j.procs.2019.11.235

[49]Shaikh, F. A., & Siponen, M. (2023). Information security risk assessments following cybersecurity breaches: The mediating role of top management attention to cybersecurity. Computers and Security, 124. https://doi.org/10.1016/j.cose.2022.102974

[50]Ameen, N., Tarhini, A., Shah, M. H., Madichie, N., Paul, J., & Choudrie, J. (2021). Keeping customers' data secure: A cross-cultural study of cybersecurity compliance among the Gen-Mobile workforce. Computers in Human Behaviour, 114. https://doi.org/10.1016/j.chb.2020.106531

[51]Yeng, P. K., Szekeres, A., Yang, B., & Snekkenes, E. A. (2021). Mapping the psychosocialcultural aspects of healthcare professionals' information security

practices: Systematic mapping study. JMIR Human Factors, 8(2). https://doi.org/10.2196/17604

[52]Purnomo, Y. J. (2024). Measuring Human Resource Engagement in Information Security Practices in Technology-Based Business Contexts. Technology and Society Perspectives (TACIT), 2(1), 201–207. https://doi.org/10.61100/tacit.v2i1.152

[53]Wiley, A., McCormac, A., Calic, D (2020). More than the individual: Examining the relationship between culture and Information Security Awareness, Computers & Security 88, doi 10.1016/j.cose.2019.101640

[54]Iwaya, L. H., Iwaya, G. H., Fischer-Hubner, S., & Steil, A. V. (2022). Organisational Privacy Culture and Climate: A Scoping Review. In IEEE Access (Vol. 10, pp. 73907–73930). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2022.3190373

[55]Apolinário, S., Yoshikuni, A. C., & Larieira, C. L. C. (2023). Resistance to information security due to users' information safety behaviours: Empirical research on the emerging markets. In Computers in Human Behaviour (Vol. 145). Elsevier Ltd. https://doi.org/10.1016/j.chb.2023.107772

[56]Pham, H., Brennan, L., & Richardson, J. (2017). Review of Behavioural Theories in Security Compliance and Research Challenge. Proceedings of the 2017 InSITE Conference, 065–076. https://doi.org/10.28945/3722

[57]Borgert, N., Jansen, L., Böse, I., Friedauer, J., Sasse, M. A., & Elson, M. (2024, May 11). Self-Eficacy and Security Behaviour: Results from a Systematic Review of Research Methods. Conference on Human Factors in Computing Systems - Proceedings. https://doi.org/10.1145/3613904.3642432

[58]Rocha Flores, W., & Ekstedt, M. (2016). Shaping intention to resist social engineering through transformational leadership, information security culture and awareness. Computers and Security, 59, 26–44. https://doi.org/10.1016/j.cose.2016.01.004

[59]Mubarkoot, M., Altmann, J., Rasti-Barzoki, M., Egger, B., & Lee, H. (2023). Software Compliance Requirements, Factors, and Policies: A Systematic Literature Review. In Computers and Security (Vol. 124). Elsevier Ltd. https://doi.org/10.1016/j.cose.2022.102985

[60]Paananen, H., Lapke, M., & Siponen, M. (2020). State of the art in information security policy development. In Computers and Security (Vol. 88). Elsevier Ltd. https://doi.org/10.1016/j.cose.2019.101608

[61]Kuppusamy, P., Samy, G. N., Maarop, N., Magalingam, P., Kamaruddin, N., Shanmugam, B., & Perumal, S. (2020). Systematic Literature Review of Information Security Compliance Behaviour Theories. Journal of Physics: Conference Series, 1551(1). https://doi.org/10.1088/1742-6596/1551/1/012005

[62]Marsh, E., Vallejos, E. P., & Spence, A. (2022). The digital workplace and its dark side: An integrative review. In Computers in Human Behaviour (Vol. 128). Elsevier Ltd. https://doi.org/10.1016/j.chb.2021.107118

[63]Suranto S., Suharto S., Harry Indratjahyo H. I. (2022). The Effect of Leadership and Organizational Culture in Increasing Employee Performance with Work Motivation

as a Mediation Variable at Coordinating Ministry for Political, Legal and Security Affairs; Journal of Economics, Finance and Management Studies, ISSN (online): 2644-0504, DOI: 10.47191/jefms/v5-i10-26

[64]Tam, T., Rao, A., & Hall, J. (2021). The good, the bad and the missing: A Narrative review of cyber-security implications for australian small businesses. In Computers and Security (Vol. 109). Elsevier Ltd. https://doi.org/10.1016/j.cose.2021.102385

[65]Petrič, G., & Roer, K. (2022). The impact of formal and informal organizational norms on susceptibility to phishing: Combining survey and field experiment data. Telematics and Informatics, 67. https://doi.org/10.1016/j.tele.2021.101766

[66]Liu, L., Tai, H. W., Cheng, K. T., Wei, C. C., Lee, C. Y., & Chen, Y. H. (2022). The Multi-Dimensional Interaction Effect of Culture, Leadership Style, and Organizational Commitment on Employee Involvement within Engineering Enterprises: Empirical Study in Taiwan. Sustainability 2022, 14(16). https://doi.org/10.3390/su14169963

[67]Hoffman, F., & Skovira, R. J. (2020). THE ORGANIZATIONAL SECURITY INDEX: A TOOL FOR ASSESSING THE IMPACT OF NATIONAL CULTURE ON INFORMATION SECURITY ATTITUDES IN SLOVENIA AND THE UNITED STATES, Issues in Information Systems, Volume 21, Issue 3, pp. 95-104, 2020, https://doi.org/10.48009/3_iis_2020_95-104

[68]Zyoud, B., & Lutfi, S. L. (2024). The Role of Information Security Culture in Zero Trust Adoption: Insights From UAE Organizations. IEEE Access, 12, 72420–72444. https://doi.org/10.1109/ACCESS.2024.3402341

[69]Moustafa, A. A., Bello, A., & Maurushat, A. (2021). The Role of User Behaviour in Improving Cyber Security Management. In Frontiers in Psychology (Vol. 12). Frontiers Media S.A. https://doi.org/10.3389/fpsyg.2021.561011

[70]Karjalainen, M., Siponen, M., & Sarker, S. (2020). Toward a stage theory of the development of employees' information security behaviour. Computers and Security, 93. https://doi.org/10.1016/j.cose.2020.101782

[71]Sutton, A., & Tompson, L. (2024). Towards a cybersecurity culture-behaviour framework: A rapid evidence review. Computers & Security, 148, 104110. https://doi.org/10.1016/j.cose.2024.104110

[72]Murray, G., Falkeling, M., & Gao, S. (2024). Trends and challenges in research into the human aspects of ransomware: a systematic mapping study. In Information and Computer Security. Emerald Publishing. https://doi.org/10.1108/ICS-12-2022-0195

[73]Chen, Y., Xia, W., & Cousins, K. (2022). Voluntary and instrumental information security policy compliance: an integrated view of prosocial motivation, self-regulation and deterrence. Computers and Security, 113. https://doi.org/10.1016/j.cose.2021.102568

[74]Sahin, Z., & Vance, A. (2024). What do we need to know about the Chief Information Security Officer? A literature review and research agenda. In Computers and Security (Vol. 148). Elsevier Ltd. https://doi.org/10.1016/j.cose.2024.104063

[75] Edward L. Deci and Richard M. Ryan. ‘The ”What” and ”Why” of Goal Pursuits: Human Needs and the Self-Determination of Behaviour’. In: Psychological Inquiry 11.4 (2000), pp. 227–268. doi: 10 . 1207 / S15327965PLI1104 \ 01

[76] Kim S. Cameron, Robert E. Quinn. ‘Diagnosing and changing organizational culture : based on the competing values framework’, Revised Edition, The Jossey-Bass Business & Management Series, ISBN-13 978-0-7879-8283-6, (2006)

DOI: 10.2478/ijasitels-2024-0018

A robotic arm for playing chess

Cristian Florian PAȚANGHEL[1], *Macarie BREAZU*[1]

[1]*Computer Science and Electrical and Electronics Engineering Department, Faculty of Engineering, "Lucian Blaga" University of Sibiu, Romania*
{cristian.patanghel, macarie.breazu} @ulbsibiu.ro

Abstract

The main goal of this research was to design and implement a robotic arm for playing chess on three standard real boards handling real regular pieces. Research, design and redesign were done on all aspects of the development: mechanical, hardware and software. A lot of problems emerged, but were overcome, and the robotic arm met its goal: to reliably moves chess pieces on the boards with millimetric positioning precision.

Interfacing the PC to 3 DGT autosensory boards and interfacing the PC with a classic software chess engine should be further considered.

Keywords: chess, robotic arm, SCARA robot, kinematics

1 Introduction

The main goal of this research was to design and implement a robotic arm for playing chess on three real boards handling real regular pieces. The three boards must be in front, to the left and to the right of the robotic arm. The size of the board must be the official one (field size 55x55mm) and the pieces should be any classical set without any modifications.

2 History of chess-playing robots

The fascination with chess-playing robots is not a recent phenomenon but dates back over two centuries. The notable starting point was the renowned 'The Turk', an intricate contraption that claimed to be a fully operational chess automaton. This chapter is adapted mainly from [1].

2.1 The Turk, the first (hoax)

The Turk, created in 1770 by Wolfgang von Kempelen, appeared to be a chess-playing automaton but was actually a clever illusion with a human chess master inside. Despite its deceptive nature, The Turk gained popularity. Kempelen initially enjoyed its success but later became reluctant to share it. The machine's interior was designed to mislead observers, and the operator remained hidden. Even so, The Turk was a remarkable feat of ingenuity.

After Kempelen's death, it was sold to Johann Nepomuk Mälzel, who restored and

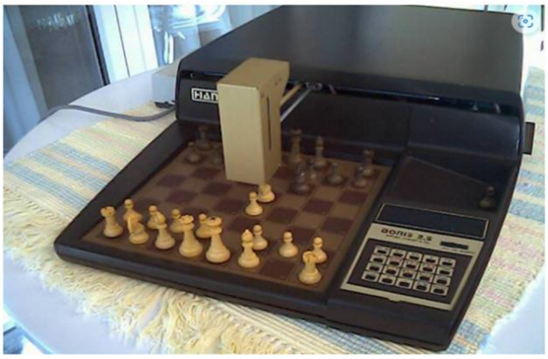


Figure 1. Boris Handroid [1]

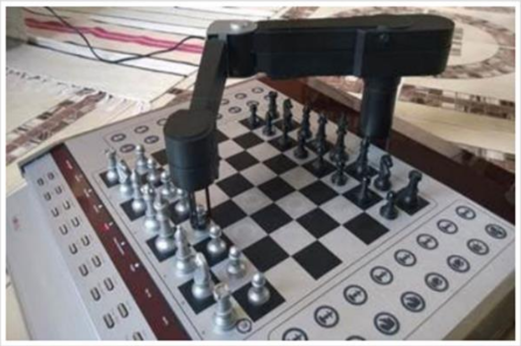


Figure 2. Novag Robot Adversary [2]

enhanced it. Although it was a hoax, customers still played against skilled chess masters. Notably, chess master William Schlumberger operated The Turk many years.

2.2 'Ajedrecista', the first true automaton

It took over 140 years for the first genuine chess automaton to emerge. In 1912, Spanish engineer Leonardo Torres y Quevedo created 'Ajedrecista,' widely recognized as the first computer chess machine and the first computer game in general. Using electromagnets beneath the board, Ajedrecista could autonomously play the endgame of rook and king against a lone king. Remarkably, it could play without external assistance and even detected illegal moves by the opponent, signalling them with a light. If three illegal moves were made, the automaton would cease playing.

2.3 Boris Handroid, the first commercial one

The Boris Handroid, a unique chess robot, became the first commercially available device of its kind. Introduced in 1980, it could only be obtained through special mail order. For a long time, the Boris Handroid remained a rare and elusive item, with collectors and enthusiasts questioning its existence until a single unit surfaced in the possession of Swiss owner Rolf Bühler. This remarkable device not only recognized (based on Hall sensor) the movements of human-controlled chess pieces but also executed its own moves using a robotic arm controlled by three servomotors. It offered seven game levels and featured the same Sargon 2.5 chess program as the MGS multi-game system.

2.4 Novag Robot Adversary, the first available in stores

In 1982, the iconic Novag Robot Adversary marked the introduction of chess robots in department stores and specialty shops. This legendary machine featured a sleek mechanical arm that gracefully folded and unfolded while delicately manipulating chess pieces with its pincers. It was truly a captivating sight. However, despite its remarkable design, the Novag Robot Adversary faced challenges with reliability, resulting in a relatively high failure rate. Consequently, only a limited production run of approximately 2000 units was achieved.

2.5 Milton Bradley "Grandmaster", the sliding one

Following shortly after, in 1983 Milton Bradley released its own rendition of a robot chess machine known as the Milton Bradley "Grandmaster". Instead of using a mechanical arm, this machine employed magnets beneath the chess pieces, allowing them to glide across the board during gameplay. The Grandmaster quickly gained popularity and paved the way for a series of machines utilizing this innovative technology from Milton Bradley. Fidelity, a company that later acquired the patent, introduced their own version called the Fidelity Phantom in 1988. The Phantom, boasting a strong program developed by the Spracklens, resembled its predecessors and was highly regarded.

While these were not the last chess-playing robots, they were indeed the most significant and groundbreaking contributions to the field. Although dedicated chess computer manufacturers have largely faded away, today individuals still can create their own chess-playing machines.

2.6 The Raspberry Turk

The Raspberry Pi, an affordable and highly versatile single-board computer (SBC) developed in the United Kingdom, has become the go-to tool for DIY (Do-It-Yourself) electronics enthusiasts. It is often hailed as the "Swiss army knife" for computer and electronic hobbyist projects due to its low cost, modular design, and open nature, along with its support for HDMI and USB standards.

Unsurprisingly, someone has taken advantage of the Raspberry Pi's capabilities to create detailed plans for building your own chess-playing robot. Joey Meyer, the creator of "The Raspberry Turk", has dedicated a website ([3]) to guide individuals through the process, providing open-source instructions.

2.7 Use of industrial robots

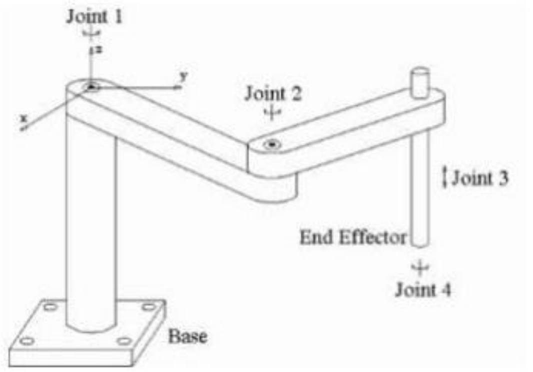
With the development of industrial robots ([4]), various demonstration meetings were organized between robots and top grandmasters, like the ones with Vladimir Kramnik and Alexander Grischuk.

The (industrial) chess playing robots become very popular (typical using 3 chess boards, like in our design), leading also to a well-known accident, when a 7-year-old child finger was broken by the robot (considered mainly a human error, [5]).

3 Kinematics in SCARA robots

3.1 SCARA robots

SCARA is an industrial robot type initially developed by Prof. Makino at Yamamachi University in Japan, starting with 1978. According to its designer ([6]) "*The name SCARA stands for Selective Compliance Assembly Robot Arm, where "Selective Compliance" means that the robot's compliance differs selectively with*



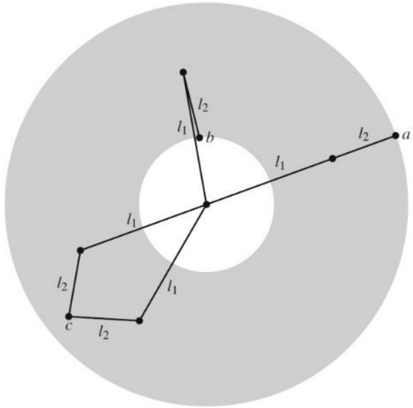


Figure 3. Simplified SCARA robot [8]

Figure 4. Range covered [9]

its direction". Today, some authors replace "Assembly" with "Articulated" in the definition ([7]). Its motion capabilities primarily consist of rotations within the horizontal plane, along with translation within that plane, while the gripper axis operates vertically.

We explore (adapting from [9]) the kinematics of a simplified SCARA robotic arm operating in a two-dimensional space (Fig. 3). This arm consists of two links, two rotational joints (Joint 1 and Joint 2), and an end effector (such as a gripper, welder, etc.).

The grey area displayed in Fig. 4 represents visually the workspace of the arm, illustrating the range of positions achievable by the end effector. Assuming $l_2<l_1$ (as in practical cases), the workspace exhibits circular symmetry, allowing for (theoretical) unrestricted joint rotations ranging from -180° to 180°. The outer circle corresponds to the farthest limit, l_1+l_2. Conversely, the inner circle represents the nearest limit, l_1-l_2. The other reachable positions, labelled as c, can be obtained by two different joint rotation configurations that position the arm at that point.

3.2 Forward Kinematics

In Fig. 5 we consider the first joint as the origin (0, 0) of the coordinate system. The lengths of the two links are l_1 and l_2, respectively. The first joint is rotated by an angle α, and the second one by an angle β (relative to the first joint).

Based on the l_1, l_2, α, β parameters, the position of the end effector results easily as:

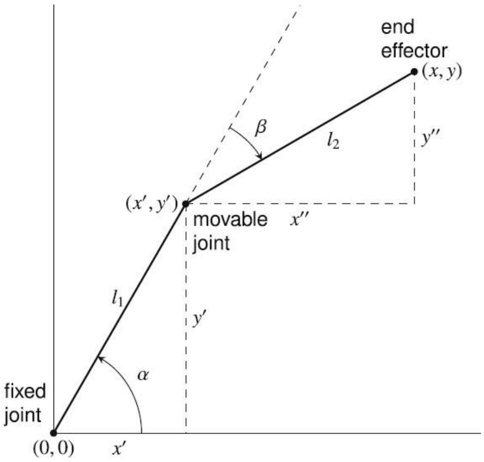
$$\mathrm{x} = \mathrm{x}' + \mathrm{x}'' = l_1 \cos(\alpha) + l_2 \cos(\alpha + \beta) \tag{1}$$

$$\mathrm{y} = \mathrm{y}' + \mathrm{y}'' = l_1 \sin(\alpha) + l_2 \sin(\alpha + \beta) \tag{2}$$

In Fig. 5 the value of β is negative (corresponding to a clockwise rotation).

3.3 Inverse kinematics

In this case we must solve the inverse problem: given the desired coordinates (x, y) and the values for l_1 and l_2 what are the values for α and β to reach the desired position.



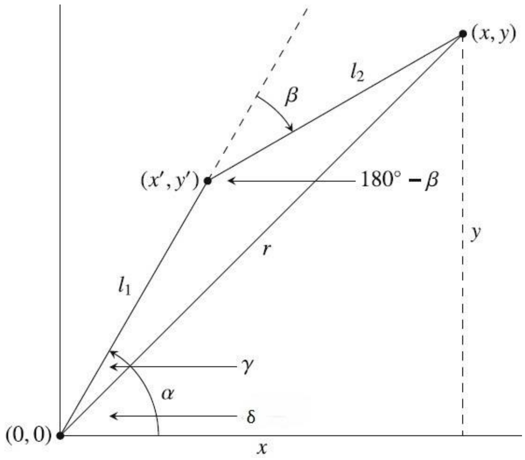


Figure 5. Forward kinematics [9] *Figure 6. Inverse kinematics [9]*

As discussed before, the inverse problem can have no solution (outside the grey area), a single solution (on the borders of the grey area) and two solutions (inside the grey area). When we have two possible solutions, we will select the one so that the arm avoids negative y coordinates as much as possible.

In Fig. 6, by applying the law of cosines in the triangle l_1, l_2, r, we get

$$l_1{}^2 + l_2{}^2 - 2l_1l_2\cos(180° - \beta) = r^2 \tag{1}$$

which can be rearranged to solve for β

$$\cos(180° - \beta) = \frac{l_1{}^2+l_2{}^2-r^2}{2l_1l_2} \tag{2}$$

$$\beta = \pm\left(180° - \text{acos}\left(\frac{l_1{}^2+l_2{}^2-r^2}{2l_1l_2}\right)\right) \tag{3}$$

To obtain γ, and subsequently α, we utilize the law of cosines again, this time considering γ as the central angle

$$\cos\gamma = \frac{l_1{}^2+r^2-l_2{}^2}{2l_1r} \tag{4}$$

$$\gamma = \text{acos}\left(\frac{l_1{}^2+r^2-l_2{}^2}{2l_1r}\right) \tag{5}$$

Also, we have

$$\tan(\delta) = \frac{y}{x} \tag{6}$$

So, we get

$$\delta = \text{atan}\left(\frac{y}{x}\right) \pm \gamma \tag{7}$$

thus

$$\alpha = \text{atan}\left(\frac{y}{x}\right) \pm \text{acos}\left(\frac{l_1{}^2+r^2-l_2{}^2}{2l_1r}\right) \tag{8}$$

To restrict, as much as possible, the arm from moving to negative y coordinates (where a wall could be), we choose a specific solution from the two that are available, according to the quadrants we must reach. For quadrants I and IV (the right ones) we add the γ in (7) and consider negative values for β in (3). For quadrants II and III (the left ones) we subtract γ in (7) and consider positive values for β in (3).

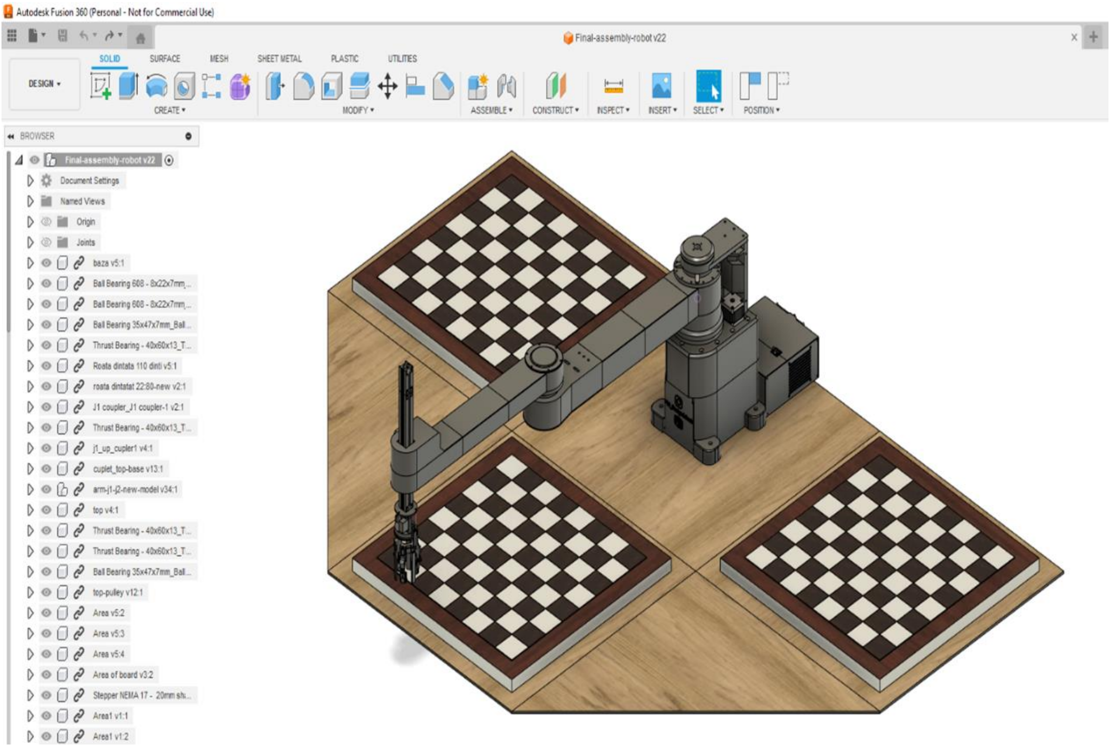

Autodesk Fusion 360 (Personal - Not for Commercial Use)
Final-assembly-robot v22
SOLID
SURFACE
MESH
SHEET METAL
PLASTIC
UTILITIES
DESIGN ▾
CREATE ▾
MODIFY ▾
ASSEMBLE ▾
CONSTRUCT ▾
INSPECT ▾
INSERT ▾
SELECT ▾
POSITION ▾
BROWSER
Final-assembly-robot v22
Document Settings
Named Views
Origin
Joints
baza v5:1
Ball Bearing 608 - 8x22x7mm_...
Ball Bearing 608 - 8x22x7mm_...
Ball Bearing 35x47x7mm_Ball...
Thrust Bearing - 40x60x13_T...
Roata dintata 110 dinti v5:1
roata dintatat 22:80-new v2:1
J1 coupler_J1 coupler-1 v2:1
Thrust Bearing - 40x60x13_T...
j1_up_cupler1 v4:1
cuplet_top-base v13:1
arm-j1-j2-new-model v34:1
top v4:1
Thrust Bearing - 40x60x13_T...
Thrust Bearing - 40x60x13_T...
Ball Bearing 35x47x7mm_Ball...
top-pulley v12:1
Area v5:2
Area v5:3
Area v5:4
Area of board v3:2
Stepper NEMA 17 - 20mm sh...
Area1 v1:1
Area1 v1:2

Figure 7. View of assembled robot in Fusion 360

4 Design of robotic arm

4.1 Mechanical design

The arm was designed in compliance with the board requirements: board size of 520x520mm with field size 55x55mm (as for the autosensory DGT Smart board [11]). The robot was designed to have one board in front, one on left and one on right, thus being able to play chess with three human players simultaneously. The mechanical part of the robot (Fig. 7), a classic SCARA one, was developed using Fusion 360 ([11]) powered by Autodesk.

The robot has a total length of 129 cm and is fully 3D printed. We chose the 3D solution because it is a low-cost and accurate solution. The model is split in more pieces to fit for printing using a small 3D printer (up to 22x22x25 cm). Total amount of time to print the robot was around 300 hours.

The motor responsible for rotating joint 1 is situated within the base. By utilizing two gears a ratio of 1:22 can be achieved. These gears are driven by the motor via toothed belts. To enhance the resistance of the plastic arm, a clamp-type coupling is used at the base, meaning that arm 1 is secured to the base at both top and bottom.

To minimize the weight of the arm, the motor responsible for rotating joint 2 is positioned at the centre of joint 1. Two gears are utilized to achieve a combined ratio

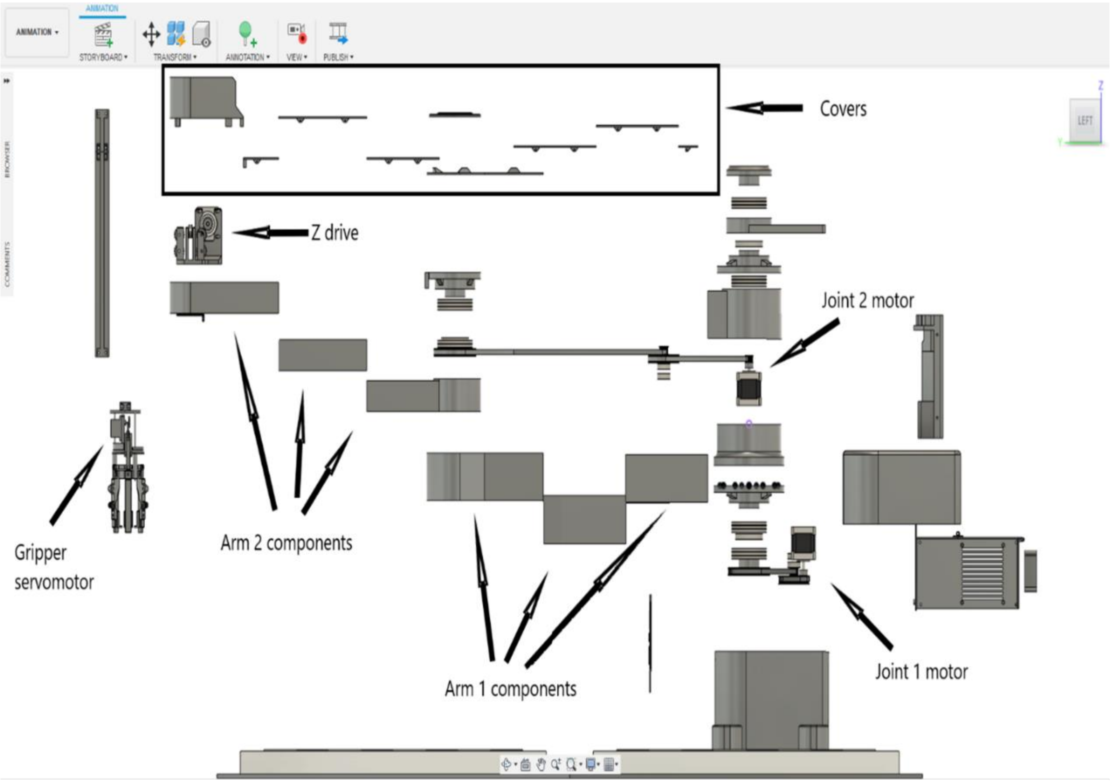


Figure 8. Components of the robotic arm

of 1:14.6. Additionally, two toothed belts are employed to drive these gears and rotate the wheels.

Joint 3 is specifically designed for the linear movement on the Z axis. It features a 1:1 ratio and consists of four gears for driving the axis and bearings for ensuring smooth linear movement. The Z-axis has a total length of 30 cm, but only 20 cm are usable.

At the base of the Z-axis the gripper is positioned, specially designed to securely hold all chess pieces. In the centre of the gripper a camera is strategically placed, enabling future projects involving automatic corrections based on visual location of

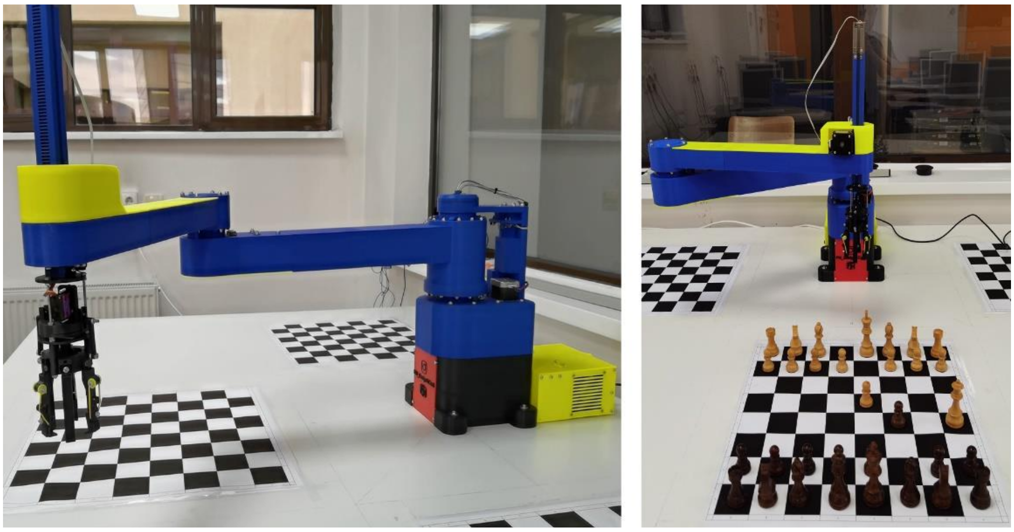


Figure 9. The final stage of the robotic arm

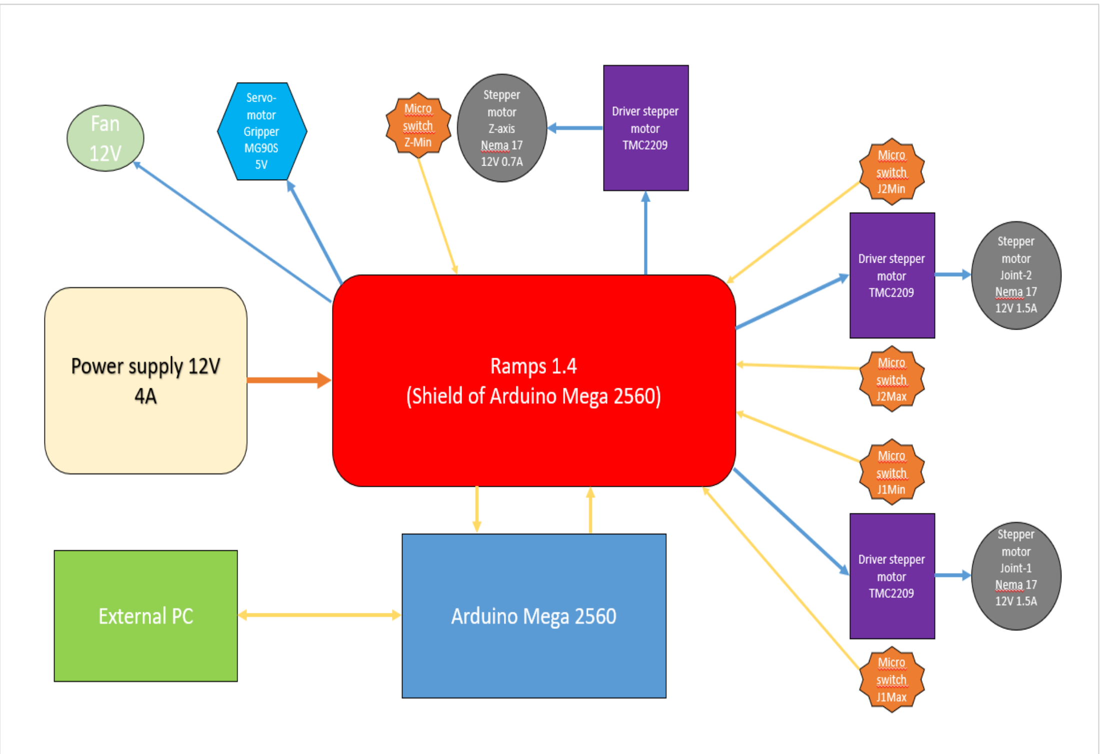

Fan 12V
Servo-motor Gripper MG90S 5V
Micro switch Z-Min
Stepper motor Z-axis Nema 17 12V 0.7A
Driver stepper motor TMC2209
Micro switch J2Min
Driver stepper motor TMC2209
Stepper motor Joint-2 Nema 17 12V 1.5A
Power supply 12V 4A
Ramps 1.4 (Shield of Arduino Mega 2560)
Micro switch J2Max
Micro switch J1Min
Driver stepper motor TMC2209
Stepper motor Joint-1 Nema 17 12V 1.5A
External PC
Arduino Mega 2560
Micro switch J1Max

Figure 10. Block diagram of the hardware part of the robot

pieces on the board. To control the opening and closing of the gripper a servomotor is utilized, providing precise control over its movements.

The robot has a maximum operational radius of 95 cm (l_1=50 cm, l_2=45 cm) and a minimum radius of 20 cm (because of the limiters). Joint 1 is equipped with two position limiters, enabling the arm to rotate up to 272 degrees. Similarly, at the base of joint 2 there are also two position limiters that allow the arm to rotate up to 312 degrees. The robot was specially designed with internal wire integration in mind.

4.2 Hardware design

The board used as the controller of this robot is the Arduino Mega 2560 ([12]). With an impressive array of features, it offers 54 digital input/output pins, among which 15 can function as PWM outputs. Additionally, it boasts 16 analog inputs, 4 UARTs, a 16 MHz crystal oscillator and a USB connection. We use it in combination with a Ramps 1.4 shield ([13]) than offers precise control over up to five stepper motors, allowing for 1/16 stepping precision. Moreover, it seamlessly interfaces with various components, being used by RepRap enthusiasts to build and customize their 3D printing machines with ease. The choice of this board and shield was made considering also future developments of this project.

The block diagram reveals that the robot controller operates under the command of an external computer, establishing a connection with the Arduino board. To operate the robot only three stepper drivers are required, one each for Joint 1, Joint 2, and Joint 3. Additionally, five position limiters were used, the microswitches defining the range of motion. The gripper function is controlled by a servo motor, providing precise control over its movements.

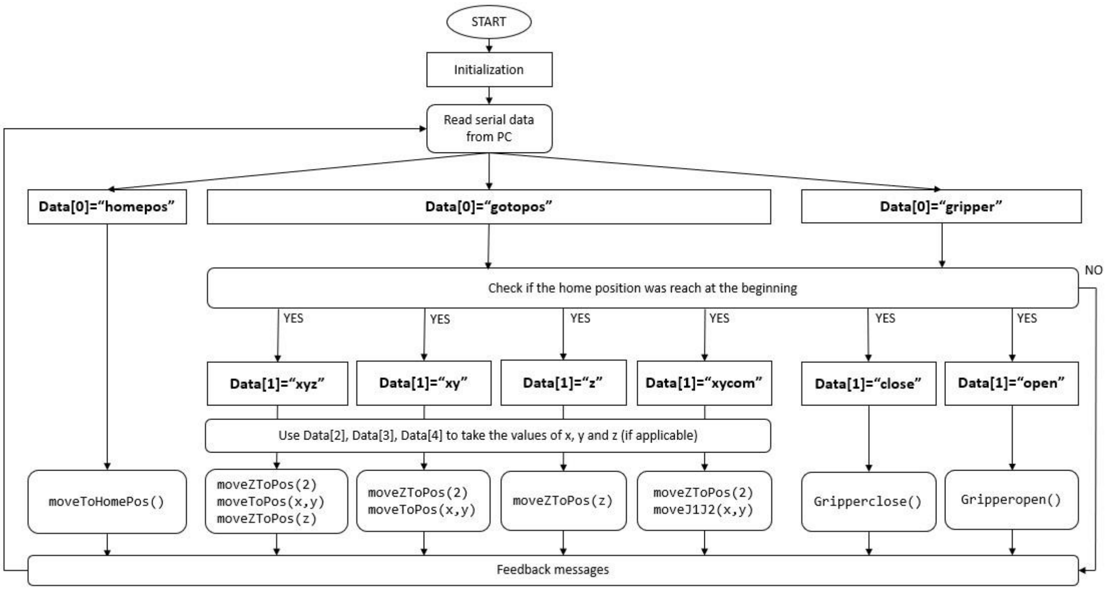


Figure 11. Flowchart of the source code (simplified)

For Joint 1 and Joint 2, Nema 17 motors (12V,1.5A) were selected. These motors belong to the category of stepper motors and provide 200 steps per rotation. To avoid excessive strain on the end effector of the arm, a smaller Nema 17 motor (12V,0.7A) was chosen for joint 3. This motor also offers 200 steps per rotation.

To effectively control these motors TCM2209 bipolar stepper motor drivers were employed. This type of driver is renowned for its silent and precise operation. It enhances the motor's performance by multiplying the number of steps by 8 (microsteps), resulting in an increased precision during movement and positioning.

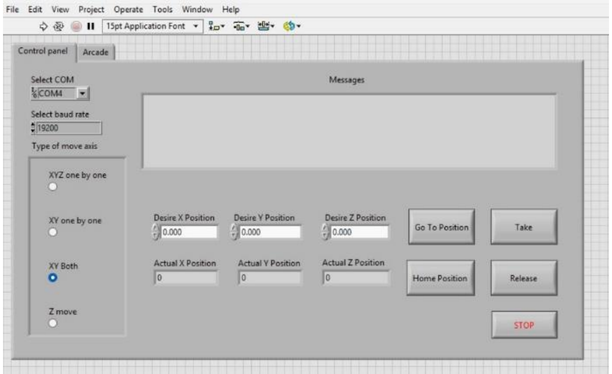
For the gripper mechanism a 5V MG90S servomotor was employed. This specific servomotor offers ease of control and adjustment, allowing for precise gripping of chess pieces based on their respective sizes. The MG90S servomotor can be easily tightened to ensure a secure grip on the chess piece during manipulation. Its reliability and versatility make it an ideal choice for the gripping function of the robot.

4.3 Software implementation

The source code of the robot has been developed in the C programming language using the Arduino IDE. The code incorporates various control functions necessary for the operation of the robot. Additionally, a custom protocol has been created to receive commands through the serial port and process them effectively, simplifying the control of the robot. The combination of the C language code and the custom protocol enhances the versatility and ease of use of the robot, enabling seamless control and interaction with the external computer.

The logic diagram of the robot’s software (Fig. 11) illustrates the sequence of actions within the main loop of the source code. After initialization, the code awaits incoming commands through serial communication.

The software had to solve 3 main problems:



Select COM: COM4

Select baud rate: 19200

Type of move axis

○ XYZ one by one
○ XY one by one
● XY Both
○ Z move

Desire X Position: 0.000
Desire Y Position: 0.000
Desire Z Position: 0.000

Actual X Position: 0
Actual Y Position: 0
Actual Z Position: 0

Figure 12. Control Panel interface

- Design and implement a protocol to communicate via serial line with the PC.
- Implement the actions described in the protocol, based on the available hardware and inverse kinematics (`moveToHomePos()`, `moveToPos(x,y)`, `moveZToPos(z)`, `moveJ1J2(x,y)`, `Gripperclose()`, `Gripperopen()` being the most important functions).
- Design and implement the main loop of operation (described in Fig. 11).

The command `moveZToPos(2)` reflects the need to elevate the gripper to avoid hitting pieces during the movement (the gripper z-coordinate is measured from the home position at the upper limit).

4.3.1 Communication protocol

The communication between the command computer and the controller is established through a serial connection using an USB cable. In the main loop of the source code we have implemented a function that reads the incoming data string received from the computer. This string is structured to transmit five groups of data separated by commas. The format of the message is as follows:

```
data[0], data[1], data[2], data[3], data[4]
```

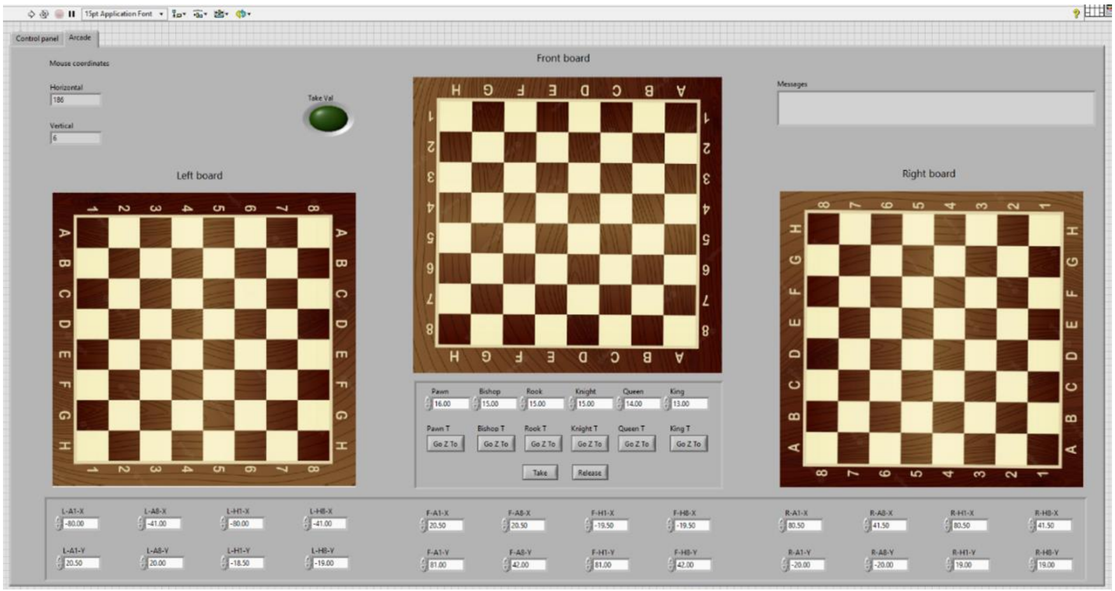
The first data element, `data[0]`, represents the primary command. Depending on the values of this field, the other data fields are interpreted accordingly. The complete protocol is described in Fig. 11. The `data[2]`, `data[3]` and `data[4]` coordinates represents positions with a millimetric accuracy.

We highlight only the difference between `gotopos` command with subcommands `xyz` and `xy` (where the movement is done first for Joint 1 and only after for Joint 2) and the subcommand `xycom` (where the movement of Joint 1 and Joint 2 is done at the same time as much as possible). This was an improvement introduced in the later stages of the project, to speed up the movement.

4.3.2 User interfaces

To interface with the user, we have developed an application using LabVIEW software. The application features two menus: Control Panel and Arcade.

The first one, the Control Panel interface (Fig. 12), is a general one. Once connected to the appropriate COM port, users have the flexibility to execute various commands



Horizontal 186
Vertical 6

Pawn	Bishop	Rook	Knight	Queen	King
16.00	15.00	15.00	15.00	14.00	13.00

L-A1-X	L-A8-X	L-H1-X	L-H8-X	F-A1-X	F-A8-X	F-H1-X	F-H8-X	R-A1-X	R-A8-X	R-H1-X	R-H8-X
-80.00	-41.00	-80.00	-41.00	20.50	20.50	-19.50	-19.50	80.50	41.50	80.50	41.50

L-A1-Y	L-A8-Y	L-H1-Y	L-H8-Y	F-A1-Y	F-A8-Y	F-H1-Y	F-H8-Y	R-A1-Y	R-A8-Y	R-H1-Y	R-H8-Y
20.50	20.00	-18.50	-19.00	81.00	42.00	81.00	42.00	-20.00	-20.00	19.00	19.00

Figure 13. The Arcade Interface (boards arranged from the robot's point of view)

and perform different actions to control the robot's movements and operations. This is a general interface, not dedicated to chess playing.

The second one, the Arcade interface (Fig. 13), is specific to the use of the robotic arm for chess. There are three chessboards available, and users provide commands to the robot, instructing it to move to specific positions on the chessboards (by clicking the desired fields). This feature enables the robot to autonomously move and play chess on the 3 boards simultaneously.

To compensate problems related to the positioning of the boards relative to the robot position, the position of each board is configured by the x,y coordinates of the chess board corner fields (A1, A8, H1, H8).

To adapt to different chess sets, the z coordinate where a specific chess piece must be gripped, we configure that value for each type of piece individually.

5 Conclusions and future work

The design and implementation of the robotic arm for playing chess was not as trivial as it might look at first view. Research, design and redesign were required in all aspects of the robot: mechanical, hardware and software. For example: the motor used in the early stages of the project for the Z-axis needed to be upgraded and, for the gripper, the initial stepper motor needed to be replaced with a servo motor. Redesigning the gripper and the z-axis occurred multiple times, a total of five iterations took place until the final stage was reached. The first driver considered also had to be improved. In the software part the need to move Joint 1 and Joint 2 at the same time emerged and was implemented.

But, in the end, all these problems were overcome, and the robot met its goal: **to reliably move chess pieces on the boards with millimetric positioning precision**.

Certainly, the project can be further developed, mainly in 2 directions:

- Firstly, to use the camera that was installed in the gripper. This will enable real-time visual feedback to the PC, allowing for corrections to be made in case the chess pieces are not precisely positioned on the chessboard by the human player. The camera can contribute also to the safety aspects of the robotic arm, preventing accidents or injuries during operation.
- Secondly, the main development needed by the project is to complete the whole system by interfacing the PC to 3 DGT autosensory boards to read the positions of the pieces and by interfacing the PC with a classic software chess engine, to decide the moves to be made by the chess playing robot when playing against 3 human opponents.

References

[1] *Robots and chess*, Available from: https://en.chessbase.com/post/robots-and-chess

[2] Mark Weeks, *Novag Robot Adversary*, Available from: https://chessforallages.blogspot.com/2015/07/novag-robot-adversary.html

[3] https://www.raspberryturk.com/

[4] A. Gasparetto, L. Scalera *A Brief History of Industrial Robotics in the 20th Century*, Available from: https://air.uniud.it/retrieve/e27ce0c7-1e99-055e-e053-6605fe0a7873/AHS_2019021414562634.pdf

[5] https://www.chess.com/news/view/7-year-olds-finger-broken-by-chess-playing-robot

[6] H. Makino, "Development of the SCARA," J. Robot. Mechatron., Vol.26 No.1, pp. 5-8, 2014, DOI: 10.20965/jrm.2014.p0005

[7] https://www.fanuc.eu/de/en/robots/robot-filter-page/scara-series/selection-support

[8] https://robohub.org/how-many-axes-does-my-robot-need/

[9] Ben-Ari, M., Mondada, F. (2018). *Kinematics of a Robotic Manipulator*. In: *Elements of Robotics*. Springer, https://doi.org/10.1007/978-3-319-62533-1_16

[10] https://www.chessshop.eu/electronic-chess-boards/45-smart-board-usb.html

[11] https://www.autodesk.com/products/fusion-360

[12] *Mega 2560 Rev3*, Available from: https://docs.arduino.cc/hardware/mega-2560

[13] https://reprap.org/wiki/RAMPS_1.4

Arduino Data Acquisition System for Monitoring Quality of Life Parameters in a Room

Alexandru MATEI[1]

[1]Computer Science and Electrical and Electronics Engineering Department, Faculty of Engineering, "Lucian Blaga" University of Sibiu, Romania

Abstract

Data acquisition is the process of collecting information from various sources through sensors or instruments. After obtaining the data, it is usually converted from the analogue format to a digital format.

In this paper an application that gathers room-related information is presented. The purpose of such a system would be to be able to monitor certain parameters of interest in a living space. Afterwards this data can be used in various manners. One of the most important results of this data is being able to identify the habitability of a room. It can point out potential health hazards before they cause actual health problems. An additional objective is for the system to be on an accessible budget and to be easy to understand and implement.

The hardware system using Arduino is presented in detail, as well as the graphical user interface created in Visual Studio.

The system works as intended for the purposes stated above. It measures: temperature, humidity, light intensity, sound, and air quality. It works on its own if it is supplied with a 9-volt power source. The Graphical User Interface can be used to visualise and analyse the collected data. The system can still be improved by adding wireless communication such as Wi-Fi and allowing it to be integrated into Internet of Things systems. The power consumption could be improved if needed for higher sampling frequency demands.

Keywords: Arduino, data acquisition, quality of life, habitability

1 Introduction

1.1 Description of the project

This paper presents a data acquisition system using the Arduino platform. The intent is to create an inexpensive system that is relatively easy to implement and understand. This is where the Arduino system shines, as it provides an abstraction between the hardware, the microcontroller, and the programming environment. This abstraction is, however, entirely optional, the ATmega328 controller of the Arduino being able to be programmed in a more traditional manner as well, if desired. In addition, the supplemental hardware used should be easy to come by or replaced if needed. Another convenient aspect is the community built around this platform. Due to its popularity the Arduino development environments are well documented. This makes it even more accessible than other possible alternatives.

From a hardware point of view, we just have the Arduino Uno Rev3 board, a microSD card module, the card itself, a RTC (Real Time Clock) module, a display made of 7 segment displays, an infrared receiver paired with a remote, and various sensors. The sensors record the following parameters: temperature, humidity, light intensity, sound, air quality, and the position of a potentiometer. Thus, the system is used for monitoring various parameters of a living space.

In addition to the hardware system, a software GUI (Graphical User Interface) was implemented using Microsoft's Visual Studio Integrated Development Environment (IDE). The program was built for a Windows 64-bit operating system. The application shows the acquired data in real time both visually, as a graph, as well as numerically, in a table. The application also allows configuring the system, saving the recorded data in a file, and loading a previously saved file. Furthermore, the application also has a serial console where the user can see the raw message sent by the microcontroller.

The interface allows for the following operations:

- Choosing and enabling the parameters to be measured;
- Setting the sampling frequency;
- Selecting the operating mode of the system: online or offline;

Also, the programs allows the user to save the configuration parameters in the EEPROM memory of the controller. If they are not saved, on the next power on, the microcontroller will load the old data saved in the EEPROM.

Overall, the system has 2 operating modes: paired , connected through an USB-A male to USB B male cable, or offline supplied with 9 volts either from an electric socket or, by using an adapter, from a 9 volt battery.

1.2 Topic relevance

The purpose of this work is to reflect the importance of data acquisition system and open-source hardware projects. By each passing day IoT (Internet of Things) systems play a bigger and bigger role in our society. In addition, the importance of data is increasing by each passing day, making data acquisition systems even more relevant. Platforms such as Arduino that cater to new technologies and concepts such as IoT are proving to be more and more relevant in our technological ecosystem. Arduino are already offering boards with Wi-Fi, Lora, GSM, and other technologies which helps them get chosen for IoT applications.

1.3 Objectives

The principal objective of this work is to create a data acquisition system that can process different parameters given from sensors. The system should be able to keep track of the date and hour of the data entries. In addition, the system should be able to work both paired, connected to a computer running a GUI, as well as offline, completely independent from external systems. In both modes, the system should confer to its users the option to configure the system as well as provide feedback after each provided command. Therefore, there are 2 main objectives: creating the hardware system and creating the GUI for the user. For the hardware part it is important for certain parameters to be chosen that reflect the functionality of the

system. As for the software, it should allow some standard features such as: saving and loading data, visualising data, and being able to configure the hardware system.
note

2 Project overview

2.1 Project presentation

The project was done in 2 big steps: developing the physical acquisition system and realizing the graphical user interface.
For the purposes of creating the hardware system the following components were used:

- The Arduino Uno Rev3 board, based on the ATmega 328 microcontroller from Microchip;
- A USB A cable male to USB B connector;
- The following sensors:
  - DHT22 – sensor module for measuring temperature and humidity;
  - MQ135 – sensor module for measuring air quality;
  - KY-038 – sensor module for measuring sound;
  - A photoresistor for measuring light intensity;
  - A linear rotary potentiometer of 50k Ohms;
- An infrared receiver and an infrared remote;
- A display made from 5 7 segment display elements and a 74HC595 8-bit shift register;
- DS3231 – a Real Time Clock module;
- A microSD module and the card itself (a 32 GB variant from Kingston);
- A 9-volt adapter accompanied by a 9-volt battery.
- A microSD USB card reader used for transferring the files generated by the microcontroller, in the offline operating mode, to the computer;

Following the hardware step, the microcontroller was programmed using the Arduino IDE. The following functionalities were added:

- Executing the commands given either by the interface or the remote control;
- Providing feedback when receiving a command, either through the serial port or through the display, depending on the operating mode;
- Implementing the communication protocol used to talk with the computer interface;
- Creating a data packet that contains the sensor data bundled with the date and hour at which said data was recorded;
- Either sending the data packets generated to the user interface, through the protocol, or saving it locally, on the microSD card;

The graphical user interface was done in Microsoft's Visual Studio IDE and programmed in the C# programming language. It has the following features:

- Communication protocol implementation to communicate with the microcontroller;
- Being able to configure the hardware system;

- Being able to view the received messages through a custom implementation of a serial interface;
- Graphs and tables to visualize the data from the microcontroller, both in real time and from loaded data;
- Saving and loading data in CSV (Comma Separated Values) or XLSX files;

3 Implementation details

3.1 Hardware implementation

In the Fig. 1 the block diagram of the hardware system can be seen. Here it is shown how the Arduino Uno Rev3 board is connected to all the other hardware elements of the system. The 4 analogue sensors (light, air quality, position, and sound) are connected to the analogue inputs of the Arduino. The temperature and humidity sensor are connected to a digital input, as the module is digital. The infrared receiver as well is connected to a digital input. The DS3231 RTC module is connected to the I²C interface. The microSD module is connected to SPI. Of particular importance is the 5-digit 7 segment elements display. Here we used a 74HC595 8-bit shift register to reduce the number of digital inputs needed on the Arduino to only 3 pins. Otherwise, the board would not have enough digital inputs for the display.

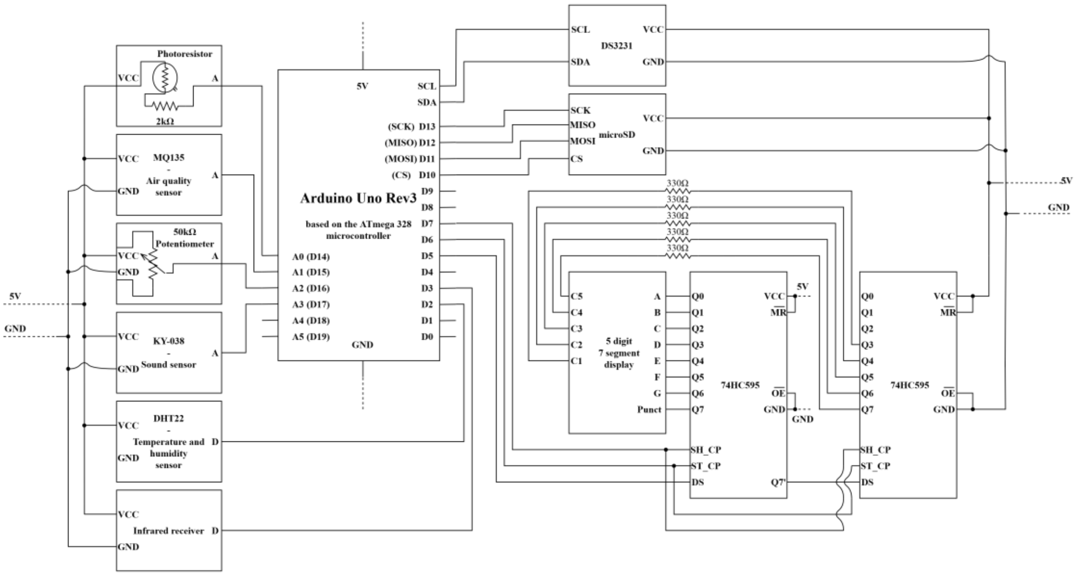


Figure 1. Block diagram of the hardware system

3.2 Communication protocol

3.2.1 System structure

This chapter has the role of providing an overview of how the system works. I chose presenting the system through the communication protocol used because this is what ties the hardware system with the graphical interface. In Fig. 2 and Fig. 3 the overall structure of the system and data flow of can be observed. The system can operate in 2 different modes, online, or paired, and offline.

In the online mode, the system is connected directly to a computer through a USB cable. The Arduino transmits data to the computer through serial, the interface then displays the received data in graphs and tables.

In the offline mode, the hardware system is decoupled from the graphical interface. In this mode it must be supplied with 9 volts. Instead of actively transmitting data, while offline the system saves the data locally on a microSD card in a CSV file. Afterwards this file can be taken and loaded manually in the graphical interface for review.

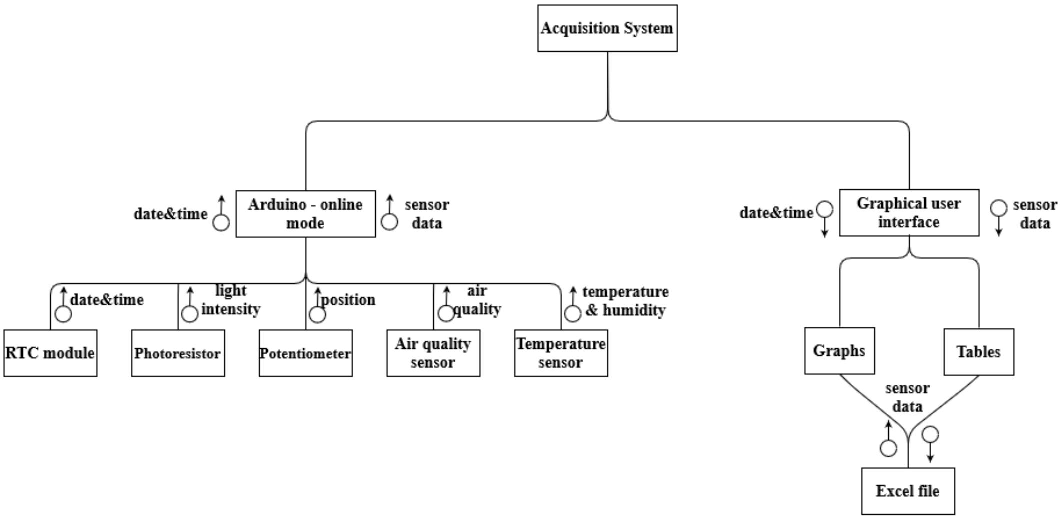


Figure 2. System structure – online mode

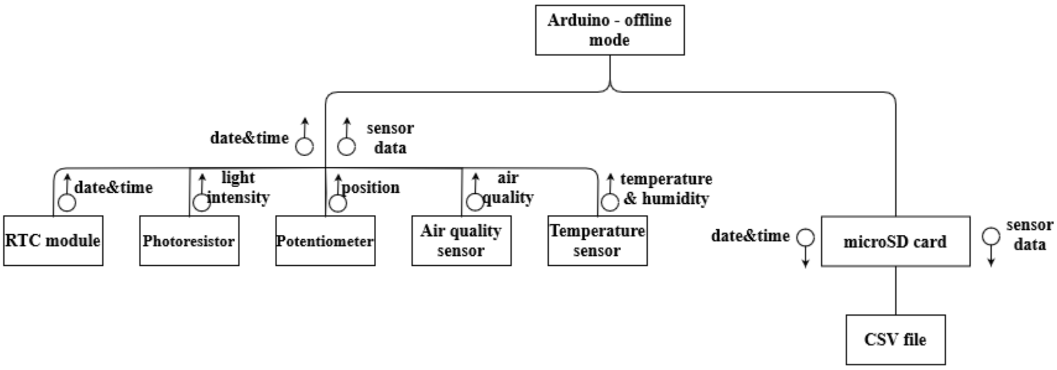


Figure 3. System structure - offline mode

3.2.2 Data transmission

The communication protocol used is built on top of the UART communication protocol. Over UART a secondary protocol was created for communication of the data acquisition system. It is represented through the structure of the data packets sent and the system configuration commands. The packets were designed to be compatible with the CSV file type. Therefore, each comma inside the packet indicates the end of a value and the beginning of a new one. The data packets have the following format:

YYYY-MM-DD,HH-MM-ss-mmm,T----,U----,S---,P---,A---,L---

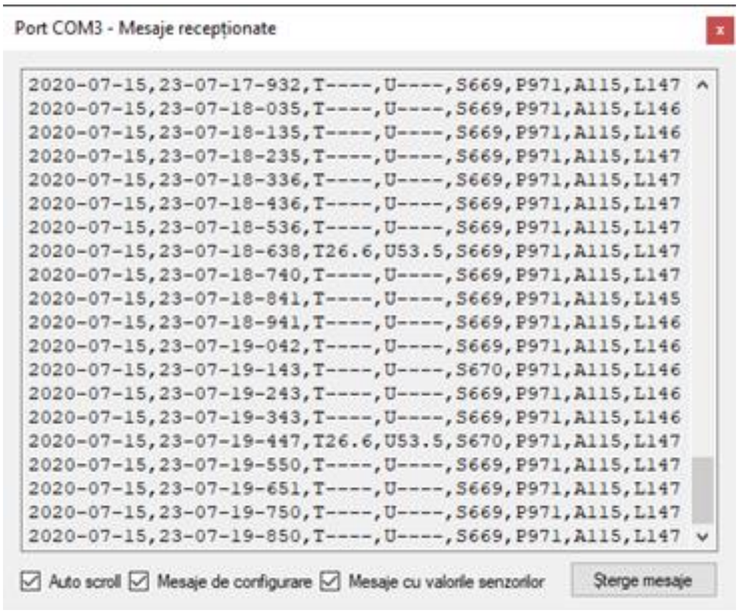
As it can be observed, the packet has 8 different values. The first 2 are the date and time while the following 6 are values obtained from the sensors.

- YYYY-MM-DD represents the year, month, and day when the data packet was generated;
- HH-MM-ss-mmm represents the hour, minute, second and millisecond when the data packet was generated;
- T---- is the value of the temperature, which is represented with 1 decimal point precision;
- U---- is the value of the humidity, represented with the same precision;
- The following values indicate the sensor and the value generated by it in the interval 0 – 999:
    - S---- is the value of the sound;
    - P---- is the value of the position;
    - A---- is the value of the air quality;
    - L---- is the value of the light intensity;

The various sensor used have a different sampling frequency. The data acquisition system works at a sampling frequency of maximum 10 data packets per second. However, not all sensors used can keep up with this pace. According to the data sheet, the temperature and humidity sensor DHT22 has a sampling frequency of 0.5 Hz, which means it measure once every 2 seconds. The system was configured not to collect data from this sensor unless at least 2 seconds have passed from the last obtained value. If the overall system has a sampling frequency greater than 0.5 Hz and the DHT22 sensor has not generated a new value, in the data packet for the parameters of this sensor a null value will be transmitted. In Fig. 4 the serial console of the graphical application can be seen with data packages received in real time.

3.2.3 System configuration – graphical user interface

Depending on the operating mode of the system, it can be configured in one of two methods. In the online, paired, mode, the configuration can be done directly in the interface by using the buttons in the GUI. The application sends serial messages to the controller which then executes the configuration commands it receives. In the offline mode, the system can be configured by using the remote control.


Port COM3 - Mesaje recepționate
x
☑ Auto scroll ☑ Mesaje de configurare ☑ Mesaje cu valorile senzorilor
Șterge mesaje

```
2020-07-15,23-07-17-932,T----,U----,S669,P971,A115,L147
2020-07-15,23-07-18-035,T----,U----,S669,P971,A115,L146
2020-07-15,23-07-18-135,T----,U----,S669,P971,A115,L146
2020-07-15,23-07-18-235,T----,U----,S669,P971,A115,L147
2020-07-15,23-07-18-336,T----,U----,S669,P971,A115,L147
2020-07-15,23-07-18-436,T----,U----,S669,P971,A115,L147
2020-07-15,23-07-18-536,T----,U----,S669,P971,A115,L147
2020-07-15,23-07-18-638,T26.6,U53.5,S669,P971,A115,L147
2020-07-15,23-07-18-740,T----,U----,S669,P971,A115,L147
2020-07-15,23-07-18-841,T----,U----,S669,P971,A115,L145
2020-07-15,23-07-18-941,T----,U----,S669,P971,A115,L146
2020-07-15,23-07-19-042,T----,U----,S669,P971,A115,L146
2020-07-15,23-07-19-143,T----,U----,S670,P971,A115,L146
2020-07-15,23-07-19-243,T----,U----,S669,P971,A115,L146
2020-07-15,23-07-19-343,T----,U----,S669,P971,A115,L146
2020-07-15,23-07-19-447,T26.6,U53.5,S670,P971,A115,L147
2020-07-15,23-07-19-550,T----,U----,S669,P971,A115,L147
2020-07-15,23-07-19-651,T----,U----,S669,P971,A115,L147
2020-07-15,23-07-19-750,T----,U----,S669,P971,A115,L147
2020-07-15,23-07-19-850,T----,U----,S669,P971,A115,L147
```

Figure 4. Serial console – viewing data packages

Here are the configuration commands that the controller has programmed into it:

- Config – asks the microcontroller for the configuration data saved within its EEPROM memory;
- Online – if possible, the controller switches to the online operating mode;
- Offline – the microcontroller switches to the offline operating mode;
- Start – begin data transmission;
- Stop – stop data transmission;
- CanTUSPAL – activating or deactivating certain channels. The letters ‘TUSPAL’ indicate the positions where, writing a 0 or a 1 enables or disables the corresponding sensor. For example, Can111111 is used to enable all the channels;
- A numerical value in the interval [0, 10]
  - This represents the sampling frequency in message per seconds. For example, to transmit a message once every minute we will give the command 0.01666;
- Salveaza – saves the configuration data in the EEPROM memory of the microcontroller;

Initially, when the microcontroller is first powered on, it is in the idle mode where it does not transmit data. Data transmission must be explicitly enabled.

3.2.4 System configuration – remote control

For the remote-control configuration of the system, the display comprised of 7 segment elements is important. It gives feedback to the user that the commands are received. In Fig. 5 are some of the messages the display can show. In order, they represent the following commands: Arduino online mode, Arduino offline mode, data transmission enabled, data transmission disabled, remote selected channel enabled, remote selected channel disabled, configuration saved, pre-set sampling frequency 01 enabled, and finally an error message.

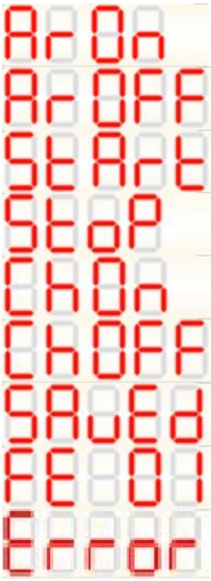


Figure 5. 7 segment display command feedback messages

3.2.5 Communication protocol implementation

In Fig 6. and Fig. 7 the control flow and logic of the software programs can be seen. For the microcontroller software, the diagram is valid for both online and offline modes, the single difference being the different mode of transmitting the data packets. Otherwise, the behaviour is identical.

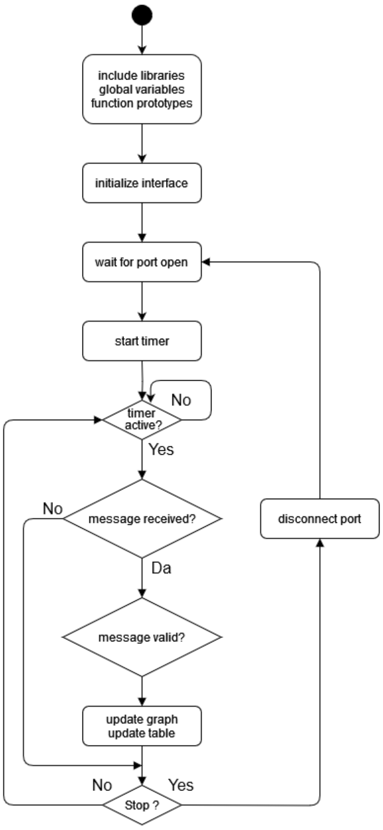


Figure 6. Activity diagram of the interface logic

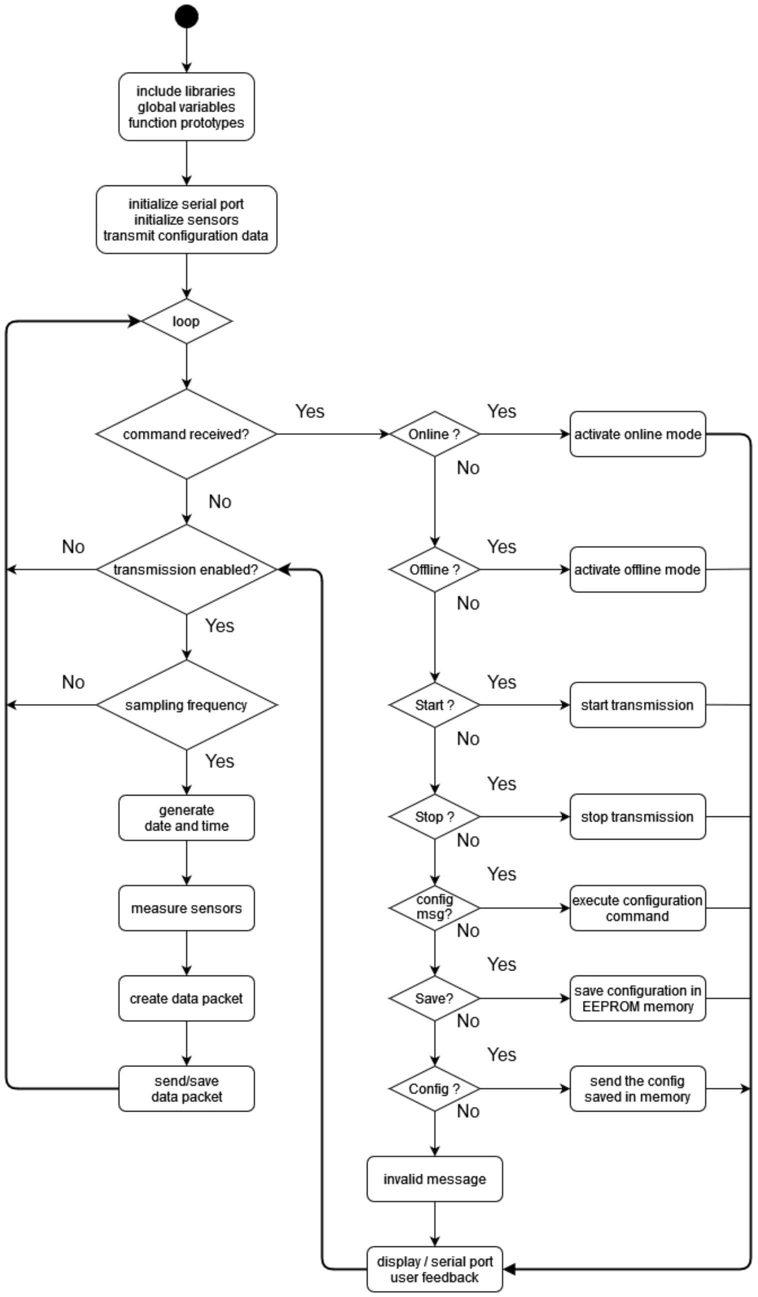


Figure 7. Activity diagram of the microcontroller software

4 Conclusions

The system works as intended and can be a suitable solution for tracking parameters of interest in a living environment. It can operate in two modes: online, or paired, and offline. The system gathers data from sensors measuring various parameters from a room: temperature, humidity, light intensity, sound, and air quality.

One disadvantage with the current implementation is the power usage. Configuring the system with a high sampling frequency consumes a lot of power, in this state the system can’t be used for a long period of time. The system can be improved by optimizing power usage, increasing power supply, as well as adding additional functionality such as Wi-Fi, Bluetooth, or other wireless communication protocol. With internet communication capabilities such a system can be used and integrated with other Internet of Things systems to provide additional benefits.

An advantage is the lower budget, as the resources of the Arduino Uno Rev3 were used to their fullest for this project, without being wasteful. The project could also be changed to use an ASIC (Application Specific Integrated Circuit), using just the ATmega 328, to reduce costs. But for the purposes of making an accessible, easy to use and understand system, that would hinder that goal.

References

[1] DUKA Adrian-Vasile, JOVREA Titus, *Sisteme cu microprocesoare. Microcontrolerul PIC18F4455*, Universitatea "Petru Maior" din Târgu Mureş, Târgu Mureş, 2010.

[2] GERMÁN-SALLÓ Zoltán, *Achiziţia şi prelucrarea datelor: Curs*, Universitatea "Petru Maior" din Târgu Mureş, Târgu Mureş, 2008.

[3] Jeremy Blum, *Exploring Arduino: Tools and Techniques for Engineering Wizardry 2nd Edition*, John Wiley & Sons, Inc., Indianapolis, ISBN 978-1-119-40537-5, 2019.

[4] Steven F. Barrett, Daniel J. Pack, *Microcontrollers Fundamentals for Engineers And Scientists (Synthesis Lectures on Digital Circuits and Systems)*, Morgan & Claypool, California, ISBN 978-1598290585, 2006.

[5] Steven F. Barrett, *Arduino Microcontroller Processing for Everyone!: Third Edition (Synthesis Lectures on Digital Circuits and Systems) 3rd Edition*, Morgan & Claypool, California, ISBN 978-1627052535, 2013.

DOI: 10.2478/ijasitels-2024-0009

Conceptual Modelling of a Laboratory's Products and Some Assisted Management Applications

Dragoș Iliescu [1,2]

[1]*National University of Science and Technology POLITEHNICA Bucharest, Romania*

[2]*SC Nuclear NDT Research and Services SRL, Bucharest, Romania*

Abstract

The research investigates by conceptual modelling tools the requirements needed for a proper definition of a laboratory's products or services. By using a high degree generalization modelling language, under the research scope we investigate the product/service as a material test and/or calibration for measuring equipment. The research results expressed in modelling language are later implemented in an Assisted Management Application.

Keywords: Conceptual Modelling, Laboratory, ISO 17025, Management Integrated System

1 Introduction

A laboratory of interest for the present paperwork is an organization that performs various tests and/or calibrations and extracts a set of interrelated data [1]. Gathered data later transformed into information, the laboratory issues a report stating the test or calibration results, and a declaration of conformity if such a request exists. The laboratory is expected to guard against data corruption by any means [2]. As a general perspective, a laboratory may act under accreditation as a degree of its testing and/or calibration competency and performance, accreditation being issued by the national accreditation body following the standard General requirements for the competence of testing and calibration laboratories - EN ISO/IEC 17025:2017 [3].

From another perspective, the laboratory, like any other organization, pursues financial health, and across different laboratories, different methods may be applied for the purpose. The fields to improve, according to American Productivity & Quality Centre (APQC), are Process Management, Continuous Improvement, Data and Measurement [4]. One may observe that intended improvements are also stated in the standard EN ISO/IEC 17025:2017 as requirements. It is the goal of this research to introduce a starting point for a management system application by modelling the products/services of a material testing and calibration laboratory.

2 Literature Review

2.1 The term of "eLIMS"

The term of "eLIMS" is broadly used over the Internet. Searching the terms of "eLIMS software" or "eLIMS laboratory" will return several results containing some exemplification, mostly having a commercial or marketing aspect. The term of "eLIMS" stands for "Laboratory Information Management System", the e-particle indicating an electronic system as opposite to a classical, e.g. paper-based, system [5,6].

The "eLIMS" is approached by the literature as being a backbone of a laboratory [6] such that, from a management perspective, the importance of such a system is significant at least to say. Such a management system is currently described by literature in connection with other different systems, such as: i) Laboratory notebook, ii) Laboratory execution system, iii) Laboratory scientific data management, among other [1,5,6].

Despite its today importance, at the beginning, the interest in such a software system was low, but with the rapid growth of data volume and complexity, the eLIMS had become an important tool into the laboratory management [5]. Changing the perspective, i.e., the field in which the laboratory is acting, one may observe the implementations and solutions for the cases of medical or bio-medical laboratories [18], but for the cases of industrial material testing and calibration laboratories, the spread of eLIMS solutions are relatively low even nowadays.

2.2 Product versus Service

The general meaning of the term product refers to "something which is produced" with the specialization "something resulting from or necessarily following from a set of conditions" [7]. The general perspective over a product represents a physical object that poses a specific set of characteristics such as one may classify the product as "a product, a service, a process, a person, an organization, a system, a resource" [8,9,10,11]. In a special context, the product may be analysed as an abstract or a social object [10]. In the same direction, the term service is explained as of the product equivalence [9]. However, the term service returns more meanings [13,14,15] and compared with a physical object, the evaluation of its quality and/or effectiveness becomes difficult to evaluate as long the service is intangible [14].

For the cases of a management system complying in a unitary manner with a set of standards (known as a management integrated system), the definition of a product or a service can be even more complex [16,17,18,19].

2.3 Product modelling

The general perspective over a product represents a physical object but a larger perspective, including a service, or an abstract or social object, are possible objectives for modelling. The conceptual modelling represents the most general view [20] of a product for which it can be under study.

A conceptual model has two components according to DiSalvo cited by [12,21,22] - the inner and the outer models.

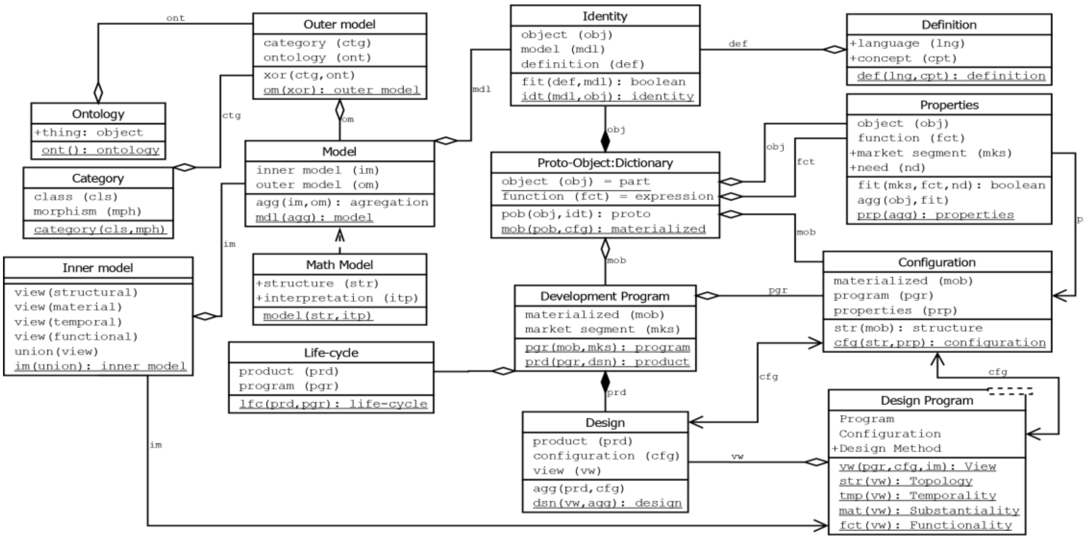


Figure 1. Conceptual model for product [10]

Regarding the outer, one may observe from the literature the category and the ontology theories [12]. For the inner model, there are four views to depict: (i) Topological, (ii) Substantial, (iii) Temporal, and (iv) Functional [12,21,22].

A specific aggregation of the enumerated views will supply in the end an identity to the object of which the model is under consideration [10,20].

A conceptual model can be formulated by a specific language, for exemplification in Fig. 1 the product conceptual model is graphically expressed by the Unified Modelling Language. Also, a formal language represents a possible formulation as in (1) [22].

$$P = \{S(O_k), O_0, \{\delta_k\}\}, \delta_k \in (h_s \circ h_t \circ h_m \circ h_f), 1 \leq k \leq (|c| - 1) \qquad (1)$$

With P as a generic product (1) and Ƥ as a category ($P \in$ Ƥ), $S(O_k)$ is the structure of an object $O_k \in P$, O_0 is the root object (in particular, the null product [20]), $|c|$ is the cardinal of P, and h_s, h_t, h_m, h_f are the functions defining the topological, temporal, substantial and functional views of δ_k.

With δ_k as a transitive transformation, then $U_k(P)$ is the orbit starting from P, and $O_k = O_0 + \delta_k$ [20].

The aggregation of the object, model, and definition results in the object's identity, and the proto-object implements it [12,23,24]. The proto-object is essentially a mental object and it can be further deployed by the concepts of potential components and null product [20]. The materialization of a proto-object is known as a prototype.

2.4 Product’s orbit

Product modelling includes concepts of transformation, hypostases, and orbit [20]. The product model starts from a root object, named by different authors as "thing", "object", "artefact", "endurant", etc. The root object poses the most general and abstract meaning [10]. The root object lacks the identity, components and qualities, only the function may be, at least partial, known.

In Fig. 1 the root object is illustrated as “part”, defined by a dictionary (i.e., a restricted language).

To the root object, one may associate an identity by considering of a definition and a model combining outer and inner views. The object here is known also as a null product [20] and in association with a given identity the proto-object results (Fig. 1). It is to be noted that identity function may associate a function of fit (Fig. 1) which represents the evaluation of a knowledge space [21].

The proto-object represents the foundation of some other product's model components such as development program, design, and configuration (Fig. 1), each of these having a precise contribution in creating the complete product model.

The proto-object's components evolve by various transformations resulting in various hypostases of the product and later in a product orbit [20]. A product orbit represents a complete view, all perspectives included, of a product [25].

3 Research Objectives and Methods

Under the research scope, we consider the laboratory of interest as an information-based organization. With the assumed laboratory typology, the laboratory's management system becomes the subject for research from the effectiveness perspective.

The research focuses on the laboratories acting on the field of material or product testing by destructive and/or non-destructive methods. A typical example represents a set of tests performed on metallic materials of any kind, e.g. – steels, iron, copper-alloys, aluminium-alloys, etc. The ultrasonic examination of tubes (i.e., tubular products), magnetic particle testing for forged, cast or rolled products, etc. may be considered as a typical example.

The research focuses also on the laboratories acting in the calibration of industrial measuring equipment, e.g. – gages, micrometres, temperature measurement systems, hardness testers, etc.

It is the goal of the present research to propose and to build a computer-based assisted management application with the goal of management system's effectiveness improvement in direct response to the ISO/IEC 17025:2017 requirements.

The research preliminaries count on the elements of conceptual modelling introduced in section 2 and intend as outputs to define the general framework for the laboratory's product definition, such that to model not only the product, but the product's orbit.

The research focuses on product definition and on the method to implement such a product definition into a laboratory's management integrated system. Under the scope of the present study, the term of product modelling refers to all the meanings the term of product poses (§2.3, §2.4).

For a finite set of abstract objects standing as laboratory's products, $P = \{p_i\}$, $i = 0 \dots n$, $n = |P|$, (2)-(4) define the characteristics of reflexivity, anti-symmetry, and transitivity, where · represents a relation of any kind between elements of P.

$$p_m = p_m \tag{2}$$

$$p_m \cdot p_r = p_r \cdot p_m \rightarrow p_m = p_r \tag{3}$$

$$p_m \cdot p_r \wedge p_r \cdot p_s \rightarrow p_m \cdot p_s \tag{4}$$

4 General Frameworks

With p as an abstract object ($p \in P$), considering (1) and the model depicted in Fig. 1 we will define the model of a laboratory's service (s) as in (5).

$$s \in p \tag{5}$$

We define s as a function of category (C) and the views of topology, temporality, substantiality and functionality, i.e., δ_k as defined in (1). Then, C and δ_k classify s (6).

$$s = C \times \delta_k \tag{6}$$

Following the model in Fig. 1, Table I (see §4.2) summarizes the parameters of service identity introduced by (6).

4.1 Outer identity implementation

The term of category introduces the theory of classes (objects and morphisms) into the laboratory's service model. A specific classification, i.e., a category-based classification, poses a high degree of generality. The current research focuses on identity morphisms, which by definition, is unique for every given object in the category, i.e., to every class. For exemplification, in Table 1 (see §4.2), one will expect that the designation of "Tensile tests" indicates a unique class of testing. A category has classes as elements (in this case the term of large categories is applied) or sets (small categories).
This specialization introduced a tree-like structure in which the research focuses on verifying the relations in (2) – (4).

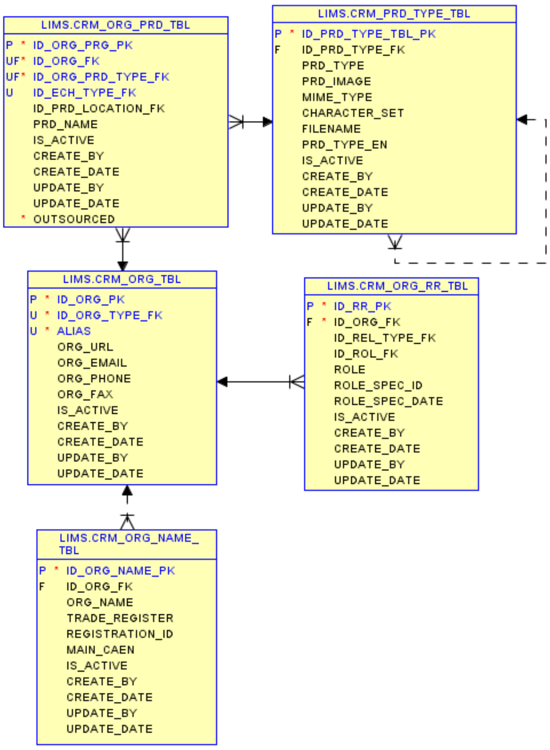
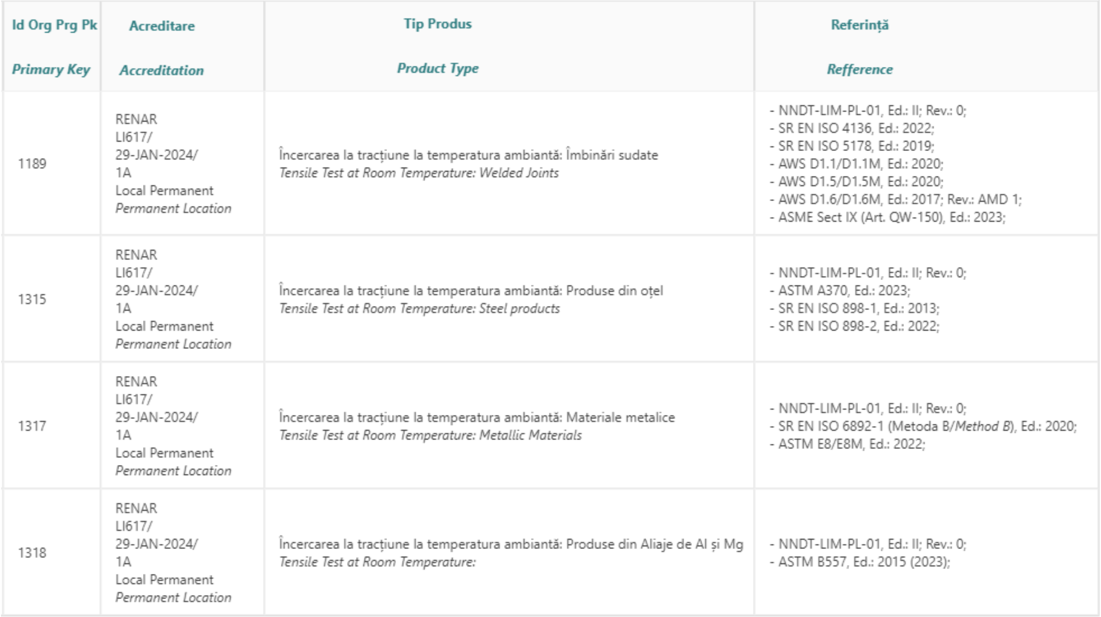


Figure 2. Database implementation of outer model (screen capture Oracle SQL Developer)



Id Org Prg Pk *Primary Key*	Acreditare *Accreditation*	Tip Produs *Product Type*	Referință *Refference*
1189	RENAR LI617/ 29-JAN-2024/ 1A Local Permanent *Permanent Location*	Încercarea la tracțiune la temperatura ambiantă: Îmbinări sudate *Tensile Test at Room Temperature: Welded Joints*	- NNDT-LIM-PL-01, Ed.: II; Rev.: 0; - SR EN ISO 4136, Ed.: 2022; - SR EN ISO 5178, Ed.: 2019; - AWS D1.1/D1.1M, Ed.: 2020; - AWS D1.5/D1.5M, Ed.: 2020; - AWS D1.6/D1.6M, Ed.: 2017; Rev.: AMD 1; - ASME Sect IX (Art. QW-150), Ed.: 2023;
1315	RENAR LI617/ 29-JAN-2024/ 1A Local Permanent *Permanent Location*	Încercarea la tracțiune la temperatura ambiantă: Produse din oțel *Tensile Test at Room Temperature: Steel products*	- NNDT-LIM-PL-01, Ed.: II; Rev.: 0; - ASTM A370, Ed.: 2023; - SR EN ISO 898-1, Ed.: 2013; - SR EN ISO 898-2, Ed.: 2022;
1317	RENAR LI617/ 29-JAN-2024/ 1A Local Permanent *Permanent Location*	Încercarea la tracțiune la temperatura ambiantă: Materiale metalice *Tensile Test at Room Temperature: Metallic Materials*	- NNDT-LIM-PL-01, Ed.: II; Rev.: 0; - SR EN ISO 6892-1 (Metoda B/*Method B*), Ed.: 2020; - ASTM E8/E8M, Ed.: 2022;
1318	RENAR LI617/ 29-JAN-2024/ 1A Local Permanent *Permanent Location*	Încercarea la tracțiune la temperatura ambiantă: Produse din Aliaje de Al și Mg *Tensile Test at Room Temperature:*	- NNDT-LIM-PL-01, Ed.: II; Rev.: 0; - ASTM B557, Ed.: 2015 (2023);

Figure 3. Application eLIMS (screen capture eLIMS application)

Continuing the exemplification, we can define the category of "Tensile tests" as being a small category (made by the sets of tensile tests at room temperature, and the set of tensile tests at elevated temperature), being part of a large category named, e.g., "Mechanical tests" (§4.2, Table 1). With a given large category, there should be a minimum of one large and one small, or at least two small categories. Along with the category, the location and the supplier (Table 1) will implement the outer model of a laboratory's service identity, with a possible implementation as illustrated in Fig. 2 and 3.

4.2 Inner identity implementation

Four views are under consideration (Table 1) for the inner identity. The structural view implements the identification of the test or calibration method, the accuracy and/or uncertainty, and the associated performance indicators.

The method and the work procedure are equipment-dependent. The collection data type means that the structural view may be optimally implemented by a relational database. Here the term of collection poses the characteristics of a configuration, with the specialization of configuration management (ISO 10007:2017 supplies more information on the topic). As a consequence, the laboratory's service structural identity is under construction by a unique collection (configuration) as illustrated in Fig. 4 and 5.

The configuration of a laboratory's service is under control with the help of the table ERP_PL_VER_TBL (Fig. 4 and 6). It controls the exact version of each standard or working procedure declared by the table of ERP_PRG_PL_TBL by edition, revision, and transaction type. Every document declared in association with a given service may have different versions, but only one may be an active document. This constraint is implemented with the help of a trigger attached to the table ERP_PL_VER_TBL.

A declared configuration is then kept under control by the tables CRM_PRD_ACRD_TBL and CRM_PRD_ACRD_POZ_TBL (Fig. 4).

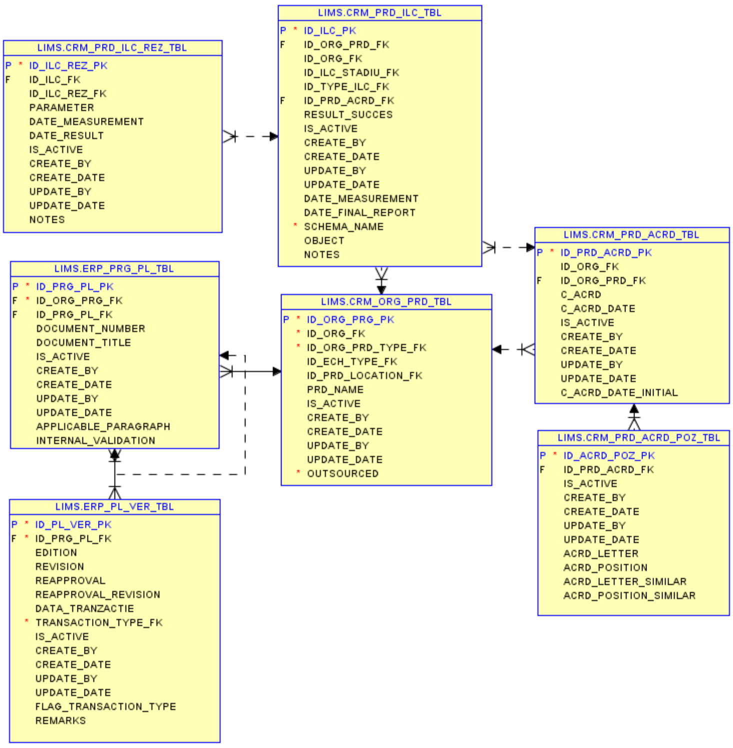
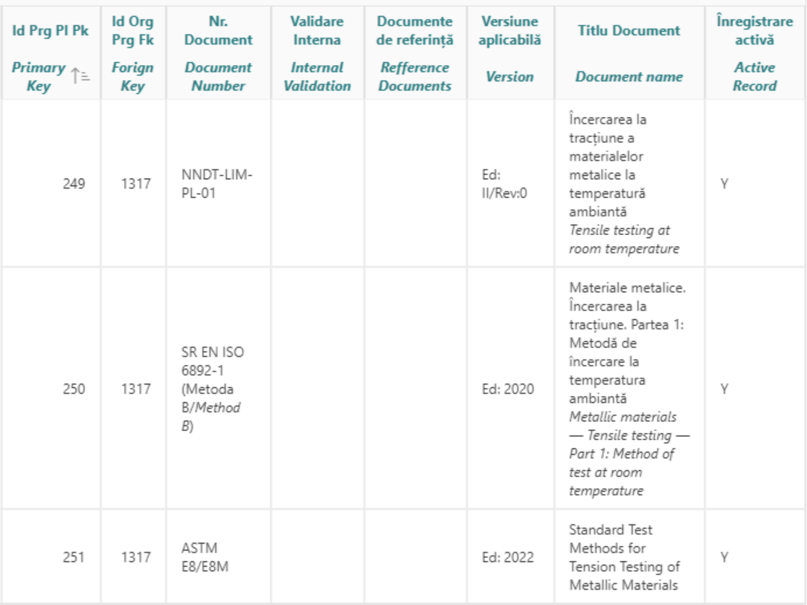


Figure 4. Database implementation of structural view
(screen capture Oracle SQL Developer v. 3.2, simplified view)

The material view classifies the measured values in primary and secondary data depending on whether the gathered data comes direct from the measurement process, or it is the result of a transformation. The material view operates with both primary and secondary data. It implements and controls the "Technical records" as per ISO/IEC 17025:2017. The material view also supplies the control of data transformation to information, i.e. supplying the unit of measure for both standard and reading values.



Id Prg Pl Pk *Primary Key* ↑≡	Id Org Prg Fk *Forign Key*	Nr. Document *Document Number*	Validare Interna *Internal Validation*	Documente de referință *Refference Documents*	Versiune aplicabilă *Version*	Titlu Document *Document name*	Înregistrare activă *Active Record*
249	1317	NNDT-LIM-PL-01			Ed: II/Rev:0	Încercarea la tracțiune a materialelor metalice la temperatură ambiantă *Tensile testing at room temperature*	Y
250	1317	SR EN ISO 6892-1 (Metoda B/*Method B*)			Ed: 2020	Materiale metalice. Încercarea la tracțiune. Partea 1: Metodă de încercare la temperatura ambiantă *Metallic materials — Tensile testing — Part 1: Method of test at room temperature*	Y
251	1317	ASTM E8/E8M			Ed: 2022	Standard Test Methods for Tension Testing of Metallic Materials	Y

Figure 5. Application NNDT-eLIMS
(screen capture NNDT-eLIMS application – structural view, reference documents)

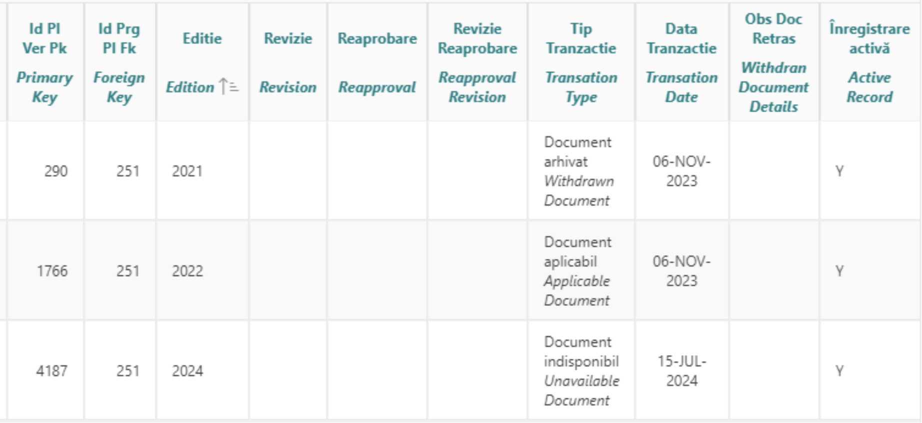
Table 1. The laboratory’s service identity

Model	**Name**	**Type**	**Quale**
Outer	Category	Natural/ Restricted language	Test designation (e.g., “Tensile tests”)
			Calibration designation (e.g., "Calliper calibration")
	Location	Natural language	Laboratory premises
			Client premises
			Mobile premises
	Supplier	Natural/ Restricted language	Laboratory identification/name/address
Inner	Structure	Collection (Configuration)	Standards for tests or calibration methods (e.g., “ISO 13385-1:2019”
			Work procedures (e.g., “WP-04 Ed. 2019, Rev 4”)
			Equipment/Domain/Range/ Accuracy/Uncertainty (e.g., “Calliper/0-300 mm/0.5%”)
	Material	Information	primary data (e.g., “standard value”,” reading value”)
			secondary data (e.g., “transformed standard value”/”reading value”)
	Temporal	Graph/ Partition	Version number (e.g. “Ed.2020, Rev. a”)
	Function	k-Matrix[c]	$\{\{Ch^{a}_{j}\ x\ r^{b}_{i}\}_{j}\}_{k}$ (e.g., “Testing/Calibration Report”)
		Communication	Communication channel (e.g., e-mail, post, etc.)

[a]. Ch_j stands for the j^{th} characteristic (quality)

[b]. r_i stands for the i^{th} result (quale)

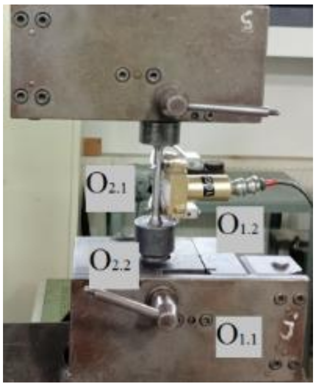
[c]. a k-dimension matrix



Id Pl Ver Pk *Primary Key*	Id Prg Pl Fk *Foreign Key*	Editie *Edition* ↑≡	Revizie *Revision*	Reaprobare *Reapproval*	Revizie Reaprobare *Reapproval Revision*	Tip Tranzactie *Transation Type*	Data Tranzactie *Transation Date*	Obs Doc Retras *Withdran Document Details*	Înregistrare activă *Active Record*
290	251	2021				Document arhivat *Withdrawn Document*	06-NOV-2023		Y
1766	251	2022				Document aplicabil *Applicable Document*	06-NOV-2023		Y
4187	251	2024				Document indisponibil *Unavailable Document*	15-JUL-2024		Y

Figure 6. Application NNDT-eLIMS
(screen capture control of applicable Standard, ASTM E8/E8M)

As already introduced, the referenced method and the corresponding work procedure are equipment-dependent. With the tensile test exemplification, in Fig. 7 and 8 there are a possible implementation of the relation of dependability and Table 2 presents the specificity of this relation.



O_1 – Tensile - Compression testing machine;

$O_{1.1}$ – fixing vice;

$O_{1.2}$ – extensometer;

O_2 – tensile sample assembly;

$O_{2.1}$ – tensile sample;

$O_{2.2}$ – fixing adapter;

Figure 7. Tensile test machine and tensile sample fixing details for tensile test

Among the objects which are part in the process of tensile test at room temperature, a series of specific relations occur (Fig. 8) as detailed in Table 2.

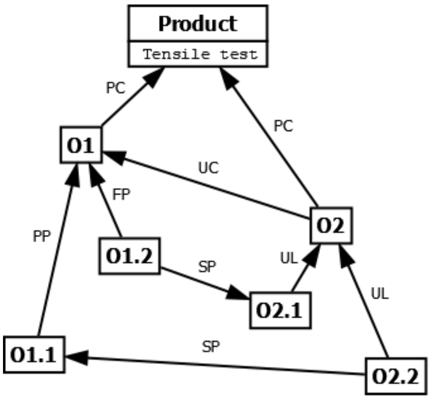


Figure 8. Topology analysis for tensile test at room temperature (simplified)

Table 2. Whole-part relations

Relation	Definition
PC (Participation)	One object participates in a process or in an event
PP (Proper part)	An object is part of a higher degree object as in assembly-subassembly
FP (Functional Part)	One object’s function participates to a process/event or can be associated to the function of another object
SP (Spatial Part)	One object is located in another object but is not part nor proper part of it
UL (Underlap)	Two objects are parts, or proper-parts, of another object
UC (Under-cross)	Two UL objects that are not part nor proper-parts one another

5 A case study

The concepts expressed in this paper were tested by the creation of an information application called Nuclear NDT eLIMS and NNDT-eLIMS stands for its alias.
The goal of the NNDT-eLIMS application is to build an electronic computer-based information management system. In the pursuit of the goal, the research uses the following electronic infrastructure:

- Oracle XE database, version 11.2;
- Oracle SQL Developer, version 3.2;
- Oracle APEX, version 20.2;
- Oracle REST Data Services, version 20.4.

5.1 The case study construction

5.1.1 NNDT-eLIMS and management system effectiveness

With R = {r_i}, i = 1 … n, as the set of ISO/IEC 17025:2017 requirements, and D = {d_j}, j=1 … m, as the class of documents/records in response to R, the effectiveness of the management system, E = {0,1}, is defined as in (7):

$$E \leftrightarrow \forall r \in R, \exists \Gamma \to R \xrightarrow{\Gamma} D \qquad (7)$$

In (7), with Γ, we denote a morphism dependent to the actual management system implementation, i.e., it may vary from one organisation to another, but as Γ is a morphism, the R and D structures are preserved. It is the goal of NNDT-eLIMS application to materialise Γ as for the case of SC Nuclear NDT Research and Services SRL, and to maximise the effectiveness of the management system, i.e., E = 1.
For the case study of NNDT-eLIMS, we consider the product accredited by the National Accreditation Body - RENAR named as “Tensile test at room temperature” (original, “Încercarea la tracțiune la temperature ambiantă”) as defined by [26], position 1A.

Table 3. Case study product definition

Parameter/definition	**Setting for case study**
Market	Destructive tests
Market segment	Mechanical tests
Specific need	Measure or determine the Ultimate Tensile Strength, Yield Strength, Reduction of Area, Elongation, and Modulus of elasticity for Metallic materials/ Products made in steel
Product function	Determine the metallic materials' behaviour under applied tensile force and a predetermined temperature
Object	Tensile test standard sample
Concept	ISO 6892-1, ASTM E8 (metallic materials) versus
	ISO 898-1, ISO 898-2, ASTM A370 (products made in steel)
Language	ISO 6892-1, ASTM E8 (metallic materials) versus
	ISO 898-1, ISO 898-2, ASTM A370 (products made in steel)
Topology	Fig. 7 and 8
Temporality	$t_{\|M(o)\|}$= 2024-05
	Test rate → strain rate versus stress rate versus rate of separation
Substantiality	Type and dimensions of product versus type and dimensions of tensile sample
	Type of response → ISO 6892-1 Fig. 2 versus Fig. 3
	R_m, $R_p/R_{eH}/R_{eL}$ → [MPa]
	Elongation, Reduction of Area → [%]
	Modulus of elasticity (Young's) → [GPa]
Functionality	Out of scope

Expanding the definition of the selected product within the case study had resulted in the matrix illustrated in Table 3.

The matrix in Table 3 should be under consideration for every object in the product of interest, e.g., the objects illustrated in Fig. 7.

Finally, with all details settled, the product definition result as in Fig. 3, where the unique identity is represented by the column "ID_ORG_PRG_PK", i.e., the primary key. The primary keys were generated and queried later without errors as an evidence of success in the implementation of the concepts depicted within this research.

6 Conclusion

Within the present paperwork, we study the conceptual modelling technique as applied to the organizations, the scope of the study being with the material testing and calibration laboratory.

We investigate the conceptual modelling application starting with a general product model, with the help of different views introduced by the study – inner and outer views. By using the suggested procedure, we establish an information structure to determine the identity of a given service (tensile test at room temperature was considered for exemplification) provided by our laboratory of interest. Keeping the traceability of records (the orbit concept) represents an objective of high interest as this is a requirement of the reference standard.

The proposed concepts, as defined in the general framework, and implemented into the NNDT-eLIMS application by the case study, validate the research's goal and results, in term of the management system effectiveness. A mathematical equation expressing the effectiveness of the information management system is proposed for this specific subject. The NNDT-eLIMS application implements and tests the concept of effectiveness expressed by (7).

The effectiveness was controlled by an information structure (the Γ set of relations) which was transformed into a database scheme to test the validity of the introduced concepts. The product identity in this case was substituted by a primary key ID as quale for $D = \{d_j\}$. The unity of $\{d_j\}$, i.e., the definition of a primary key, is the method to get the effectiveness of the management system.

Based on the database schema, an application (NNDT-eLIMS) was built to manage the information within the work scope. The NNDT-eLIMS is currently used within the SC Nuclear NDT Research and Services SRL management system to demonstrate that the necessary conditions for accreditation are preserved. The NNDT-eLIMS application is employed when RENAR evaluate SC Nuclear NDT Research and Services SRL for accreditation under the scope of ISO/IEC 17025:2017 requirements.

The utility of the present research goes in a better understanding of the work specificity of a material test and calibration laboratory as an information-based organization. The utility of the research goes further as the proposed framework can be extended to accommodate an electronic management of the information with specificity for the laboratory of interest, and even further, it permit extension of data and information management with new modules to connect the laboratory with its clients.

Acknowledgment

The present research is currently supported by SC Nuclear NDT Research and Services.

References

[1] Gerlach B., Untucht C., Stefan A.: Electronic Lab Notebooks and Experimental Design Assistants. In Good Research Practice in Non-Clinical Pharmacology and Biomedicine: Bespalov A., Martin C.M., and Steckler T. (eds.): Springer International Publishing, vol. 257, pp. 257-275, 2019.

[2] EN ISO/IEC 17025:2017 General requirements for the competence of testing and calibration laboratories. ISO, Geneve, 2017.

[3] RENAR, https://www.renar.ro/index.php/acreditarea, last accessed 2024/03/20.

[4] APQC: 2024 Process & Performance Management. APQC, Houston, 2024.

[5] Sun D., Wu L., Guomei Fan. Laboratory information management system for biosafety laboratory, Journal of Biosafety and Biosecurity, no. 3, pp. 28 – 34, 2021.

[6] Schmieder F. et al. Universal LIMS based platform for the automated processing of cell-based assays, Current Directions in Biomedical Engineering, no. 5(1), pp. 437 – 440, 2019.

[7] Merriam-Webster Dictionary, https://www.merriam-webster.com/dictionary/product, last accessed 2024/03/20.

[8] Tsimillis K.C.: Eurachem Training Programme - Data and information management, Pancyprian Union of Chemists, Nicosia, 2019.

[9] ISO 9000:2015 Quality management systems -- Fundamentals and vocabulary. ISO, Geneve, 2015.

[10] Borgo S., Vieu L.: From Physical Artefacts to Products. http://www.loa.istc.cnr.it/old/Papers/ProductFomi06-CR.pdf, last accessed 2006/11/15.

[11]Vernadat F.: Enterprise Modelling: Research review and outlook. Computers in Industry, vol. 122, p. 103265, 2020.

[12]Iliescu D.: Some Considerations on Universal Product Model (2) - Toward the Universal Product Model by UML. https:/dx.doi.org/10.13140/RG.2.2.30297.88162, last accessed 2024.02.18.

[13]Merriam-Webster Dictionary, https://www.merriam-webster.com/dictionary/service, last accessed 2023/03/10.

[14]Liu X.X., Chen Z.Y.: Service quality evaluation and service improvement using online reviews: A framework combining deep learning with a hierarchical service quality model. Electronic Commerce Research and Applications, vol. 54, p. 101174, 2022.

[15]Jaekel F.W., Zelm M., Chen D.: Service Modelling Language Applied for Hyper Connected Ecosystem. In Proceedings of the 2nd International Conference on Innovative Intelligent Industrial Production and Logistics 2021, pp. 209-215, 2021.

[16]Barbosa A.S., da Silva L.B., Morioka S. N., da Silva J.M.N, de Souza V.F.: Item response theory-based validation of an integrated management system. Journal of Cleaner Production, vol. 328, 2021.

[17]Biggeri M., Borsacchi L., Braito L., Ferrannini A.: Measuring the compliance of management system in manufacturing SMEs. Journal of Cleaner Production, vol. 382, 2023.

[18]Khalifa A. et al.: A qualitative study of prevalent laboratory information. Genetics in Medicine, vol. 23, pp. 2171-2177, 2021.

[19]Tomlinson E. et. al.: A Model for Design and Implementation of a Laboratory Information-Management System Specific for Molecular Pathology Laboratory Operations. The Journal of Molecular Diagnostics, vol. 24, no. 5, pp. 503-514, 2022.

[20]Gheorghe M. and Iliescu D.: Mathematical Modelling of Product Representations. U.P.B. Sci. Bull., Series A, vol. 81, no. 4, pp. 189-200, 2019.

[21]Iliescu D.: Toward the Universal Product Model by Knowledge Space. http:/dx.doi.org/10.13140/RG.2.2.33101.79840, last accessed 2019/10/10.

[22]Iliescu D.: Some Considerations on Universal Product Model. https://dx.doi.org/10.13140/RG.2.2.10165.22243, last accessed 2019/10/10.

[23]Dirk W., Christof K.: Modelling attention to salient proto-objects. Neural Networks, vol. 19, p. 1395–1407, 2006.

[24]Rensink A. R.: Seeing, sensing, and scrutinizing. Vision Research, vol. 40, p. 1469–1487, 2000.

[25]ISO 19440:2020 Enterprise modelling and architecture — Constructs for enterprise modelling. ISO, Geneve, 2020.

[26]Certificat de Acreditare LI 617/29.01.2024, RENAR, 2024. https://www.renar.ro/index.php/oec/get_oec_details/42927, last accessed 2024/05/10.

DOI: 10.2478/ijasitels-2024-0011

Categorical Mechanisms in Modelling Multiagent Systems

Daniel C. Crăciunean[1]

[1]*Computer Science and Electrical and Electronics Engineering Department,*
Faculty of Engineering, "Lucian Blaga" University of Sibiu, Romania
daniel.craciunean@ulbsibiu.ro

Abstract

The main contribution that this paper brings is the specification of multiagent systems, at the metamodel level, using an appropriate categorical sketch. We will see that category theory provides all the necessary ingredients for the formal specification of multiagent systems. In our approach, a multiagent system is specified by a static dimension and a behavioural dimension. For both dimensions we have defined a metamodel based on the categorical sketch. To specify the static dimension, we used a categorical sketch whose models are the states of the multiagent system. To specify the behavioural dimension, we introduced a Kripke-type categorical metamodel, which is based on a categorical sketch with constraints equivalent to the specified logical axioms.

Keywords: multiagent system, categorical sketch, Kripke model, metamodel, modal logic

1 Introduction

A formal definition of the concept of agent, unanimously accepted by the community in the field, does not exist, but a multitude of characterizations have been issued, each with their pluses and minuses [1]. Although these characterizations of the agents are different, depending on the specific applications, they all include the notions of the environment in which the agents evolve and their autonomy. Autonomy means the ability of an agent to respond to environmental changes through various actions.

If object-based modelling is characterized by the encapsulation of an object's attributes and access to them only through methods, agent-based modelling goes a step further and encapsulates the methods as well, so that access to the methods can only be done indirectly through messages. This approach leaves the freedom of the agent to decide which methods to use to achieve an objective.

Because the property of autonomy was quite confusing, in the definition of the concept of agent, it was later replaced with the property of flexible autonomy. The resulting agent concept is that of an intelligent agent that has three important characteristics [2]: reactivity, pro-activeness and social ability.

When designing a system, the designer seeks to achieve some global objectives that the resulting system must fulfil. Most of the time there is no agent capable of fulfilling these general objectives. An agent is able to fulfil local objectives. Global objectives can only be achieved by aggregating several agents, so that the fulfilment of their local objectives leads to the achievement of these global objectives. The result of this operation of adequate aggregation of agents, in order to fulfil some general objectives, is a multiagent system.

In general, the achievement of some objectives can be expressed by satisfying some logical formulas. One of the important objectives of logic is to provide languages and formal mechanisms for specifying reasoning on specific situations for the development of models. These languages and mechanisms must be, on the one hand, as intuitive as possible and on the other hand, endowed with rigorous syntax and semantics, without ambiguities so that they can be executed by a machine. In this context, modal logic plays a decisive role because it reduces the complexity of the language and at the same time preserves rigor.

Most of the time, in these models, besides the concepts of space, time, events, human or artificial agents appear and therefore the concepts of knowledge, action, belief. Each of these concepts requires its own way of reasoning and as a result of this difference, multimodal logic was introduced, which allows reasoning with several modal operators.

The reference model for specifying modal semantics is the Kripke frame, introduced by Saul Kripke in 1959, which nowadays has become a standard in specifying the semantics of multiagent systems, based on modal logic. We will introduce, in this paper, a categorical Kripke model for specifying the behavioural dimension of a multiagent system.

The main contribution that this paper brings is the specification of multiagent systems, at the metamodel level, using an appropriate categorical sketch. Category theory provides all the necessary ingredients for the formal specification and analysis of models [6, 4]. In our approach, a multiagent system is specified by two dimensions, namely: the static dimension and the behavioural dimension. For both dimensions we have defined a metamodel based on the categorical sketch.

Section 2 introduces some general notions and notations used in section 3 which presents the metamodel and the categorical model of a multiagent system. Section 4 concludes the paper with some conclusions and future papers.

2 Overview

A category $\boldsymbol{C}$ is a mathematical construct made up of two types of atomic components, namely formal functions that we call arrows, and objects that are the domains and codomains of the formal functions to which the function composition operation is added. In addition, the multitude of functions together with the composition operation form a monoid structure, i.e., it respects the associativity property and there is an identity function for each object. If $\boldsymbol{C}$ is a category, we will denote with $\boldsymbol{C}_0$ the set of objects of this category, and with $\boldsymbol{C}_1$ the set of arrows of the category.

In this paper, we will use especially the category that has sets as objects and as arrows functions with domains and codomains these sets, which we denote with Set, and we will also use the category that has graphs as objects and as arrows homomorphism between these graphs, which we denote by Grf. In this context, we will specify the multiagent systems, using as a metamodel, the categorical sketch, which is a mathematical object with precise syntax and implicit semantics. To specify the semantics of these systems, we will use the Kripke structures, which we will specify, also through a categorical sketch.

A categorical sketch is defined as a graph $\mathcal{G}$, together with a set of constraints $\mathcal{C}(\mathcal{G})$ imposed on the models through the components of the sketch graph. So, the categorical sketch $\mathcal{S}$ is a tuple $\mathcal{G}$=($\mathcal{G}$, $\mathcal{C}(\mathcal{G})$) [5, 3]. The constraints on the models can be specified by commutative diagrams, limits and colimits in the classical form of the sketch or by logical predicates in the case of the generalized sketch [5, 3]. Commutative diagrams, limits and colimits have a great advantage in modelling, because they are generic constructs and can also be used in the generalized sketch for specifying predicates.

A model of a categorical sketch $\mathcal{S}$=($\mathcal{G}$, $\mathcal{C}(\mathcal{G})$), is the image of the graph $\mathcal{G}$, through a homomorphism of graphs, in the category Set; M:$\mathcal{G}\to$Set, image that is subject to the constraints $\mathcal{C}(\mathcal{G})$. A wide range of diagrammatic models used in software engineering can be defined as categorical sketch models [3].

Functors are similar to graph homomorphisms only that they respect the monoid structure of the set of functions with the composition operation, i.e., associativity and conservation of the identity function. Graphs can also be extended to free categories by composing arrows and adding identity arrows to nodes. To simplify the exposition, we will continue to use the name functor even when it is a homomorphism of graphs.

To define the constraints, on the graph structure of the models, we need the diagram concept. A diagram D is a functor d, defined on a shape graph $\mathcal{P}$, with values in a category $\mathcal{C}$, i.e., a functor d:$\mathcal{P}\to\mathcal{C}$. A diagram has the property that its image in the category $\mathcal{C}$ preserves the shape graph $\mathcal{P}$, even if several nodes have the same label or several vertices have the same label [5, 12]. This means that the category could only be ambient for the image of diagram D, without this image mapping exactly on a portion of the category. The role of the diagrams is to link the formulas of the first order logic (FOL) to the components of the models.

We will specify the predicates that represent the constraints of the categorical sketch through the concept of diagram predicate signature. A set of predicates Π, together with an application ar:$\Pi\to$Grf$_0$, defines a diagram predicate signature. The application ar, maps each P$\in\Pi$ to a graph, from the category Grf, which is called shape graph arity of P. The images of the application ar in Grf$_0$ will be shape graphs for the diagrams that will map them to the components of the sketch graph $\mathcal{S}$ and therefore the specified constraints by predicates will propagate on the models through the diagrams.

Example 2.1. Let's suppose that we want to set the condition that the graph structure of all models of a sketch has the property that between any two nodes there is only one arc. This condition can be put by including in the diagram predicate signature, the predicate $P_1(x,y,z,r_{zx},r_{zy})=(\forall a_1,a_2\in z\Rightarrow((r_{zx}(a_1)=r_{zx}(a_2))\wedge(r_{zy}(a_1)= r_{zy}(a_2))\Rightarrow a_1=a_2))$ where the shape graph arity is $ar(P_1(x,y,z,r_{zx},r_{zy}))=Span(x,y,z,r_{zx},r_{zy})=(x\xleftarrow{r_{zx}}z\xrightarrow{r_{zy}}y)$ and $ar(x)=x$, $ar(y)=y$, $ar(z)=z$, $ar(r_{zx})=r_{zx}$, $ar(r_{zy})=r_{zy}$.

Shape graph Span(x,y,z,r_{zx},r_{zy}), will then be mapped by a diagram to the sketch graph components. Through the functor that defines the model, these constraints will reach the components of the model. The role of the shape graph construct is to keep the shape of the graph signature at the model level.

Intuitively, we can interpret the diagram predicate signature as a collection of procedures that implement the constraints defined by predicates having as parameters, formal parameter graph, and diagrams, by means of some functors, associate formal parameter graph to actual parameter graph, nodes to nodes and arcs to arcs.

In this paper we will use the categorical sketch to specify, at the metamodel level, the static dimension of a multiagent model and also to represent the behavioural dimension of the system.

Each model of the categorical sketch that specifies the static dimension of the system is a state of the system and is characterized by the graph structure of the system and the values of some attributes attached to the components of the model, at a given moment [18]. The sketch models that represent the behavioural dimension are Kripke type models, which have the states of the system as possible worlds. The transitions of the system are the result of the actions of the agents, who act to fulfil some local objectives. Each type of agent is endowed with a specific modal logic.

From a syntactic point of view, the basic modal logic language contains well-formed formulas, with the classic propositional logic operators to which two unary modal operators $\Box$ and $\Diamond$ are added. Depending on the specified modal logic, the two modal operators can have various interpretations, for example required for $\Box$, and possible for $\Diamond$. The two operators are linked by the relation $\Diamond\varphi=\neg\Box\neg\varphi$, where φ is a logical formula, and therefore, they are not independent.

The standard for interpreting the formulas of modal logic are the Kripke models. A Kripke type model is a tuple M=($\mathcal{K}$,$\mathcal{P}$,π), where $\mathcal{K}$=($\mathcal{W}$,$\mathcal{R}$), is a graph, which is called a Kripke frame, $\mathcal{W}$ is a set of possible worlds, $\mathcal{R}$ is the accessibility relation on the set $\mathcal{W}$, $\mathcal{P}$ is a set of atomic propositions and $\pi:\mathcal{W}\rightarrow 2^{\mathcal{P}}$, is an evaluation application that returns for each possible world $w\in\mathcal{W}$, the atomic propositions satisfied in the respective world.

If we denote by $\mathcal{R}$(v, w) the arc in $\mathcal{R}$ that connects the world $v\in\mathcal{W}$ to the world $w\in\mathcal{W}$, then we can check if a well-formed formula is satisfied in the world v($v\Vdash\varphi$) of the model M, inductively as [15]:

M,v $\Vdash$ p $\Leftrightarrow$ p $\in\pi$(v);

M,v $\Vdash\neg\varphi\Leftrightarrow$ M,v $\nVdash\varphi$;

M,v $\Vdash\varphi\land\psi\Leftrightarrow$ M,v $\Vdash\varphi$ and M,v $\Vdash\psi$;

M,v $\Vdash\varphi\lor\psi\Leftrightarrow$ M,v $\Vdash\varphi$, or M,v $\Vdash\psi$;

M,v $\Vdash\varphi\rightarrow\psi\Leftrightarrow$ M,v $\Vdash\varphi$ implies M,v $\Vdash\psi$;

M,v $\Vdash\varphi\leftrightarrow\psi\Leftrightarrow$ (M,v $\Vdash\varphi\Leftrightarrow$ M,v $\Vdash\psi$);

M,v $\Vdash\Box\psi\Leftrightarrow$ (for each $w\in\mathcal{W}$ with $\mathcal{R}$(v,w), we have M,w$\Vdash\psi$);

M,v $\Vdash\Diamond\psi\Leftrightarrow$ (there is a $w\in\mathcal{W}$ such that $\mathcal{R}$(v, w) and M,w $\Vdash\psi$).

In general, the evaluation of modal formulas depends on the axioms we impose. In modal logic, one starts from an axiom, called axiom K:$\Box(\varphi\rightarrow\psi)\rightarrow(\Box\varphi\rightarrow\Box\psi)$. Other important axioms that have been imposed in modal logic are: T:$\Box\varphi\rightarrow\varphi$; B:$\varphi\rightarrow\Box\Diamond\varphi$; D:$\Box\varphi\rightarrow\Diamond\varphi$; 4:$\Box\varphi\rightarrow\Box\Box\varphi$ and 5:$\Diamond\varphi\rightarrow\Box\Diamond\varphi$.

In Kripke models, as we can see, the evaluation of a modal formula depends a lot on the accessibility relation $\mathcal{R}$. There is an equivalence between axioms, or other logical formulas and the structure of the accessibility relation $\mathcal{R}$. In our approach, we will take advantage of this equivalence and impose the axioms through constraints on the relation $\mathcal{R}$, in a categorical sketch.

3 The Categorical Model of a Multiagent System

Agents are autonomous physical or logical entities that can perform actions in a certain environment in order to fulfil certain objectives. In general, they can observe environmental changes caused by other agents in a certain context and will make decisions based on these changes. We assume that each agent is endowed with a finite set of actions that it can perform. The performance of some actions is conditional on the state of the environment, which it perceives through the associative preconditions of each action and the objectives it pursues.

Based on the preconditions and objectives pursued, the agent will have to choose from the possible actions those that satisfy his objectives in optimal conditions. This choice implies a certain logical reasoning in each state. In our approach, each agent will be endowed with a certain modal logic, based on which the agent will reason for decision-making.

When designing a multiagent system, the designer pursues a general objective, which cannot be known by each agent. Agents only have local objectives. Therefore, the system designer will have to make a convenient aggregation of a lot of local objectives to achieve the general objective.

In order to achieve the objectives, most of the time agents must communicate and cooperate with each other. Therefore, a good structuring of a multiagent system will have to allow the encapsulation of cooperating agents in appropriate substructures that could cooperate or compete with other substructures depending on the general objectives.

There are many and quite different approaches related to the organization of agents, some more flexible others less flexible [13, 14]. The categorical sketch is a formal construction, suitable for structuring agents and flexible enough to allow a great diversity of structuring multiagent systems.

Each type of component is characterized by attributes and behaviour. A state of a component is represented by the values of the attributes at a given moment. A state of the system is represented by the graph structure of the system and the states of all its components at a given moment. The transactions of the system from one state to another are done by the actions of the components that can modify both the values of the attributes and the graph structure of the model within the limit allowed by the constraints $\mathcal{C}(\mathcal{G})$.

Even if the categorical sketch will only specify static models, i.e., states of the multiagent system, it facilitates the specification of dynamic components through other related mechanisms.

Each agent acts with the aim of achieving an objective: The objectives that an agent must achieve are specified by logical formulas, which will have to be satisfied in the following states. That is, an objective will be represented by a logical formula φ which is not satisfied in the current state but will have to be satisfied in the state after the action. The possibility of performing an action is also conditioned by the satisfaction of a logical formula in the current state. These decisions can be taken by an agent through a logical reasoning with which it is endowed, as will be presented in the next section. The transformations produced by the actions of the agents will be made within the limits allowed by the constraints of the categorical sketch.

Also, agents can move from one substructure to another, or new agents can appear in the system or some agents can disappear from the system, thus modifying the initial structure of the model. These structural changes can be specified by graph transformations of the model within the limits allowed by the constraints of the categorical sketch.

In our approach, the agents' actions can have the effect of changing the values of the attributes as well as changing the graph structure of the model, within the limits allowed by the constraints of the categorical sketch. Therefore, we will specify a multiagent system on two dimensions, the static dimension and the behavioural dimension, each dimension with its own syntax and semantics.

3.1 The Static Dimension of the Multiagent Model

The syntax of the static dimension will be specified through a categorical sketch, and the semantics through the mapping of attributes to data domains and graphic structures to structures with known semantics (join, fork, etc.).

The behavioural dimension is given by the actions of the agents that can modify the values of the attributes and the graph structure of the model. The syntax of actions will be represented by action signatures, and the semantics by mapping signatures to the double pushout algorithm and to algorithms that transform the values of the attributes.

To define a categorical sketch, we need a diagram predicate signature which is a construct $\Theta=(\Pi,ar)$ formed by a set of predicates Π and an application $ar:\Pi\rightarrow Grf_0$, which associates to each predicate P from Π an object from the Grf category, object called graph arity of P.

A categorical sketch $\mathcal{S}$ is a tuple $\mathcal{S}=(\mathcal{G}, \mathcal{D}(\Theta))$, where $\mathcal{G}$ is a graph, called the sketch graph, Θ is a diagram predicate signature and $\mathcal{D}(\Theta)$ is a set of diagrams indexed by the set of predicates Π, which have as shape graph, the images of the application ar in the Grf category [3]. That is, $\mathcal{D}(\Theta)=\{d_P:ar(P)\rightarrow\mathcal{G}|P\in\Pi\}$. Thus, $\mathcal{D}(\Theta)$ defines the signatures of the predicates that represent the constraints of the sketch on the structure of the models and maps them to the components of the sketch graph. A model of a sketch $\mathcal{S}$ is the image of a functor $M:\mathcal{G}\rightarrow Set$, which satisfies all the constraints specified by $\mathcal{D}(\Theta)$ and represents an instance of the multiagent system, i.e., a state of this system.

Next, we will consider that a directed graph is specified by a set of objects X, called nodes, a set of arcs Γ, and two functions $s,t:\Gamma\rightarrow X$, which associate the source and target nodes to the arcs.

In our approach, the nodes of the sketch graph represent concepts of the model such as agents, objects, groups of agents and roles, as well as elements of organizing and structuring of these components to form an adequate system. Sketch graph arcs represent sketch operators meant to help define constraints. Therefore, the sketch graph will have to be a type graph [7, 10, 11]. We will consider that the label of each node or arc of the sketch graph will correspond to the label of its type. That is, if we have a sketch $\mathcal{S}$ and a model $M:\mathcal{S}\rightarrow Set$, then, for each component c of the model $type_M(c)=C$ where $c\in M(C)$

and C is a component of the graph $\boldsymbol{\mathcal{G}}$. Thus, this categorical sketch becomes a metamodel for the static dimension of a multiagent system.

Therefore, the sketch graph nodes represent types such as agent type, object type, group type, role type, etc., or types of relationships between them. These types can be decomposed into subtypes, if the constraints on the structure of the models require this. In Fig. 1, we have an example of a categorical sketch graph. We denoted with A the agent type, with O the object type, with G the group type and with R the role type. We also denoted with γ_{XY}, the type of relationship between type X and type Y, meaning that concepts of type X can contain concepts of type Y. Of course, we cannot present a graph of the sketch that satisfies the requirements of all applications, and we do not intend to do so.

As we have already mentioned, the generalized sketch defines the constraints in the form of signatures of logical predicates that have as variables the nodes and arcs of shape graphs. These special graphs, called shape graphs, are mapped by diagrams to the sketch graph components (Fig. 2).

Example 3.1. In Example 2.1. we specified the constraint that between any two nodes of the models there should be a single arc, by including in the diagram predicate signature, the predicate $P_1(x,y,z,r_{zx},r_{zy})$ with shape graph arity, $ar(P_1(x,y,z,r_{zx},r_{zy}))=(x\xleftarrow{r_{zx}}z\xrightarrow{r_{zy}}y)$. To map this shape graph to the components of the graph $\boldsymbol{\mathcal{G}}$ of the sketch $\boldsymbol{\mathcal{S}}$, we will use a diagram $d_1:(x\xleftarrow{r_{zx}} z \xrightarrow{r_{zy}}y)\rightarrow\boldsymbol{\mathcal{G}}$, defined as follows: $d_1(x)=d_1(y)=X$, $d_1(z)=\Gamma$, $d_1(r_{zx})=s$, $d_1(r_{zy})=t$. Note that although the nodes x,y in the shape graph are mapped to the same node X of the graph $\boldsymbol{\mathcal{G}}$, the image of diagram d_P keeps the graph shape.

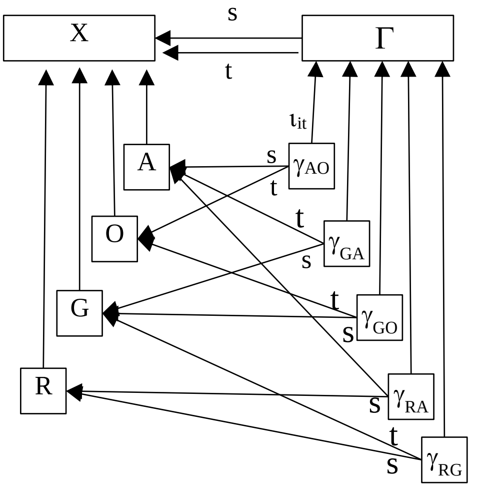


Fig. 1. An example sketch graph

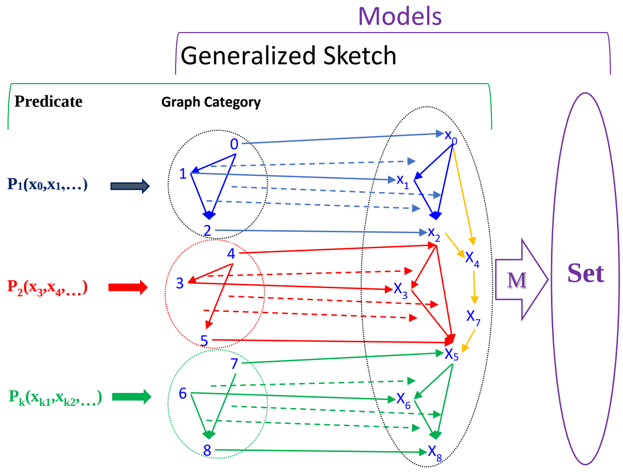


Fig. 2. Mapping diagram predicate signature to the model

Through the functor that defines the model, these constraints will reach the components of the model. The role of the shape graph construction is to keep the graph shape at the model level.

Although universal constructs such as commutative diagrams, limits and colimits, are used especially in the classical categorical sketch, many times they are also very useful in the generalized sketch to define predicates as we can see in Example 3.

Example 3.2. If we want to constrain the structure of any model of the sketch to be a connected graph, we can use a special limit, namely pushout, which is the limit of a span. It is known that the pushout of two functions f and g with the same domain of definition, which forms a span, $X \xleftarrow{f} Z \xrightarrow{g} Y$, coincides with the set of equivalence classes determined by the equivalence relation ρ induced by the relation: xρy, x∈X, y∈Y if and only if ∃ z∈Z so that f(z)=x and g(z)=y.

In the case of a graph, the pushout of the source s and target t functions from the definition of a graph coincides with the equivalence classes determined by the equivalence relation induced by s and t on the set of graph nodes, i.e., it coincides with the set of connected components of the graph. Therefore, the constraint that the graph of any model is connected can be put by the predicate $P_2(x,y,z,r_{zx},r_{zy})=|\text{pushout}(x \xleftarrow{r_{zx}} z \xrightarrow{r_{zy}} y)|=1$, i.e. the cardinal of the set $\text{pushout}(x \xleftarrow{r_{zx}} z \xrightarrow{r_{zy}} y)$, which represents the number of connected components of the graph to be 1, with the shape graph arity, $ar(P_2(x,y,z,r_{zx},r_{zy}))=(x \xleftarrow{r_{zx}} z \xrightarrow{r_{zy}} y)$, and diagram d_2, which coincides with diagram d_1, from Example 3.2.

A model of the categorical sketch represents a state of the multiagent system. The semantics of a state is characterized by the values of the attributes and the graphical structure of the model. Also, a state of the model can be endowed with a lot of atomic logical formulas that characterize the values of the attributes and the graph structure of the model.

3.2 The Behavioral Dimension of the Multiagent Model

The behavioural dimension of a multiagent system is characterized by its syntax and semantics. To specify the syntax of the behavioural dimension of a multiagent system, we will use as a metamodel, also a categorical sketch, as in the case of the static dimension, and to specify its semantics we will use a Kripke type model.

The static dimension of the categorical model specifies the set of possible states of the system. Thus, if we have a categorical sketch $\boldsymbol{\mathcal{S}}$, there are many functors M:$\boldsymbol{\mathcal{S}}$→Set, and each functor M represents a state of the model. We denote this set of functors by Mod($\boldsymbol{\mathcal{S}}$, Set). We assume that the set Mod($\boldsymbol{\mathcal{S}}$,Set) is at most countable and, therefore, we can index it with natural numbers.

The transition from one state to another of the behavioural model is done by the actions of the agents involved in this model [9]. Thus, if the initial state of the behavioural model is $\mathfrak{I}_0$, then the set of possible states $\mathfrak{I}$ of the behavioural model is: $\mathfrak{I}=\{\mathfrak{I}_k \in \mathrm{Mod}(\boldsymbol{\mathcal{S}}, \mathrm{Set}), k \geq 1 \mid \mathfrak{I}_k$ is the result of successive actions of the agents on $\mathfrak{I}_0\}$.

In this notation we will understand that a functor $\mathfrak{I}_k$, represents the image of the sketch $\boldsymbol{\mathcal{S}}$, through the functor $\mathfrak{I}_k$ in the Set category.

In our approach, an agent is characterized by a multitude of actions it can perform and by the logic it uses to make decisions. The performance of an action is conditioned by the state in which the model is located and by the objective it pursues in the following states.

An action can change the values of the attributes of the model, but it can also change the graph structure of the model because, as we mentioned before, agents can move from one substructure to another, new agents can appear or some of them can disappear. All these changes can only be made within the limits allowed by the constraints of the categorical sketch.

We will specify the structural changes of the models caused by the agents' actions through graph transformations. A graph transformation, $\boldsymbol{\tau}$=(L, R), is composed of two graphs, namely the left graph L and the right graph R, and a mechanism that specifies the conditions and the way to replace L with R. Because the structural changes, in our case, involves both deletions and additions, we will use the double-pushout variant (DPO), a graph transformation that is specified by three graphs L, R and K and two graph morphisms l and r: $\boldsymbol{\tau}=(L \xleftarrow{l} K \xrightarrow{r} R)$ where K is an interface graph contained in both R and L [10, 11, 7].

The components of the graphs L and R will be mapped to the components of the graph $\boldsymbol{\mathcal{G}}$, of the sketch $\boldsymbol{\mathcal{S}}$, by a pair of diagrams d_L and d_R, and will receive the types of the components of the graph $\boldsymbol{\mathcal{G}}$, and therefore will be shape graph typed. Now we can define an action of an agent as a pair a=($\boldsymbol{\tau}$, p), where $\boldsymbol{\tau}$ is a graph transformation $\tau=(L \xleftarrow{l} K \xrightarrow{r} R)$, and p is a procedure with the property that R=p(L). Therefore, performing an action a=($\boldsymbol{\tau}$, p), consists in applying the graph transformation $\boldsymbol{\tau}$ and calculating the values of some local attributes, associated with graphs L, R and K. Of course, the graph transformation could be an identity transformation that does not change the graph structure of the system in any way, but it is still useful because the graph L from the graph transformation $\boldsymbol{\tau}$ also has the role of locating the agent's action

To model the semantics of the behavioural dimension of multiagent systems, we will use Kripke-type structures, which have an implicit semantics. In our approach, each agent is characterized by a multitude of actions that it is capable of performing and by the logic that it uses to evaluate the logical formulas necessary for decision-making.

If all the agents use the same logic to evaluate the formulas, then, in our approach, a semantic model of a multiagent system is a tuple:

M=($\mathcal{K}$, $\mathcal{P}$, π, w_0, $\mathcal{A}$c, α, $\mathcal{A}$g, β, δ), where:

$\mathcal{K}$=($\mathcal{W}$, $\mathcal{R}$) is a graph in which the set of nodes $\mathcal{W}$ represents the set of possible worlds and $\mathcal{R}$ is a set of arcs between the elements of $\mathcal{W}$;

$\mathcal{P}$ is a set of atomic propositions, and $\pi:\mathcal{W}\rightarrow 2^{\mathcal{P}}$, is an evaluation application, which associates to each world the set of valid atomic propositions in that world.

$\mathcal{A}$c is a set of actions, and α is a surjective function α:$\mathcal{R}$→$\mathcal{A}$c, which associates an action to each arc;

$\mathcal{A}$g is a set of agents, and β is a surjective function β:$\mathcal{R}$→$\mathcal{A}$g, which associates an agent to each arc;

δ is a surjective function δ:$\mathcal{A}$c→$\mathcal{A}$g, which associates to each agent the set of actions it is capable of performing and which has the property $\beta=\delta\circ\alpha$.

We notice that the graph $\mathcal{K}$=($\mathcal{W}$, $\mathcal{R}$), is, in fact, a Kripke frame, where $\mathcal{W}$ is a set of possible worlds, and $\mathcal{R}$ is the accessibility relation on the set $\mathcal{W}$. Also, the tuple ($\mathcal{K}$, $\mathcal{P}$, π) specifies a Kripke semantic model for modal logic.

The theoretical results showed that there is a direct correspondence between the satisfaction of some schemes of modal logical formulas and the properties of the accessibility relation $\mathcal{R}$ [17, 8, 15, 16]. Thus, for axiom T to be satisfied, relation $\mathcal{R}$ must be reflexive, axiom B is satisfied if relation $\mathcal{R}$ is symmetric, axiom D is satisfied if relation $\mathcal{R}$ is serial, axiom 4 is satisfied if relation $\mathcal{R}$ is transitive, and axiom 5 is satisfied if the relation $\mathcal{R}$ is Euclidean. Also, if the relation $\mathcal{R}$ is functional, the formula $\Box\varphi\leftrightarrow\Diamond\varphi$ is satisfied, and if the relationship is linear, it satisfies the formula $\Box(\varphi\wedge\Box\varphi\rightarrow\psi)\vee\Box(\psi\wedge\Box\psi\rightarrow\varphi)$.

Although a Kripke frame $\mathcal{K}$=($\mathcal{W}$, $\mathcal{R}$), does not contain the atomic formulas satisfied in every world w∈W, it is important to be able to specify a Kripke frame that satisfies certain formula schemes as a whole. Thus, a Kripke frame $\mathcal{K}$=($\mathcal{W}$, $\mathcal{R}$), satisfies a formula scheme ϕ, i.e., $\mathcal{K}\Vdash\phi$, if for any evaluation application $\pi:\mathcal{W}\rightarrow 2^{\mathcal{P}}$, in every world $w\in\mathcal{W}$, $\mathcal{K},w\Vdash\phi$. Also, we have a theoretical result that says a Kripke frame $\mathcal{K}$, which satisfies a scheme of formulas, satisfies all substitution instances of that formula [15, 17].

Therefore, a model M=($\mathcal{K}$, $\mathcal{P}$, π, w_0, $\mathcal{A}$c,α, $\mathcal{A}$g, β, δ), of a multiagent system characterized by a certain logic can be specified by constraints on the accessibility relation. Thus, if we want the model to be characterized by the KT4 logic, we will introduce the constraints that the relation $\mathcal{R}$ be reflexive and transitive, if we want the model to be characterized by the KT45 logic, we will introduce the constraints that the relation $\mathcal{R}$ be reflexive, transitive and Euclidean, etc.

A model of a multiagent system may include concepts such as time, knowledge, belief, obligations, etc., which determine the formula schemes that must be satisfied. Therefore, in a multiagent system we will have to have for each type i of agents a distinct $\Box_i$ operator and, implicitly, a distinct $\Diamond_i$ operator. The i index of the agent type in modal operators implies distinct accessibility relations.

When designing a model for a multiagent system, it is important to establish precisely which formula schemes we need to be satisfied by each type of agent, and to specify constraints so that they are equivalent to these formula schemes.

As we have already mentioned, for the specification of a metamodel, of the behavioural dimension of a multiagent system, we will use the categorical sketch. We will denote this sketch by $\boldsymbol{\mathcal{S}}^B=(\boldsymbol{\mathcal{G}}^B, \boldsymbol{\mathcal{D}}^B(\Theta))$. In this case, the concepts involved in the metamodel are the agents Ag, with the corresponding subtypes, the relation R, with the sub-relations corresponding to each type of agent, and the actions Ac, with the subtypes corresponding to the types of agents. Thus, the graph of the categorical sketch, which specifies the behavioural dimension of a multiagent system, can be the one in Fig. 3.

In our approach, the set of agents Ag will be a disjoint union of subsets of different types of agents; Ag=$\coprod_{i=1}^{n} Ag_i$, the relation R is a disjoint union of sub-relations; R=$\coprod_{i=1}^{n} R_i$, and the set of actions Ac is a disjoint union of subsets of actions of different types; Ac=$\coprod_{i=1}^{n} Ac_i$. These constraints can be elegantly be imposed, in the categorical sketch, by colimits of discrete diagrams [5, 3], which we will not specify in this paper.

The constraints on the categorical sketch models are represented by the component $\boldsymbol{\mathcal{D}}^B(\Theta)$ by diagrams indexed by the set of predicates Π, which have as shape graph, the images of the application ar in the Grf category.

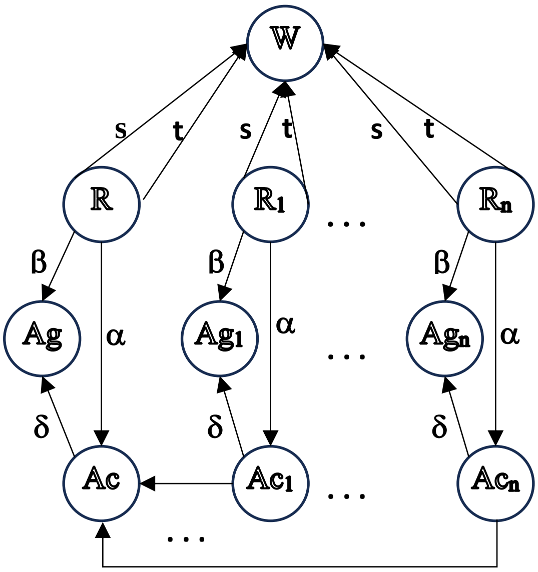


Fig. 3. The categorical sketch graph specifying the behavioral dimension

An important part of the predicates involved in such a metamodel are those that impose restrictions on the relation R and on the sub-relations R_i, 1≤i≤n. We will now exemplify some predicates that can impose constraints on the accessibility relations, conditions that can replace the logical formulas as mentioned above.

The constraint that a relation Ri, be reflexive can be put through the predicate:

$Q_1(x,y,s,t)=\forall w\in x, \exists r\in y, (s(r)=t(r)=w)$.

The symmetry of a relation Ri can be imposed by the predicate:

$Q_2(x,y,s,t)=\forall r_1\in y, (s(r_1)=v\wedge t(r_1)=w\rightarrow\exists r_2\in y.\ s(r_2)=w\wedge t(r_2)=v)$.

A relation is serial if the predicate is satisfied:

$Q_3(x,y,s,t)=\forall w\in x,\ \exists r\in y,\ \exists v\in x.\ s(r)=w\wedge t(r)=v.$

A relation is transitive if the predicate is satisfied:

$Q_4(x,y,s,t)=\forall u,v,w\in x,\exists r_1,r_2\in y.s(r_1)=u\wedge t(r_1)=v\wedge s(r_2)=v\wedge t(r_2)=w\rightarrow\exists r\in y.s(r)=u\wedge t(r)=w.$

A relation is Euclidean if the predicate is satisfied:

$Q_5(x,y,s,t)=\forall u,v,w\in x,(\exists r_1,r_2\in y.s(r_1)=u\wedge t(r_1)=v\wedge s(r_2)=u\wedge t(r_2)=w\rightarrow\exists r\in y.s(r)= v\wedge t(r)=w).$

A relation is functional if the predicate is satisfied:

$Q_6(x,y,s,t)=\forall w\in x,\ \exists! r\in y.\ s(r)=w\wedge t(r)=v.$

A relationship is linear if the predicate is satisfied:

$Q_7(x,y,s,t)=\forall u,v,w\in x,(\exists r_1,r_2\in y.s(r_1)=u\wedge t(r_1)=v\wedge s(r_2)=u\wedge t(r_2)=w\rightarrow\exists r_3\in y,s(r_3)= v\wedge t(r_3)=w).$

A relation is complete if the predicate is satisfied:

$Q_8(x,y,s,t)=\ \forall v,w\in x,\ \exists r\in y.\ (s(r)=v\wedge t(r)=w)\vee(s(r)=w\wedge t(r)=v).$

All these predicates have the shape graph arity $ar(Q)=x\underset{t}{\overset{s}{\leftleftarrows}}y$, defined as follows:

$ar(x)=x$, $ar(y)=y$, $ar(s)=s$, $ar(t)=t$. Also, all these predicates will be mapped to the components of the sketch graph through a diagram $\eta_1:ar(Q)\rightarrow\boldsymbol{\mathcal{K}}$, thus: $\eta_1(x)=W$, $\eta_1(y)=R_i$, $\eta_1(s)=s$, $\eta_1(t)=t$, where i represents the relation on which we want to impose the constraint.

Therefore, the component $\boldsymbol{\mathcal{D}}^B(\Theta)$, of the categorical sketch $\boldsymbol{\mathcal{S}}^B$, will contain the diagram η_1, indexed by the predicates Q_i, $1\leq i\leq 8$.

Of course, these constraints can also be imposed through mechanisms specific to category theory, such as limits and colimits. For example, the condition that a relationship is total is equivalent to the condition that the graph corresponding to the relationship is connected. Therefore, the constraint that the relationship is total can be put through the predicate $Q(x,y,z,r_{zx},r_{zy})=|pushout(x\xleftarrow{r_{zx}}z\xrightarrow{r_{zy}}y)|=1$, which indexes the diagram d_1, from Example 2.

Other constraints imposed for the categorical sketch $\boldsymbol{\mathcal{S}}^B$ refer to the functions α,β,δ, respectively $\alpha_i,\beta_i,\delta_i$, $1\leq i\leq n$. These functions must be epimorphisms. The necessary and sufficient condition for an application $f:A\rightarrow B$ to be an epimorphism is that the pushout of f and f is isomorphic to A. This can be specified by the predicate: $S_1(x,y,f)= |pushout(x\underset{f}{\overset{f}{\rightrightarrows}}y)|=|x|$ with shape graph arity $ar(S_1)=x\underset{f}{\overset{f}{\rightrightarrows}}y$, defined as such $ar(x)=x$, $ar(y)=y$, $ar(f)=f$. This shape graph will be mapped to the components of the graph $\boldsymbol{\mathcal{G}}^B$, from Fig. 3, by a set of diagrams as follows: diagram μ^β defined as follows $\mu^\beta(x)=R$, $\mu^\beta(y)=Ag$, $\mu^\beta(f)=\beta$; diagrams μ_i^β defined as follows $\mu_i^\beta(x)=R_i$, $\mu_i^\beta(y)=Ag_i$, $\mu_i^\beta(f)=\beta_i$, $1\leq i\leq n$; diagram μ^α defined as follows $\mu^\alpha(x)=R$, $\mu^\alpha(y)=Ac$, $\mu^\alpha(f)=\alpha$; diagrams μ_i^α defined as follows $\mu_i^\alpha(x)=R_i$, $\mu_i^\alpha y)=Ac_i$, $\mu_i^\alpha(f)=\alpha_i$, $1\leq i\leq n$; diagram μ^δ defined as follows $\mu^\delta(x)=Ag$, $\mu^\delta(y)=Ac$, $\mu^\delta(f)=\delta$ and diagrams μ_i^δ defined as follows $\mu_i^\delta(x)=Ag_i$, $\mu_i^\delta y)=Ac_i$, $\mu_i^\delta(f)=\delta_i$, $1\leq i\leq n$;

The conditions that $\beta=\delta\circ\alpha$ and respectively $\beta_i=\delta_i\circ\alpha_i$ can be specified by the predicate:

$S_2(x,y,z,f,g,h) = \forall r \in x,\ f(x)=h(g(x)$, with shape graph arity $\mathrm{ar}(S_2)=x \begin{array}{c}\xrightarrow{\quad f \quad}\\ \xrightarrow{g} z \xleftarrow{h}\end{array} y$, defined as follows: ar(x)=x, ar(y)=y, ar(z)=z, ar(f)=f, ar(g)=g, ar(f)=f, ar(g)=g, ar(h)=h. This shape graph will be mapped to the graph components by the diagram ω defined as follows ω(x)=R, ω(y)=Ac, ω(z)=Ag, ω(f)=α, ω(g)=β, ω(h)=δ, and by the diagrams ω_i, defined as follows $\omega_i(x)=R_i$, $\omega_i(y)=Ac_i$, $\omega_i(z)=Ag_i$, $\omega_i(f)=\alpha_i$, $\omega_i(g)=\beta_i$, $\omega_i(h)=\delta_i$, $1\le i\le n$.

Therefore, in this metamodel $\Pi=\{Q_i|1\le i\le 8\}\cup\{S_1,S_2\}$, and $\mathcal{D}(\Theta)=\{\eta_1:\mathrm{ar}(Q_i)\to\mathcal{G}\ |1\le i\le 8\} \cup \{\mu^\alpha:\mathrm{ar}(S_1)\to\mathcal{G},\ \mu^\beta:\mathrm{ar}(S_1)\to\mathcal{G},\ \mu^\delta:\mathrm{ar}(S_1)\to\mathcal{G},\ \omega:\mathrm{ar}(S_2)\to\mathcal{G}\} \cup \{\mu_i^\alpha:\mathrm{ar}(S_1)\to\mathcal{G}\ |\ 1\le i\le n\} \cup \{\mu_i^\beta:\mathrm{ar}(S_1)\to\mathcal{G}\ |\ 1\le i\le n\} \cup \{\mu_i^\delta:\mathrm{ar}(S_1)\to\mathcal{G}\ |\ 1\le i\le 8\} \cup \{\mu_i^\alpha:\mathrm{ar}(S_i)\to\mathcal{G}\ |\ 1\le i\le n\} \cup \{\omega_i:\mathrm{ar}(S_2)\to\mathcal{G}\ |\ 1\le i\le n\}$. Depending on the requirements of the model, other constraints can be added, such as imposing a fixed or limited number of agents or actions.

A model of the sketch B is a functor $\mathfrak{B}:\mathcal{G}\to$Set, which satisfies all the constraints specified by $\mathcal{D}^B(\Theta)$, where: $\mathfrak{B}(W)$ is a set of possible worlds; $\mathfrak{B}(R)$ is a set of arcs that define the accessibility relation on the set of worlds $\mathfrak{B}(W)$; each $\mathfrak{B}(R_i)$, $1\le i\le n$, is a set of arcs that define the accessibility relation on the set of worlds of agents of type Ag_i; $\mathfrak{B}(Ag)$ is the set of agents involved in the model; each $\mathfrak{B}(Ag_i)$, $1\le i\le n$, is the subset of agents of type Ag_i; $\mathfrak{B}(Ac)$ is the total number of actions; each $\mathfrak{B}(Ac_i)$, $1\le i\le n$, is the subset of actions of type Ac_i. By the functor $\mathfrak{B}:\mathcal{G}\to$Set, we understand its image in the Set category, which represents a behavioural model of the multiagent system. We denote the set of these models by Mod($\mathcal{S}^B$, Set).

3.3 Aggregation of the Two Models

As we saw in the previous sections, both the static and the behavioural dimensions can be specified, at the metamodel level, through appropriate categorical sketches.

The static dimension of the categorical model is represented, at the metamodel level, by a categorical sketch $\mathcal{S}=(\mathcal{G}, \mathcal{D}(\Theta))$, which specifies the set of possible states of the system that we denoted with Mod($\mathcal{S}$, Set). The transition from one state to another of the model is done by the actions of the agents involved in this model. If the initial state of a model is $\mathfrak{I}_0, \in$ Mod($\mathcal{S}$, Set) then the set of possible states $\mathfrak{I}$ of the behavioural model becomes:

$\mathfrak{I}=\{\mathfrak{I}_k\in$Mod($\mathcal{S}$, Set), $k\ge 1$ | $\mathfrak{I}_k$ is the result of successive actions of the agents on $\mathfrak{I}_0\}$.

The behavioural dimension of a multiagent system is represented, at the metamodel level, by the categorical sketch $\mathcal{S}^B=(\mathcal{G}^B, \mathcal{D}^B(\Theta))$, as we saw in the previous section.

Now we can aggregate the two models to specify the categorical model of a multiagent system. A categorical metamodel for multiagent systems is a tuple MM=($\mathcal{S}, \mathcal{S}^B$), where $\mathcal{S}$ is a categorical sketch that represents a metamodel for the static dimension of the multiagent system, and $\mathcal{S}^B$ is a categorical sketch that represents a metamodel for the behavioural dimension of the multiagent system.

A categorical model of a multiagent system is a tuple: $\mathcal{M}=(\mathfrak{I}, \mathfrak{B}, \mathfrak{I}_0, \mathcal{P}, \pi)$, where $\mathfrak{I}\subseteq$Mod($\mathcal{S}$,Set), is a set of static models of the sketch $\mathcal{S}$; $\mathfrak{B}\subseteq$Mod($\mathcal{S}^B$,Set), is a behavioural model of the $\mathcal{S}^B$ sketch, with the property that $\mathfrak{B}(W)= \mathfrak{I}$; $\mathfrak{I}_0\in\mathfrak{I}$, is the initial state of the system; $\mathcal{P}$ is a set of atomic propositions, which characterizes the state of the system, and $\pi:\mathfrak{I}\to 2^{\mathcal{P}}$, is an evaluation application, which associates to each state the set of valid atomic propositions in that state.

In this paper, we present only the general idea that is the basis of the specification of the logical language that specifies the behaviour of the agents without going into details. Each agent acts with the aim of achieving an objective. The objectives of an agent are specified by logical formulas, which it can satisfy through the actions it is able to perform. The agent pursues the satisfaction of some objective formulas that are not satisfied in the current world but can be satisfied after performing some actions, that is, in the following worlds.

For each type of Ag_i agent, there is a subset $\mathcal{P}_i$ of atomic formulas that such an agent can satisfy through its actions. Therefore, the set of logical formulas that a type of Agi agent can satisfy will all be well-formed formulas, starting from the set of atomic formulas that it can satisfy using logical operators. In a similar way, the logical formulas that can be satisfied by a group of agents can be specified.

4 Conclusions and Future Work

The most important conclusion of the paper is the fact that the categorical sketch is an appropriate mathematical mechanism for specifying models for multiagent systems, at the metamodel level. This metamodel can be the basis of the implementation of a domain-specific modelling tool endowed with a diagrammatic language.

Constraints on models can be specified by universal properties in category theory, which can be implemented in the form of generic algorithms that work on all models specified by categorical sketches.

I mention the fact that in this paper I did not deal with the problems related to parallelism and synchronization. Of course, the agents' actions can take place in parallel with the sharing of common resources, but this will be the subject of another paper.

In the semantics based on the Kripke model, the accessible worlds are fixed. In a reactive Kripke model [19], the evaluation of logical operators can cause the reconfiguration of the model in which the formula is evaluated. In our model, graph transformations can be used to model this reconfiguration, but this will be the subject of another paper.

References

[1] S. Franklin, A. Graesser, “Is it an agent or just a program?” In: Proceedings of the 3rd International Workshop Agent Theories, Architectures, and Languages. Springer, New York (1996)

[2] M. Wooldridge, “An Introduction to MultiAgent Systems”, 2nd Edition, John Wiley & Sons Ltd 2009.

[3] Uwe Wolter, Zinovy Diskin, “The Next Hundred Diagrammatic Specification Techniques, A Gentle Introduction to Generalized Sketches”, 02 September 2015 : https://www.researchgate.net/publication /253963677.

[4] Zinovy Diskin, “Tom Maibaum- Category Theory and Model-Driven Engineering: From Formal Semantics to Design Patterns and Beyond”, ACCAT 2012

[5] Michael Barr, Charles Wells, "Category Theory For Computing Science", Reprints in Theory and Applications of Categories, No. 22, 2012.

[6] Robin Milner, "The Space and Motion of Communicating Agents", Cambridge University Press, 2009. ISBN 978-0-521-73833-0

[7] Hartmut Ehrig, Claudia Ermel, Ulrike Golas, Frank Hermann, "Graph and Model Transformation General Framework and Applications", Springer-Verlag Berlin Heidelberg 2015

[8] Olivier Gasquet, Andreas Herzig, Bilal Said, François Schwarzentruber, "Kripke's Worlds, An Introduction to Modal Logics via Tableaux", Springer Basel AG 2014.

[9] J. Luo, JJ. Meyer, M. Knobbout, "Reasoning About Opportunistic Propensity in Multi-agent Systems." In: Sukthankar, G., Rodriguez-Aguilar, J. (eds) Autonomous Agents and Multiagent Systems. AAMAS 2017. Lecture Notes in Computer Science(), vol 10642. Springer, Cham. https://doi.org/10.1007/978-3-319-71682-4_13

[10] G. Campbell, "Algebraic graph transformation: A crash course", Department of Computer Science, University of York, UK, Tech. Rep., 2018. [Online]. Available: https://cdn.gjcampbell.co.uk/2018/Graph-Transformation.pdf.

[11] G. Campbell, B. Courtehoute and D. Plump, "Linear-time graph algorithms in GP2", Department of Computer Science, University of York, UK, Submitted for publication, 2019. [Online]. Available: https://cdn.gjcampbell.co.uk/2019/Linear-Time-GP2-Preprint.pdf.

[12] David I. Spivak, "Category Theory for the Sciences", The MIT Press Cambridge, Massachusetts London, England, 2014 Massachusetts Institute of Technology

[13] J. Ferber, O. Gutknecht and F. Michel, „From Agents to Organizations: An Organizational View of Multi-Agent Systems", In Proc. 4th International Workshop on Agent-Oriented Software Engineering (AOSE), volume 2935 of LNCS, pages 214–230. Springer, 2003.

[14] Virginia Dignum, "Handbook of Research on Multi-Agent Systems: Semantics and Dynamics of Organizational Models", Information Science Reference, 2009 by IGI Global.

[15] M. Huth, M. Ryan, "Logic in Computer Science, Modelling and Reasoning about Systems", Published in the United States of America by Cambridge University Press, New York , 2004.

[16] Y. Shoham, K. Leyton-Brown, "Multiagent systems_ algorithmic, game-theoretic, and logical foundations" -Cambridge University Press, 2009.

[17] Nino B. Cocchiarella and Max A. Freund, "Modal Logic, An Introduction to Its Syntax and Semantics", Published by Oxford University Press, Inc. 2008

[18] D.C. Crăciunean, D. Karagiannis, "A categorical model of process cosimulation", Journal of Advanced Computer Science and Applications(IJACSA), 10(2), (2019).

[19] D. Gabbay. "Reactive Kripke semantics", In Proceedings of CompLog 2004, W. Carnielli, ed., pp. 7–20. Centre of Logic and Computation, University of Lisbon, 2004.

DOI: 10.2478/ijasitels-2024-0012

Logistics Optimization for Resource Allocation and Scheduling Using Time Slots

Cezar-Marian Papară [1], Ștefan-Horia Schirliu [2]

[1]PhD student at "Alexandru Ioan Cuza" University, Faculty of Computer Science, 16 Berthelot St., Iasi, 700506, România, cezarmarian98@gmail.com

[2]Master's student at Faculty of Sciences, "Vasile Alecsandri" University of Bacău, 157 Cal. Mărășești, Bacău, 600115, România, stefanschirliu@gmail.com

Abstract

In the field of logistics, efficient scheduling and resource allocation are essential for ensuring the seamless flow of goods through transportation networks. This paper addresses the Interval Scheduling Problem, a combinatorial optimization challenge, in the context of logistics planning for goods transportation. The study examines how optimized appointment scheduling and resource allocation can enhance the performance of transportation networks. By combining theoretical insights, algorithmic solutions, and practical applications, this work proposes a comprehensive approach grounded in mathematical models that account for time, resource, and capacity constraints, alongside a computational implementation. Utilizing advanced computational techniques and real-time data integration, the proposed solutions aim to increase operational effectiveness and competitiveness while reducing costs in transportation logistics.

Keywords: Interval Scheduling Problem, Time Slots, Resource Allocation, Combinatorial Optimization, Logistics, Transportation Networks.

1 Introduction

In today's interconnected world, large-scale transportation networks play a critical role in facilitating the efficient movement of goods and services. These networks form the backbone of global trade, supply chains, and local distribution systems [1]. Improving the efficiency of such networks is essential, as it directly impacts economic growth, operational costs, and service quality. Consequently, optimizing logistics operations within transportation networks has become a central focus of both academic research and practical development efforts [2].

While our previous research has predominantly addressed external aspects of transportation networks—such as routing and path optimization, including the well-known Traveling Salesman Problem (TSP) and Vehicle Routing Problem (VRP) [14], [15]—a less explored area is the internal logistics within network nodes. This paper shifts the focus to understanding how goods are processed once they enter a node, such as a warehouse or distribution center, and how constraints and optimization decisions within these nodes—related to resource allocation, timing, and capacity—affect the overall flow and efficiency. Although past work has utilized heuristics, greedy algorithms, and dynamic programming for optimization [16], [17], [18], [19], there is still much to be explored regarding the internal logistics operations of these nodes.

This study seeks to address this research gap by investigating the Interval Scheduling Problem (ISP) [3] as it applies to logistics within transportation networks. The ISP focuses on allocating tasks or activities within specific time intervals to maximize operational efficiency and minimize conflicts or overlaps in resource usage. In the logistics context, this involves scheduling deliveries, transport operations, and inventory management tasks across different logistical points in a way that ensures smooth and uninterrupted operations [4].

Classified as NP-hard due to its computational complexity [5], ISP necessitates advanced algorithmic solutions for effective resolution. Researchers and practitioners rely on dynamic programming techniques to decompose the scheduling problem into smaller, manageable subproblems. By recursively solving these subproblems and storing their solutions, dynamic programming optimizes resource allocation and minimizes scheduling conflicts within the constraints of time availability and resource capacity.

The research question at the core of this study is: How can the Interval Scheduling Problem be effectively applied to optimize internal logistics within transportation network nodes, accounting for temporal, resource-based, and capacity constraints? By addressing this question, the study aims to contribute to the development of more efficient and reliable scheduling strategies that improve both operational performance and service quality in logistics operations. The findings are expected to provide practical insights into the management of internal logistics challenges and offer robust solutions for complex scheduling problems in this domain.

The structure of the paper is as follows: Section 1 provides a review of both classical and contemporary literature on scheduling problems. Section 2 presents a mathematical model that addresses key constraints, such as operating hours, national holidays, team holidays, and the prevention of overlapping appointments within the same time intervals and available resources. Section 3 introduces the implemented solution using state-of-the-art programming frameworks. Finally, Section 4 analyses data sets and results, culminating in a set of conclusions based on the study’s findings.

2 Literature Review

2.1 Challenges of Uncertainty in ISP

One of the key challenges in implementing ISP in logistics is managing uncertainty in processing and setup times. Article [6] highlights these challenges, emphasizing the need for resilient scheduling techniques such as robust optimization, stochastic programming, and scenario-based approaches to handle uncertainties effectively. These techniques are essential for minimizing disruptions and ensuring effective resource allocation in logistics operations.

2.2 Mathematical Models for Optimization in Logistics

A substantial body of literature focuses on the development of mathematical models to optimize logistics operations. Article [7] introduces several crucial models such as linear programming, integer programming, and constraint programming, which are widely used for optimizing deliveries, vehicle routes, and work shifts. These models maximize resource utilization and minimize operational costs, playing a pivotal role in ISP implementation across various logistical nodes.

2.3 Real-Time Decision Making in ISP

The integration of real-time decision-making techniques into ISP is another significant theme in the literature. Article [8] explores the use of predictive models, including machine learning, to enhance scheduling accuracy. By incorporating online algorithms with predictive analytics, schedules can be dynamically adjusted based on real-time data, allowing logistics operations to respond swiftly to unexpected changes and improve overall operational efficiency.

2.4 Heterogeneous Entities in ISP

Another critical aspect of ISP in logistics is the consideration of heterogeneous entities—both in terms of agents and machines. Article [9] focuses on how heterogeneous agents with diverse preferences and constraints can complicate scheduling, necessitating robust allocation mechanisms that ensure fairness and efficiency without overlap. Similarly, article [10] addresses the complexities introduced by nonidentical machines with varying processing capacities, demonstrating how task allocation must be optimized to fully utilize these resources and enhance scheduling efficiency.

2.5 Task Sequencing in ISP

The sequencing of tasks is crucial in optimizing logistics operations (Fig. 1). Article [11] examines the permutation flow-shop scheduling problem, where tasks must be processed in a specific order across multiple machines. By applying constrained programming techniques, the study improves the sequencing of tasks to reduce delays and maximize throughput, which is directly applicable to ISP challenges in logistics.

2.6 Resource-Related Constraints in ISP

Scheduling in logistics is also influenced by various resource-related constraints, including team capacity, availability, and holidays. Article [12] investigates these variables using integer linear optimization to develop flexible scheduling solutions. These solutions accommodate both employee needs and operational requirements, ensuring sufficient staffing levels while preventing conflicts in ISP.

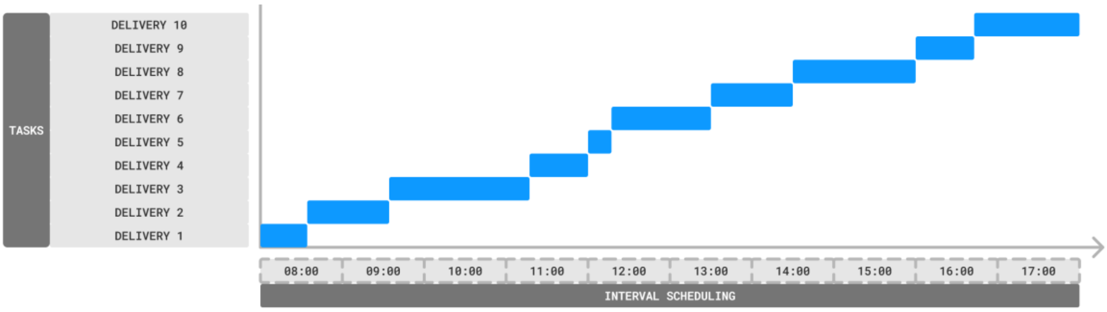


Figure 1 - A solution to an Interval scheduling problem

2.7 Heuristic Methods for ISP Optimization

Given the complexity of scheduling problems, heuristic methods play a critical role in providing efficient solutions. Article [13] demonstrates the effectiveness of heuristics in addressing NP-complete problems, such as those encountered in ISP. By using heuristic techniques, logistics operations can achieve near-optimal scheduling solutions, even in highly dynamic environments, ensuring efficient resource allocation and minimizing scheduling conflicts.

The literature review has outlined the primary challenges and existing solutions in applying ISP to logistics, including uncertainty, mathematical modeling, real-time decision-making, heterogeneous entities, and heuristic methods. Building on these findings, this paper will focus on developing a mathematical model that addresses key constraints—such as timing, resource capacity, and availability—applied to the logistics component at the entry point of a node within a transportation network. This model aims to optimize internal logistics operations, ensuring efficient resource allocation and scheduling.

3 Mathematical Model Overview and Contribution

3.1 Key Contribution and Innovation

The primary contribution of this paper is the development of a new, custom mathematical model that addresses specific constraints related to logistics within large-scale transportation networks. While existing studies on the ISP have addressed certain constraints, the novelty of this model lies in its ability to serve as a reference framework for managing the inbound flow of goods at network nodes, considering real-time constraints such as working hours, team availability, holidays, and overlapping operations.

This model is particularly useful for second-mile or middle-mile logistics operations, such as transporting goods from logistics centres to local warehouses and can be adapted for use by both private and public entities, including retail companies and courier services. A key feature of the proposed framework is its ability to dynamically assess the availability of a node based on the constraints, assigning time slots that ensure efficient processing of goods.

3.2 Future Application and Hybrid Integration

Beyond its immediate utility, this model offers significant potential as a reference for future studies and hybrid implementations. Specifically, it can be integrated with dynamic routing methods, such as those used to solve TSP. By combining the scheduling framework with routing algorithms, future systems can construct dynamic transportation routes based on real-time node availability, evaluated for each specific time slot. This integration would enable logistics operators to select optimal routes that consider not only the shortest or most cost-effective path but also the operational readiness of each node at a given moment in time.

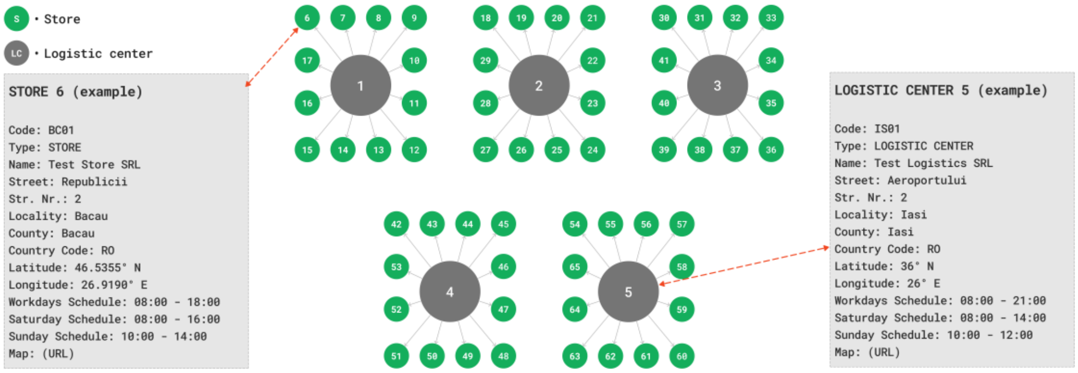
In this way, the proposed model paves the way for hybrid scheduling and routing solutions that can improve the overall efficiency of transportation networks, allowing for more intelligent decision-making and resource allocation in logistics operations.

3.3 Mathematical Model

3.3.1 Definition of the transportation network

To address the complexities of logistics scheduling within transportation networks, we must first define the transportation network itself and structure it as a graph. In this graph, nodes represent logistical points such as warehouses, distribution centres, or retail stores, while edges depict the connections or routes between these nodes (Fig. 2). For the ISP, as applied to logistics units within this network, we will construct concepts such as time slots, which refer to the flow of logistic arrivals at these points, considering various time, capacity, and resource constraints. We will analyse the problem by considering diverse constraints, detailed further, and develop mathematical models that we will use to build our solution.

When considering these constraints, it is crucial to consider a wide array of heterogeneous factors that influence operational efficiency. These factors include varied working hours depending on the logistics node, differing team sizes and availability, and distinct types of merchandise, each requiring customized processing times. Additionally, operator preferences for scheduling arrivals at logistics points at different hours or frequencies, rather than in continuous succession, must be accounted for. Our proposed solution will integrate these diverse aspects and will be applied to a large-scale transportation network.



STORE 6 (example)

Code: BC01
Type: STORE
Name: Test Store SRL
Street: Republicii
Str. Nr.: 2
Locality: Bacau
County: Bacau
Country Code: RO
Latitude: 46.5355° N
Longitude: 26.9190° E
Workdays Schedule: 08:00 - 18:00
Saturday Schedule: 08:00 - 16:00
Sunday Schedule: 10:00 - 14:00
Map: (URL)

LOGISTIC CENTER 5 (example)

Code: IS01
Type: LOGISTIC CENTER
Name: Test Logistics SRL
Street: Aeroportului
Str. Nr.: 2
Locality: Iasi
County: Iasi
Country Code: RO
Latitude: 36° N
Longitude: 26° E
Workdays Schedule: 08:00 - 21:00
Saturday Schedule: 08:00 - 14:00
Sunday Schedule: 10:00 - 12:00
Map: (URL)

Figure 2 - Representation of Transportation network

To enumerate all the constraints identified for our problem, they are as follows:

- Availability only between Logistic Point Working Hours
- Existence or absence of team holidays or national holidays set by the government
- Overlapping with other operations or deliveries already scheduled in the same slots.
- Type of merchandise involved in the transportation process (palletized, non-palletized, mixed, oversized, etc.), and different processing times for each
- Continuous system verification to ensure that time slots do not become invalid by transitioning into the past
- Real-time calculation of availability for scheduling

By addressing these factors through advanced mathematical models, we aim to develop a robust scheduling solution that enhances the overall performance of transportation networks, ensuring that goods are handled efficiently, and resources are optimally utilized.

To define the transportation network for the interval scheduling problem applied to logistics, we consider a network consisting of a total of a work points (nodes), where b represents the number of logistics centres and (a – b) represents the number of stores, forming a complete graph. This means each node is connected to every other node, allowing for logistics arrivals from one point to another. Mathematically, the transportation network is represented as a graph G = (V, E), where:

- V is the set of vertices (nodes), representing the logistics centres and stores.
- E is the set of edges (connections) between the nodes

$V = \{v_1, v_2, \dots, v_a\}$ where $v_1, v_2, \dots, v_b$ represent the logistics centers and $v_{b+1}, v_{b+2}, \dots, v_a$ represent the stores.

The set of edges E is defined as: $E = \{(v_i, v_j) \mid v_i, v_j \in V, i \neq j\}$, this means that for any two distinct nodes v_i and v_j: $(v_i, v_j) \in E$ and $(v_j, v_i) \in E$. Network Properties:

- The number of nodes (vertices) in the graph G is $|V| = a$
- The number of edges (connections) in the complete graph G is: $|E| = a \times (a - 1)$.

By defining the transportation network this way, we ensure a fully connected structure that supports robust and efficient scheduling solutions. This comprehensive framework allows us to address the interval scheduling problem in logistics effectively, accommodating the diverse and interconnected nature of logistics operations.

3.3.2 Definition of Logistics Scheduling at a Point

To define logistics scheduling mathematically, we represent it as a set named SCHEDULE with specific properties. A logistics schedule at a point involves the allocation of time slots for the arrival and departure of goods, considering various constraints such as working hours, equipment availability, and processing times for different types of merchandise. We define the logistics SCHEDULE as follows:

$$\text{SCHEDULE} = \{(O, D, T_s, T_e, Ct, St) \mid O \in V_{centers}, D \in V_{stores}, (v_O, v_D) \in E, T_s, T_e \in T, Ct \in C, St \in \text{STATUSES}\} \quad (1)$$

, where:

- O is the origin point (logistics center) from the subset $V_{centers} \subset V$ where $V_{centers} = \{v_1, v_2, \dots, v_b\}$

- D is the destination point (store) from the subset $V_{stores} \subset V$ where $V_{stores} = \{v_{b+1}, v_{b+2}, \ldots, v_a\}$
- T_s is the start time of the logistics schedule, from the set of time slots T
- T_e is the end time of the logistics schedule, from the set of time slots T
- Ct is the cargo type from the set of cargo types C
- St is the schedule status from the types of STATUSES = {“New”, “Processing”, “Canceled”, “Completed” etc.}

3.3.3 Dependencies (indirect constraints)

3.3.3.1 Dependency on Cargo Type

The logistics SCHEDULE depends on the cargo type Ct, which influences the processing time. Different types of merchandise require different handling times, affecting the start and end times of the logistics schedule. $Ct \in C = \{C_1, C_2, \ldots, C_m\}$, where each C_i represents a different type of cargo (e.g., palletized, non-palletized, mixed, oversized) with specific processing times τ (C_i).

$$\tau(Ct) = T_e - T_s \qquad (2)$$

3.3.3.2 Dependency on Inventory Source Teams

The logistics SCHEDULE also depends on the availability of teams at inventory sources. Let Lt be the set of all teams at logistics nodes V. $Lt = \{Lt_1, Lt_2, \ldots, Lt_n\}$, where each Lt_j is associated with a specific node $v \in V$ and has its own availability. It’s necessary to have at least one team defined for the destination node (point) ($\forall\ v_j \in V$, $\exists\ Lt_k$, with Lt_k (node) = v_j), and not be on holiday. For checking an existing team availability, we assume that team is available every day, except when there are team holidays. Team holidays are defined as a set of associations between teams and specific days intervals:

$LtH = \{(LtH_k, x_l) \mid LtH_k \in LtH, x_l \in X\}$, where x_l represents a team holiday period
X = {(startTime, endTime) | startTime, endTime ∈ T (time slots)
, endTime > startTime} (3)

3.3.3.3 National Holidays

National government holidays can also impact the scheduling process. Let H be the set of national holidays for a given year. $H = \{h_1, h_2, \ldots, h_m\}$. Similarly, as team holidays, we define a set of associations between national holidays and specific days intervals:

$H = \{(H_n, x_o) \mid H_n \in H, x_o \in X\}$, where x_o represents a national holiday period (4)

3.3.4 Logistic constraints

Availability only between Logistic Point Working Hours:
$(T_s, T_e) \cap (Lwh_s, Lwh_e)$!= $\emptyset$, $\forall\ T_s, T_e \in T$
, where $Lwh_s, Lwh_e \in$ {hours} represents the (start and end) working hours
(for a specific weekday) of (logistic) point Lt_j (5)
Overlapping with Other Unfinished Operations:
$(T_s, T_e) \cap (T_{s'}, T_{e'}) = \emptyset$

$, \forall(T_s, T_e), (T_{s'}, T_{e'}) \in T$ and $\forall\ St, St' \in STATUSES \setminus \{$“Finished”, “Canceled”$\}$ (6)

Continuous System Verification:

$T_s \geq$ current time, $\forall\ T_s \in T$ (7)

Real-Time Calculation of Availability:

$A\ (T_s, T_e) \in \{true, false\}$, availability status is based on real-time data (8)

3.3.5 Comprehensive Scheduling Equation

Finally, we define the comprehensive scheduling equation, incorporating all constraints ((1), (2), (3), (4), (5), (6), (7), (8)) for a specific logistics SCHEDULE:

$SCHEDULE = \{(O, D, T_s, T_e, Ct, St) \mid (v_o, v_d) \in E, (T_s, T_e) \cap (Lwh_s, Lwh_e) \neq \emptyset$
$, (T_s, T_e) \cap H = \emptyset$, count$(Lt) \geq 1$, with $Lt(node) = D, (T_s, T_e) \cap LtH = \emptyset$
$, (T_s, T_e) \cap (T_{s'}, T_{e'}) = \emptyset, \tau(Ct) = T_e - T_s, T_s \geq$ current time, $A(T_s, T_e) = true$
$, \forall\ O \in V_{centers} \subset V$ where $V_{centers} = \{v_1, v_2, \ldots, v_b\}, \forall\ D \in V_{stores} \subset V$
, where $V_{stores} = \{v_{b+1}, v_{b+2}, \ldots, v_a\}, \forall(T_s, T_e), (T_{s'}, T_{e'}) \in T, T_e > T_s, T_{e'} > T_{s'}$
$, \forall\ Ct \in C, \forall\ St, St' \in STATUSES \setminus \{$“Finished”, “Canceled”$\}\}$. (9)

This comprehensive mathematical framework (represented visually in Fig. 3) ensures that logistics scheduling is efficient, considering all necessary constraints and dependencies.

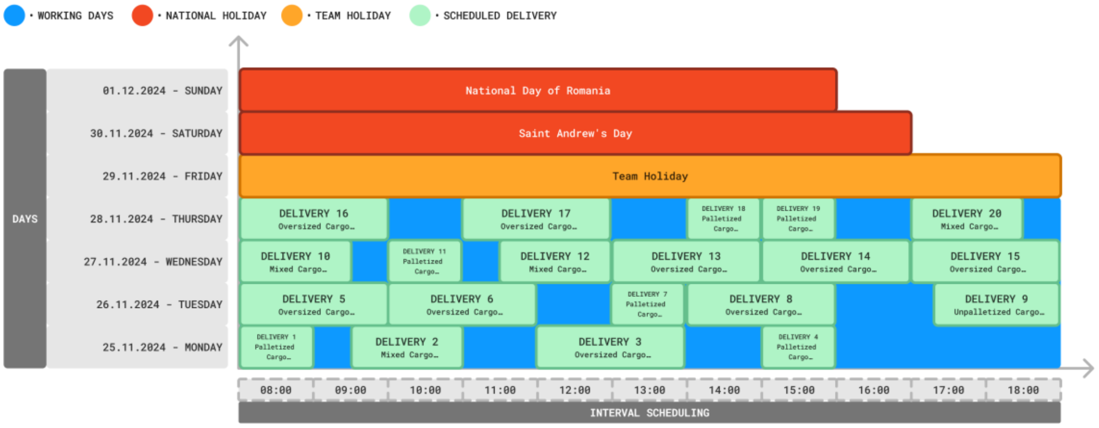


Figure 3 - Visual representation of proposed framework for restricted Interval Scheduling Problem

4 Implementation of the Proposed Solution

In this section, we will discuss the implementation of the solution for the interval scheduling problem described mathematically in the previous section. The application we have developed leverages well-known and widely used frameworks to ensure robustness and efficiency. The back-end processing is managed using PHP Symfony, a powerful framework that facilitates scalable and maintainable web applications. For data storage, we utilize a well-structured MySQL database, ensuring reliable and efficient data management. At the front-end, our application employs a variety of technologies to enhance the user interface and user experience. These include Node.js, JavaScript, jQuery, AJAX, Twig, Bulma and CSS. These technologies collectively contribute to a responsive and interactive user interface. To ensure portability and streamline development and deployment processes, the application architecture is containerized using Docker. This approach not only enhances portability but also simplifies the management of dependencies and the deployment process across different environments. The application's functionality relies on the seamless communication between the frontend and backend via RESTful APIs. This architecture allows for efficient data exchange and ensures that the application can dynamically respond to user interactions and real-time updates. By integrating these advanced frameworks and technologies, our solution provides a comprehensive and effective tool for optimizing appointment scheduling and resource allocation in transportation networks, ultimately enhancing operational efficiency and reducing costs.

The database has been populated with all streets, localities, and counties in Romania using a public governmental data source [20]. Subsequently, we constructed a transportation network that spans the entire country, encompassing 5 regional logistics centres and 55 stores, each dynamically linked to an address. The transportation network is inspired by the distribution network of Dedeman stores [21]. Therefore, the variables from the mathematical model Section, have the following values: a = 60, b = 5. Major components of the applications are presented in Figure 4.

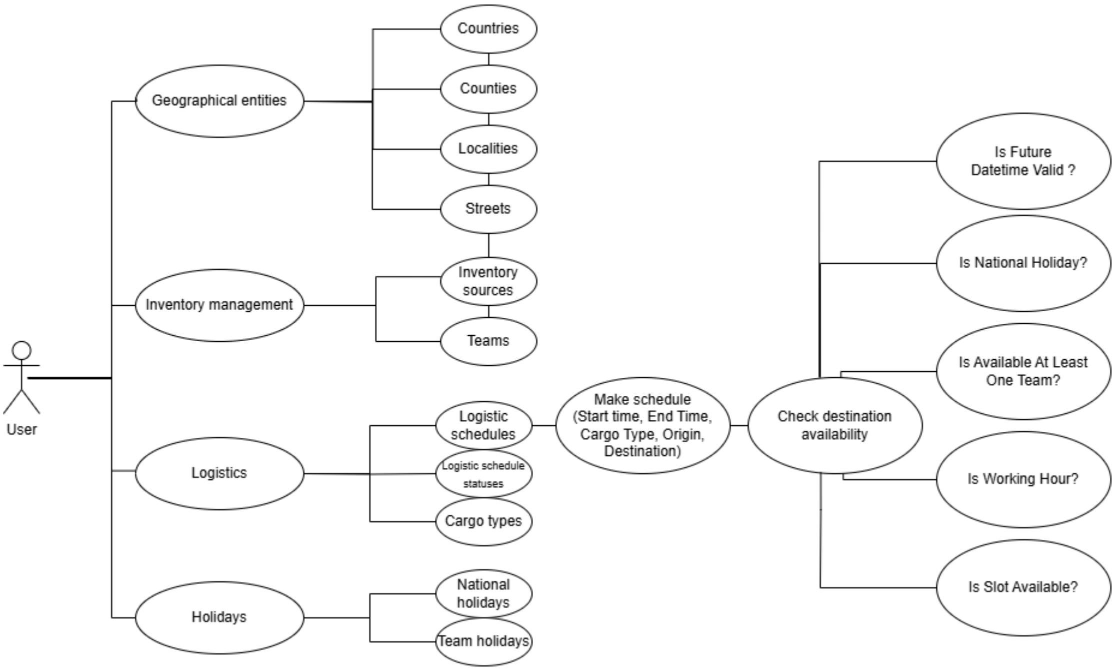


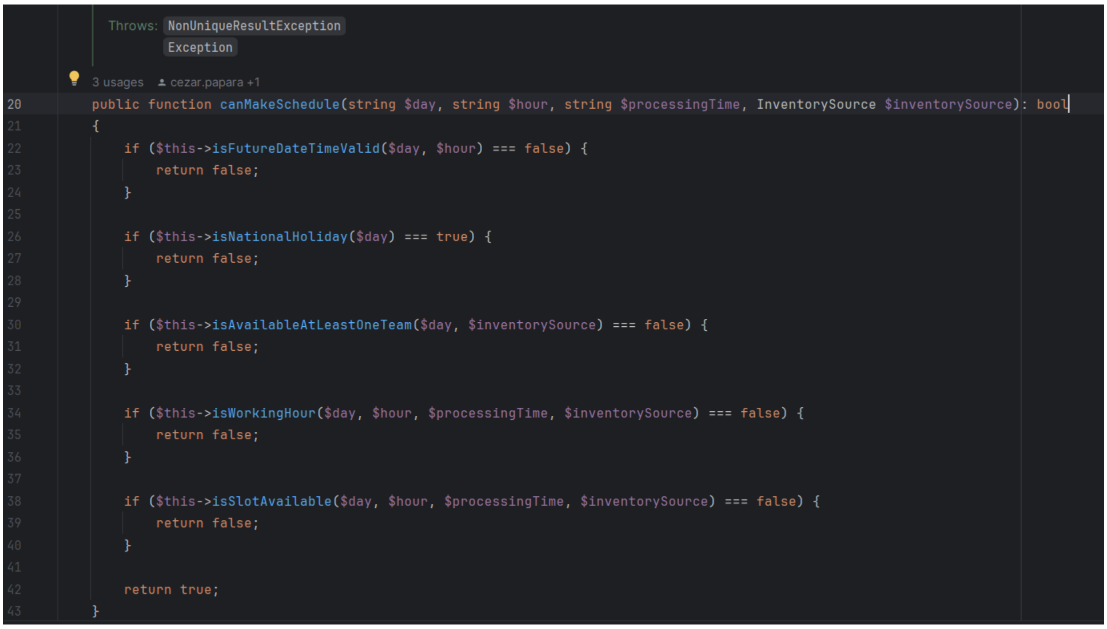
Figure 4 - Major components of proposed framework

The relationships between entities in the database are established using foreign keys to ensure dynamic connections and to enforce constraints that allow only valid values to be inserted. We have defined the LogisticSchedule entity, which serves as the main implementation of our problem. This entity is linked to the Status entity, as schedules can be in various processing states. It is also connected to the CargoType entity, where custom processing times are defined for each type of cargo. The LogisticSchedule entity is further related to the InventorySource entity, which defines the working schedule for weekdays and weekends, and is associated with teams (InventorySourceTeam) that may have specified holidays (TeamHoliday). Moreover, there is an entity for national holidays provided by the government. All these relationships and entities in the database will help us validate when a new logistic schedule can be recorded in a specific slot, that is, for a certain logistics point, on a particular day, at a specific hour, and for a defined type of cargo.

As discussed, the core implementation of our solution revolves around the LogisticSchedule entity, which will store all scheduled plans for a logistics point, only after all validations defined mathematically in the previous section have been verified. The proposed solution is based on a validation function that checks the availability status for scheduling a logistic appointment within a certain time interval, on a specific day, at a particular logistics point (the destination), which has assigned teams that may or may not have holidays.

These teams can also be affected by national holidays. The scheduling is influenced by the type of cargo (considering that each specific type of cargo has different processing times and thus non-uniform slot occupancy), as well as the existence of other schedules with the condition that they are not in the final status. Additionally, it is impacted by the working hours of the store and the validity of the time moment.

As mentioned, we have developed a function that validates the eligibility of creating a new schedule. This function is depicted in Figure 5. It returns a Boolean value. This function in turn calls five different singular functions, each of which also returns a Boolean value. We will present each of these functions immediately to highlight their significance and importance. If, in the end, each of these functions returns an appropriate result, then our main function will return true, indicating that we can record the schedule for the specified parameters at this time.



```
Throws: NonUniqueResultException
        Exception

3 usages   cezar.papara +1
public function canMakeSchedule(string $day, string $hour, string $processingTime, InventorySource $inventorySource): bool
{
    if ($this->isFutureDateTimeValid($day, $hour) === false) {
        return false;
    }

    if ($this->isNationalHoliday($day) === true) {
        return false;
    }

    if ($this->isAvailableAtLeastOneTeam($day, $inventorySource) === false) {
        return false;
    }

    if ($this->isWorkingHour($day, $hour, $processingTime, $inventorySource) === false) {
        return false;
    }

    if ($this->isSlotAvailable($day, $hour, $processingTime, $inventorySource) === false) {
        return false;
    }

    return true;
}
```

Figure 5 - Validation function for presented solution of Logistic Scheduling Interval Problem

The **Function isFutureDateTimeValid** ensures that the scheduled date and time for the logistics appointment are set in the future. It validates that the provided date and time are not in the past by comparing them with the current date and time. This is essential to avoid scheduling appointments for times that have already passed.

The **Function isNationalHoliday** checks if the proposed scheduling date falls on a national holiday. This validation is crucial because national holidays typically affect the availability of logistics services. The function queries the database to determine if the date lies within the range of any national holiday period. If it does, scheduling on that day is not permitted.

The **Function isAvailableAtLeastOneTeam** ensures that at least one logistics team is available on the specified date at the given logistics point. It checks the availability of teams by considering their assigned holidays and other constraints. This function is fundamental to guarantee that there is sufficient human resource capacity to handle the logistics operations on the chosen date.

The **Function isWorkingHour** verifies that the scheduling falls within the working hours of the specified logistics point. It retrieves the working schedule for the day (weekdays or weekends) and confirms that the logistics appointment can be accommodated within these hours. This ensures that the logistics operations are planned within the operational hours, avoiding conflicts with non-working periods.

The **Function isSlotAvailable** checks the availability of the time slot for the new logistics appointment. It evaluates whether the proposed time slot overlaps with any existing schedules that are still active and not finalized. This function is crucial for preventing double-booking and ensuring that each time slot is dedicated to a single logistics operation, maintaining an organized and conflict-free scheduling system.

By systematically applying these validation functions, the canMakeSchedule method ensures that all logistical appointments adhere to defined constraints and are feasible within the operational framework. Each function plays a vital role in maintaining the integrity and efficiency of the scheduling system.

5 Results Assessment

This section evaluates the performance and effectiveness of our application in solving logistics scheduling problems. We assess how well the solution adheres to defined constraints and improves the efficiency and reliability of the transportation network.

Validation of Theoretical Models: We validated our theoretical models against real-world logistics scenarios by constructing a graph representing logistics centers and stores, incorporating constraints such as working hours, team availability, and cargo types. This approach, using real data, confirmed the accuracy and practicality of our models.

Performance of the Scheduling Algorithm: The scheduling algorithm was tested across various logistics scenarios. Key performance metrics include:

Accuracy: The algorithm consistently generated valid schedules, respecting all constraints.

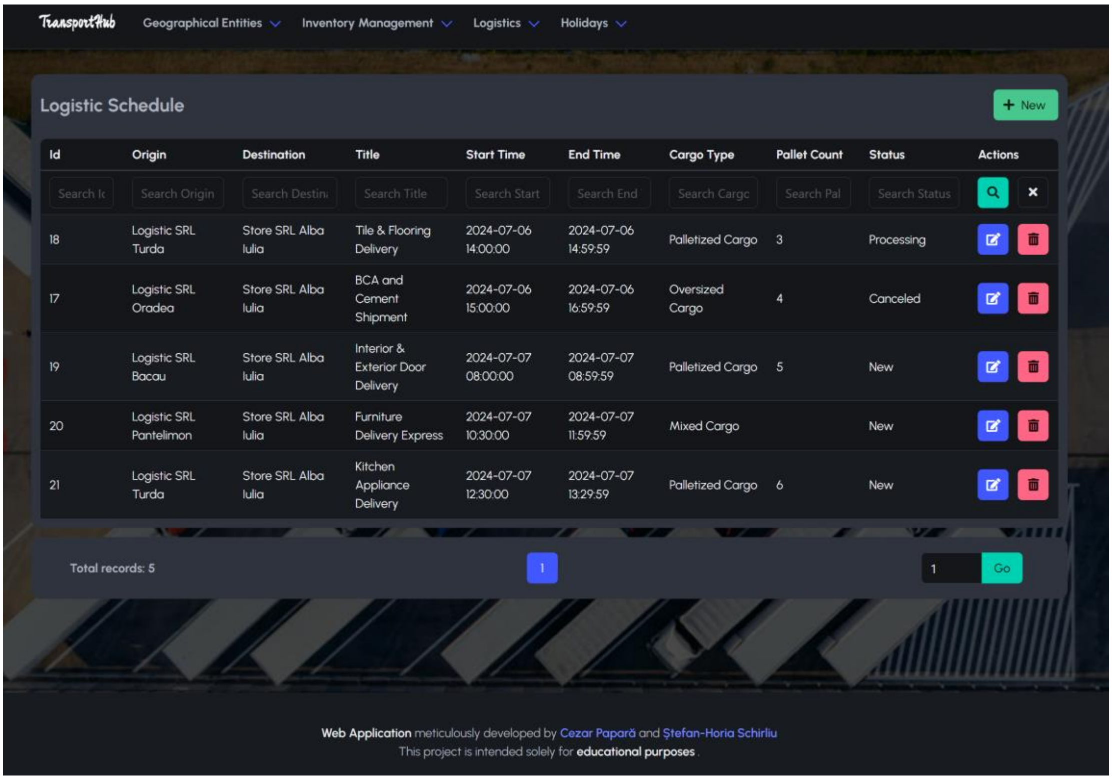
Efficiency: It handled numerous scheduling requests promptly, suitable for high-demand environments.

Scalability and Performance Analysis: The application demonstrated exceptional scalability and performance, maintaining reliability even as the complexity of the logistics network increased. Initially tested on a transport network with 60 nodes, the system was further validated experimentally with up to 500 nodes, representing logistics centers and stores across Romania. Under these expanded conditions, it continued to perform real-time validation for creating new schedules while accounting for existing ones.

Performance metrics confirm the system's efficiency, with response times consistently under 10 seconds, even when handling up to 50 simultaneous user requests generated using automated testing tools. Frontend AJAX requests ensured seamless real-time validation, effectively preventing invalid data entries.

Practical Applicability: Our solution was assessed for practical applicability in real-world logistics operations. It effectively integrated diverse constraints, including team holidays and national holidays, demonstrating adaptability across different regions and industries. Real-time validation functions ensured schedules remained feasible and up to date.

Analysis with Numerical Data: To illustrate, we consider the existing schedules for destination Store SRL Alba Iulia from Figure 6.



Logistic Schedule

Id	Origin	Destination	Title	Start Time	End Time	Cargo Type	Pallet Count	Status
18	Logistic SRL Turda	Store SRL Alba Iulia	Tile & Flooring Delivery	2024-07-06 14:00:00	2024-07-06 14:59:59	Palletized Cargo	3	Processing
17	Logistic SRL Oradea	Store SRL Alba Iulia	BCA and Cement Shipment	2024-07-06 15:00:00	2024-07-06 16:59:59	Oversized Cargo	4	Canceled
19	Logistic SRL Bacau	Store SRL Alba Iulia	Interior & Exterior Door Delivery	2024-07-07 08:00:00	2024-07-07 08:59:59	Palletized Cargo	5	New
20	Logistic SRL Pantelimon	Store SRL Alba Iulia	Furniture Delivery Express	2024-07-07 10:30:00	2024-07-07 11:59:59	Mixed Cargo		New
21	Logistic SRL Turda	Store SRL Alba Iulia	Kitchen Appliance Delivery	2024-07-07 12:30:00	2024-07-07 13:29:59	Palletized Cargo	6	New

Total records: 5

Web Application meticulously developed by Cezar Papară and Ștefan-Horia Schirliu
This project is intended solely for educational purposes .

Figure 6 - Example of existing schedules for specific Logistic Point

On July 6, 2024, at 13:55:00, we attempted to schedule deliveries to destination Store SRL Alba Iulia with various start times on different days and hours, considering the cargo type as Oversized Cargo, which requires a processing time of 2 hours. The table below presents the scheduling attempts and their outcomes. In the column "Available Time Slot" we noted the system's response: "TRUE" if the scheduling was possible and "FALSE" otherwise, accompanied by the reason for the invalid scheduling.

The operating schedule for the Alba Iulia store is Monday to Friday from 08:00 to 18:00, Saturday from 08:00 to 17:00, and Sunday from 08:00 to 16:00. Additionally, according to the national holidays table, November 30, 2024, is Saint Andrew's Day. Furthermore, a vacation period for the Alba Iulia team has been set from August 15 to August 20, 2024. Table 1 presents the attempts and outcomes.

Table 1 - Validation of different schedule attempts

Nr.	Start Time	End Time	Cargo Type	Available Time Slot	Reason
1	2024-07-06 08:00:00	2024-07-06 09:59:59	Oversized Cargo	FALSE	Invalid moment (past moment)
2	2024-07-06 13:55:00	2024-07-06 15:54:59	Oversized Cargo	FALSE	Invalid moment (past moment)
3	2024-07-06 14:00:00	2024-07-06 15:59:59	Oversized Cargo	FALSE	Slot unavailable, overlapping with another active schedule
4	2024-07-06 15:00:00	2024-07-06 16:59:59	Oversized Cargo	TRUE	-
5	2024-07-06 16:00:00	2024-07-06 17:59:59	Oversized Cargo	FALSE	Slot unavailable, outside working hours
6	2024-07-07 08:30:00	2024-07-07 10:29:59	Oversized Cargo	FALSE	Slot unavailable, overlapping with another active schedule
7	2024-07-07 13:30:00	2024-07-07 15:29:59	Oversized Cargo	TRUE	-
8	2024-07-07 13:45:00	2024-07-07 15:44:59	Oversized Cargo	TRUE	-
9	2024-07-07 14:00:00	2024-07-07 15:59:59	Oversized Cargo	TRUE	-
10	2024-07-06 14:15:00	2024-07-06 16:14:59	Oversized Cargo	FALSE	Slot unavailable, end time will be outside working hours
11	Any time between 2024-08-15 and 2024-08-20	Any start time + processing time	Oversized Cargo	FALSE	Team is on holiday
12	2024-11-30 (any time)	Any start time + processing time	Oversized Cargo	FALSE	National holiday (Saint Andrew's Day)

6 Discussion: Comparison Between Our Method and TSP with Time Windows

Advantages of Our Interval Scheduling Problem (ISP) Approach with Custom Constraints:

- **Dynamic Constraints Handling**: Our method efficiently integrates various real-world constraints, such as working hours, holidays, overlapping operations, and merchandise types. These are highly relevant to logistics nodes and are checked in real-time, ensuring that the availability of time slots is accurately reflected.
- **Customized to Logistic Needs**: By considering node-specific factors, such as team holidays and merchandise types, our solution is more tailored to the internal operations of logistic points compared to generic TSP with time windows. This allows for more precise planning and resource allocation within nodes.

- **Real-Time System Monitoring**: Our approach continuously verifies that time slots remain valid, preventing issues caused by scheduling into past time slots. This real-time verification offers a higher level of accuracy and responsiveness than traditional TSP models.

Disadvantages or Challenges:

- **Complexity**: The introduction of multiple custom constraints increases computational complexity, especially when compared to TSP with time windows, which typically only accounts for a few constraints, such as travel times and delivery windows.
- **Scalability**: Handling dynamic constraints in real-time, such as continuously monitoring holiday schedules or overlapping operations, may reduce scalability for larger networks. TSP with time windows is a more established solution that tends to scale better with larger networks due to its simplicity.

Comparison with TSP with Time Windows:

- **Flexibility**: While TSP with time windows primarily focuses on optimizing routes and delivery schedules within fixed windows, it does not account for the internal dynamics of logistics nodes, such as staff availability or different processing times for goods. Our ISP-based approach is more flexible in adapting to these internal constraints.
- **Real-Time Decision Making**: TSP with time windows often works with pre-set schedules, whereas our method incorporates real-time availability checks, offering more dynamic decision-making capabilities. This can improve efficiency in unpredictable environments but comes at the cost of increased system complexity.
- **Scope**: TSP with time windows is primarily focused on external routing optimization. In contrast, our approach shifts focus to internal node operations, making it more suitable for environments where bottlenecks within nodes, like warehouses, are a significant factor in overall performance.

7 Conclusions

In conclusion, our analysis of the logistic interval scheduling problem highlights that the more detailed and real-world constraints we incorporate into our transportation network, the more complex the problem becomes. However, this increased complexity results in more adaptable and realistic solutions. By integrating custom constraints such as working hours, holidays, overlapping operations, and merchandise-specific processing times, our method creates solutions that are not only theoretically sound but also highly practical and robust in real-world logistics.

Comparing our approach to the traditional Traveling Salesman Problem (TSP) with time windows, we observe that while TSP provides simpler, scalable solutions focused on external routing, it does not address the internal logistical complexities within network nodes. Our Interval Scheduling Problem (ISP)-based approach, on the other hand, captures these internal dynamics and offers real-time adaptability, though at the cost of increased computational overhead.

Looking ahead, we plan to explore further the connections between TSP, Vehicle Routing Problem (VRP), and Logistic Interval Scheduling. Investigating how logistical constraints—such as capacity, resource availability, and time limitations—affect the selection of optimal routes will be central to our future research. We aim to develop heuristic solutions that account for all relevant constraints and variables, ensuring these theoretical ideas can be applied in practice. This work will focus on addressing the inherent uncertainty and complexity in transportation networks, with the goal of creating dynamic and scalable solutions for modern logistics.

Acknowledgment

The research performed Ștefan-Horia Schirliu was funded by the Ministry of Education, through the National Council for the Financing of Higher Education, Romania, grant number CNFIS-FDI-2024-F-0492: Integrated actions to increase the quality of the research process at "Vasile Alecsandri" University of Bacau (acronym: UBc4Research).

References

[1] Wang, Yu & Liu, Haoxiang & Fan, Yinchao & Ding, Jianxun & Long, Jiancheng. *Large-scale multimodal transportation network models and algorithms-Part II: Network capacity and network design problem*, Transportation Research Part E: Logistics and Transportation Review, 167, 102918, Pages 1-7, 2022, 10.1016/j.tre.2022.102918

[2] Lumír Pečený, Pavol Meško, Rudolf Kampf, Jozef Gašparík. *Optimisation in Transport and Logistic Processes*, Transportation Research Procedia, Volume 44, Pages 15-22, ISSN 2352-1465, 2020

[3] James R. Jackson. *Scheduling a Production Line To Minimize Maximum Tardiness*, Research Report, Issue 43, Pages 1-72, Management Science Research Project, University of California, Los Angeles, 1955

[4] Min Ji, Ji Fang, Wenya Zhang, Lijuan Liao, T.C.E. Cheng, Yuanyuan Tan. *Logistics scheduling to minimize the sum of total weighted inventory cost and transport cost*, Computers & Industrial Engineering, Volume 120, Pages 206-215, ISSN 0360-8352, 2018

[5] Mallek, Amin & Boudhar, Mourad. *Scheduling on uniform machines with a conflict graph: complexity and resolution*, International Transactions in Operational Research. 31, Pages 1-4, 2024, 10.1111/itor.13170

[6] Allahverd, Ali. *A survey of scheduling problems with uncertain interval/bounded processing/setup times*, Journal of Project Management, Volume 7, Pages 255-264, 2022, 10.5267/j.jpm.2022.3.003

[7] Niraj Ramesh, D., Krishnamoorthy, M., Ernst, A.T. *Efficient models, formulations and algorithms for some variants of fixed interval scheduling problems*, in: Sarker, R., Abbass, H.A., Dunstall, S., Kilby, P., Davis, R., Young, L. (eds.) Data and Decision Sciences in Action. LNMIE, Pages 43–69, Springer, Cham, 2018, https://doi.org/10.1007/978-3-319-55914-8_4

[8] Boyar, J., Favrholdt, L.M., Kamali, S., Larsen, K.S. *Online interval scheduling with predictions*, in: 18th International symposium on algorithms and data structures (WADS), Lecture Notes in Computer Science, Volume 14079, Pages 193–207, Springer, Berlin, Heidelberg, 2023

[9] Xu, Xinping & Zhang, Jingwen & Li, Minming & Duan, Lingjie & Xie, Lihua. *A Family of Strategyproof Mechanisms for Activity Scheduling* *, Autonomous Agents and Multi-Agent Systems, Volume 37, 2, Pages 1-8, 2023, 10.1007/s10458-023-09624-7

[10]Zhou, Haohao & Bai, Guanghan & Deng, Su. *Optimal interval scheduling with nonidentical given machines*, Cluster Computing, Volume 22, 5, Pages 1-6, 2019, 10.1007/s10586-018-02892-z

[11]Gogos, Christos. *Solving the Distributed Permutation Flow-Shop Scheduling Problem Using Constrained Programming*, Applied Sciences, Volume 13, 12562, Pages 1-26, 2023, 10.3390/app132312562

[12]Kolker, Alexander. *Staffing and Scheduling Problems with Sliding Days Off: Solutions Using Integer Linear Optimization*, in book: Healthcare Management Engineering In Action, Pages 1-15, 2024, 10.1007/978-3-031-53663-2_6

[13]A. Sri Harsha, P. Vijaya Kumar, Arumalla Nagaraju, K. Prakash Babu, N. R. Medikondu, B. V. Dharmendra. *Mining knowledge for NP-complete scheduling problems*, AIP Conference Proceedings, At: Nandyal City, Volume 2821 (1), Pages 1-5, 070014, 2023, https://doi.org/10.1063/5.0158474

[14]G. Reinelt. *The Traveling Salesman Problem: Computational Solution for TSP Applications*, LNCS, Springer-Verlag, vol. 840, Pages 4-30, 1994

[15]G. B. Dantzig and J. H. Ramser. *The Truck Dispatching Problem*, Management Science, Vol. 6, No. 1, Pages 80-91, 1959

[16]Cezar-Marian Papară. *An empirical study of large transportation networks and solutions for the cost optimization*, International Journal of Advanced Statistics and IT&C for Economics and Life Sciences, Vol. XII, no. 2, Pages 41-52, 2022

[17]Cezar-Marian Papară. *Enhancing Education Quality in Rural Settings through Steiner-Based Societal Partitioning Optimization*, Journal of Innovation in Psychology, Education and Didactics ISSN: 2247-4579, E-ISSN: 2392-7127, 2023, Volume 27, Issue 2, Pages: 241-248, 2023, doi:10.29081/JIPED.2023.27.2.08

[18]Cezar-Marian Papară. *Digital frontiers in logistics: A scalable approach to wide-area transportation network optimization*, "Vasile Alecsandri" University of Bacău, Faculty of Sciences, Series Mathematics and Informatics, Vol. 33, No 2, Pages 63 – 80, 2024

[19]Cezar-Marian Papară, Alex-Andrei Rîpan. *Strategies for Optimizing Pharmaceutical Logistics: A Focus on Transportation Efficiency*, Eighth International Conference on Applied Informatics, Imagination, Creativity, Design, Development - ICDD, Sibiu, Romania, https://conferences.ulbsibiu.ro/icdd/2024/program.php, 2024

[20][Online] https://data.gov.ro/dataset?tags=nomenclator, [Accessed 01.04.2024]

[21][Online] https://www.dedeman.ro/ro/suport-clienti/magazine-dedeman, [Accessed 03.04.2024]

DOI: 10.2478/ijasitels-2024-0015

Implementation of sensorless indirect vector control of induction motor with closed-loop current control

Gabriela Crăciunaş[1], *Alina Cristina Viorel*[1]

[1]*Computer Science and Electrical Engineering Department, "Lucian Blaga" University of Sibiu, Romania, E-Mail: gabriela.craciunas@ulbsibiu.ro; alina.viorel@ulbsibiu.ro*

Abstract

The paper focuses on the development and testing of an advanced induction motor control method. This method, known as indirect rotor flux-oriented vector control, allows independent control of motor torque and flux, providing high dynamic performance similar to that of a DC motor. Additionally, the implemented system is sensorless, eliminating physical rotor position sensors, which reduces complexity and cost while increasing system reliability. The proposed system uses a rotor flux estimation algorithm, based on mathematical models, which provides the necessary information for vector orientation in the absence of a sensor. Additionally, closed-loop current control improves stability and accuracy of the control by correcting output current deviations from reference values. Thus, the system provides fast and accurate response in the presence of load variations and external disturbances. Simulations and experimental tests were carried out in the Simulink environment, where the proposed model was evaluated under various operating conditions. The results demonstrate the efficiency of the system and the validity of the method for industrial applications where precise and robust induction motor control is required.

Keywords: three-phase induction machine, vector-control, rotor flux estimator

Introduction

Induction motors have been and continue to be widely used in industrial applications due to their robustness, durability and low maintenance compared to other types of electric machines. However, precise control of these motors presents significant challenges, as motor parameters vary with operating conditions and dynamics are complex. Indirect rotor flux oriented vector control has emerged as an effective solution to enable superior dynamic performance comparable to that of DC motors by decoupling torque and flux components, thus providing better motor handling in transient regimes and variable load.

The literature includes numerous studies and researches that explore various methods of implementing vector control for induction motors. The flux-orientation method was originally proposed by Blaschke in the 1970s, and has since been extended and improved in many directions, including sensorless implementations to reduce the complexity and cost of industrial systems. The removal of flux sensors, for example, has become an important research direction due to the potential to simplify the control system and increase its reliability in difficult operating environments. Estimating the rotor flux through mathematical models of the motor and using adaptive algorithms has proven to be an effective method to eliminate flux sensors, with sufficient accuracy for most industrial applications, [1], [2].

Closed-loop current control is also a key element in ensuring the stability and accuracy of the system. This approach allows rapid compensation of variations and disturbances by continuously adjusting the current, thus providing superior performance in induction motor control, [3]. Recent studies show that the integration of closed-loop current control with sensorless vector control methods leads to improved system performance under real operating conditions, [4].

With the help of the Matlab/Simulink platform, it is possible to develop and simulate the control models of induction motors, thanks to its ability to represent the mathematical complexity of the systems and to allow testing of different operating scenarios before physical implementation. This paper aims to study and implement indirect vector control of a flux sensorless induction motor with closed-loop current control using advanced simulations in Simulink to evaluate the performance and robustness of the proposed method.

Mathematical model

The mathematical model of the asynchronous motor is written in stator notation and describes the dynamic behavior of the machine based on electrical and mechanical equations. Typically, the vector representation for currents, voltages and fluxes is used within a system of fixed axes relative to the stator (d,q), which allows voltages and currents to be treated as scalar quantities for each of these two perpendicular axes. Thus, the stator and rotor voltage equations are, [4],

$$\begin{aligned} u_{ds} &= R_s \cdot i_{ds} + \frac{d\psi_{ds}}{dt} \\ u_{qs} &= R_s \cdot i_{qs} + \frac{d\psi_{qs}}{dt} \\ 0 &= R_r \cdot i_{dr} + \frac{d\psi_{dr}}{dt} + \omega_r \cdot \psi_{qr} \\ 0 &= R_r \cdot i_{qr} + \frac{d\psi_{qr}}{dt} - \omega_r \cdot \psi_{dr} \end{aligned} \tag{1}$$

The stator and rotor flux components are,

$$\begin{aligned} \psi_{ds} &= L_s \cdot i_{ds} + L_{sr} \cdot i_{dr} \\ \psi_{qs} &= L_s \cdot i_{qs} + L_{sr} \cdot i_{qr} \\ \psi_{dr} &= L_r \cdot i_{dr} + L_{sr} \cdot i_{ds} \\ \psi_{qr} &= L_r \cdot i_{qr} + L_{sr} \cdot i_{qs} \end{aligned} \tag{2}$$

Where, $U_{ds}, I_{ds}, \psi_{ds}$ - d-axis components,

$U_{qs}, I_{qs}, \psi_{qs}$ - q axis components.

The general equation of motion and the expression of the electromagnetic torque are,

$$\frac{d\omega_r}{dt} = \frac{p}{J}\left(T_e - T_L\right) \tag{3}$$

$$T_e = \left(\frac{3p}{2}\right)\left(\frac{L_m}{L_r}\right)\left(\psi_{rd} i_{sq} - \psi_{rq} i_{sd}\right) \tag{4}$$

Where, T_e, T_L - electromagnetic torque and load torque.

Rotor flux estimator

For the indirect vector control with rotor flux orientation of the asynchronous motor, the proposed model uses a rotor flux estimation algorithm, based on mathematical models derived directly from the machine equations. With such models the partial state estimators are determined. These rotor flux estimation models provide the information needed for vector orientation in the absence of a sensor. For this paper, the "VI" estimator variant was chosen, an estimator that uses the stator voltage (V) and current (I) measurements to approximate the rotor flux. This type of estimator is relatively simple to implement and is used in applications where flux sensors are not available but indirect control of the induction motor is required.

In such situations, the mathematical model is determined from the motor equations written in the state space, [5], [6],

$$\dot{x} = A \cdot x + B \cdot u \tag{5}$$

Where: x – the vector of state variables,

u – system input (control).

If the stator currents and rotor fluxes (6) are chosen as state variables, then equations (1) are modified as follows,

$$x = \begin{bmatrix} i_{sd} & i_{sq} & \psi_{rd} & \psi_{rq} \end{bmatrix}^T \tag{6}$$

$$u = \begin{bmatrix} u_{sd} & u_{sq} \end{bmatrix}^T$$

$$A = \begin{bmatrix} -\left(\frac{1}{\sigma T_s} + \frac{1-\sigma}{\sigma T_r}\right) & 0 & \frac{L_m}{\sigma L_s L_r T_r} & \frac{L_m}{\sigma L_s L_r}\omega \\ 0 & -\left(\frac{1}{\sigma T_s} + \frac{1-\sigma}{\sigma T_r}\right) & -\frac{L_m}{\sigma L_s L_r}\omega & \frac{L_m}{\sigma L_s L_r T_r} \\ \frac{L_m}{T_r} & 0 & -\frac{1}{T_r} & -\omega \\ 0 & \frac{L_m}{T_r} & \omega & -\frac{1}{T_r} \end{bmatrix} \tag{7}$$

$$B = \begin{bmatrix} \frac{1}{\sigma L_s} & 0 \\ 0 & \frac{1}{\sigma L_s} \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \tag{8}$$

Where, $T_s = \frac{L_s}{R_s}$ - stator time constant,

$T_r = \frac{L_r}{R_r}$ - rotor time constant,

$\sigma = 1 - \frac{L_m^2}{L_s L_r}$ - leakage coefficient.

Following the performed calculations, the mathematical model of the rotor flux estimator "VI" becomes,

$$\frac{d\hat{\Psi}_{rd}}{dt} = \frac{L_r}{L_m} u_{sd} - \frac{L_r}{L_m} R_s i_{sd} + \frac{L_m^2 - L_s L_r}{L_m} \frac{di_{sd}}{dt} \tag{9}$$

$$\frac{d\hat{\Psi}_{rq}}{dt} = \frac{L_r}{L_m} u_{sq} - \frac{L_r}{L_m} R_s i_{sq} + \frac{L_m^2 - L_s L_r}{L_m} \frac{di_{sq}}{dt} \tag{10}$$

Where, $\hat{\Psi}_{rd}$ and $\hat{\Psi}_{rq}$ - d and q-axis estimated rotor linkage fluxes.

From the mathematical model (9) and (10) we can see the dependence of the "VI" estimator on all the main parameters of the asynchronous motor, except for the rotor resistance. However, the involvement of the stator resistance in the mathematical model of the flux estimator may lead to errors in the estimation of the rotor flux. Being a temperature-dependent parameter, correction filters are used in practical applications to compensate for errors in the estimation of the rotor flux. Fig.1 shows the "VI" estimator, where the stator currents are in the derivative form in relation to time and also the internal reaction term is missing.

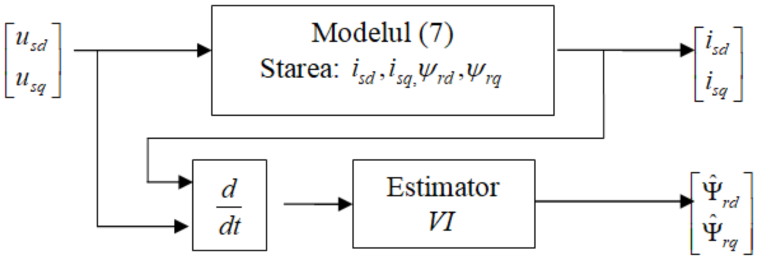

$\begin{bmatrix} u_{sd} \\ u_{sq} \end{bmatrix}$
Modelul (7)
Starea: $i_{sd}, i_{sq}, \psi_{rd}, \psi_{rq}$
$\begin{bmatrix} i_{sd} \\ i_{sq} \end{bmatrix}$
$\frac{d}{dt}$
Estimator VI
$\begin{bmatrix} \hat{\Psi}_{rd} \\ \hat{\Psi}_{rq} \end{bmatrix}$

Figure 1. “VI” Estimator

Simulation results

The purpose of testing in the Matlab/Simulink simulation environment is to evaluate the performance and robustness of an advanced control method for induction motor. Specifically, the aim is to validate the operation of the indirect vector control, with rotor flux orientation, implemented in sensorless mode, with closed circuit current control, [7].

The objectives of the simulation are: checking the stability of the flux estimation in various operating regimes, including load and speed variations; closed loop current control validation; the dynamic response of the induction motor to speed and load control under different operating conditions. All these tests were done on a three-phase induction motor, with known parameters, Table1.

Table 1. The three-phase induction motor parameters

Parameter	Value
p	2
J	0,04 [kgm^2]
R_s	2,71[Ω]
R_r	3,53[Ω]
L_s	0,268[Ω]
L_r	0,274[Ω]
L_m	0,265[Ω]

The general block diagram of this paper can be structured to reflect the main functional components of the control system, Fig.2.

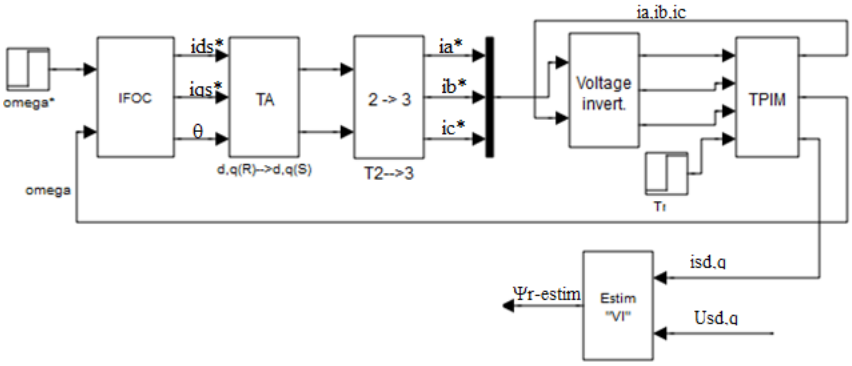


Figure 2. The block diagram of the control system

In the "IFOC" control block, the two components that produce the electromagnetic torque (active quantities) are decoupled by orthogonality from the quantities that produce the magnetization flux (reactive quantities), (12) and (13). This block has no information about the flux, but its value is imposed by the system design through a flux generator and also in this block its position is determined in relation to the fixed stator system θ_e (11), Fig.3.

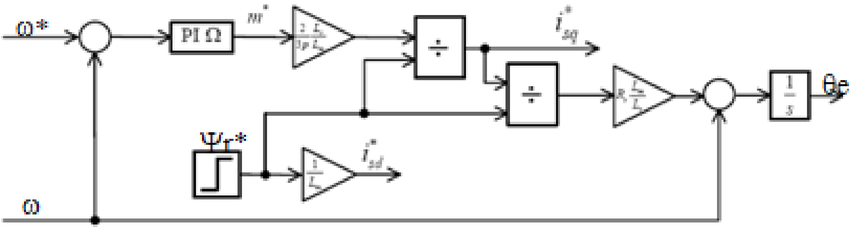


Figure 3. The block “IFOC” using rotor flux and torque to produce command currents

$$\theta_e = \int \left(\omega_r + \frac{R_r L_m}{L_r} \frac{i_{sq}^*}{\Psi_r^*} \right) dt \tag{11}$$

$$i_{sd}^* = \frac{1}{L_m} \Psi_r^* \tag{12}$$

$$i_{sq}^* = \frac{2}{3p} \frac{L_r}{L_m} \frac{1}{\Psi_r^*} m* \tag{13}$$

The "TA" block performs the coordinate transformation of the stator electric current from the mobile rotor system to the fixed stator system, using the position θ_e, determined in the command block.

The "TPIM" asynchronous motor is fed from a time-modulated, current-controlled voltage inverter. In this specific case, a bang-bang control is used for the motor currents. At the entrances of the block “Voltage Invert.” the prescribed instantaneous values of the currents on the three phases and the measured values of the currents must be applied. When the actual current deviates from the prescribed current, the bang-bang control rapidly turns on and off the transistors in the inverter to correct this error, resulting in a rapid variation of the voltage applied to the motor. Thus, this type of inverter generates phase voltages applied to the motor through closed-loop current control, ensuring accurate performance and rapid adaptation to dynamic load requirements.

The "Estim VI" rotor flux estimator block, Fig.4, has as inputs the stator quantities on the two orthogonal axes d-q of the electric voltages and currents. The output is an estimated value of the rotor flux, under the conditions of sensorless system operation.

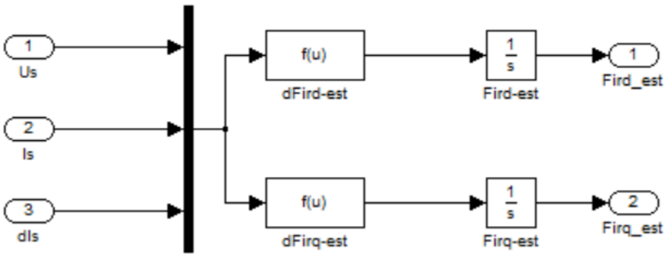
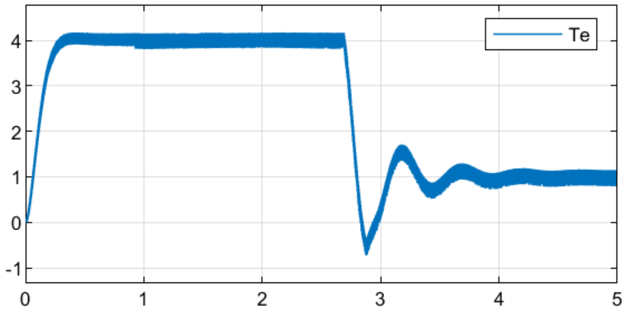


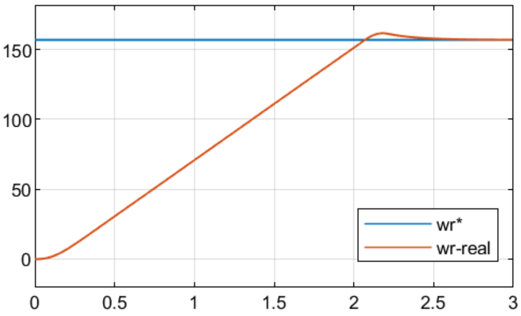
Figure 4. Rotor flux estimator block "VI"

To carry out the simulations, it was considered that the induction motor starts at idle, and after $t = 0{,}5s$ a load torque is applied $T_L = 1Nm$. In the evaluation of the performance of the system operation, the prescribed values of the speed and the rotor flux are made through a step type signal. The interpretation of the graphs regarding the variation of the speed and the estimated rotor flux are important for understanding the performances of the indirect vector control system, in a sensorless system. The three analyzed scenarios assume variations of stator resistance (R_s), rotor resistance (R_r) and supply frequency (f), respectively their impact on the magnetic and mechanical characteristics.

With the prescribed values of the asynchronous motor from Table 1, the mechanical characteristics of Fig. 5 were initially obtained and the magnetic characteristic from Fig. 6, in transitory regime. It can be seen that both the speed and the estimated rotor flux converge towards the imposed reference value, $\omega_r^* = 157 rad/sec$ and $\psi_r^* = 0{,}8Wb$.



a) electromagnetic torque



b) rotor speed

Figure 5. Mechanical characteristics of the asynchronous motor

Stator resistance (R_s) plays a critical role in maintaining stability of rotor speed and estimated rotor flux in an indirect vector control system. Its value influences both the accuracy of rotor flux estimation and the overall system performance. After performing several simulations in which the value of R_s changed by (+-50%) from the prescribed value, $R_s = 1{,}35\Omega; R_s = 4{,}06\Omega$, it was found that the system can maintain the stability of the rotor speed, its characteristic not being influenced, Fig.5. Also, the voltage drop across R_s of the "VI" estimator is correctly compensated, so that the estimated rotor flux is constant and its variations are minimal, Fig.6.

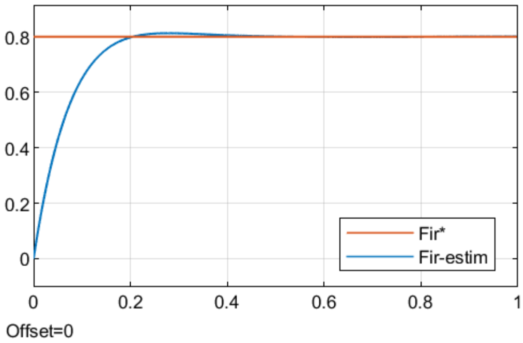


Figure 6. Characteristic of the estimated rotor flux

Rotor resistance (R_r) is an essential parameter in induction motor operation, having a significant impact on torque, speed characteristic and accuracy of rotor flux estimation. In an indirect vector control system, R_r is used in the calculation of synchronous speed and rotor flux orientation. If the value of R_r changes by ($\pm 50\%$) of the prescribed value, ($R_r = 1{,}77\Omega$; $R_r = 5{,}3\Omega$), the characteristic from Fig.7 and Fig.8 are obtained. Thus, for a lower value ($R_r = 1{,}77\Omega$), the motor shows a faster response of the rotor speed, in transient mode. If the resistance value is higher, ($R_r = 5{,}3\Omega$), then the system response is slower, Fig.7.

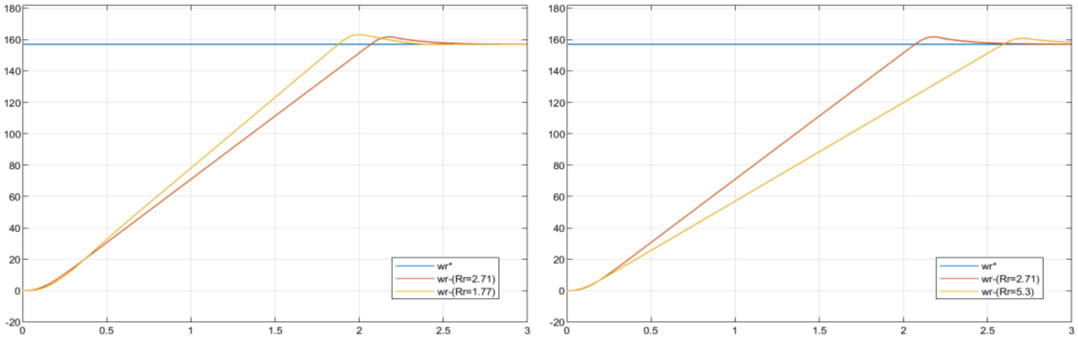


Figure 7. Rotor speed characteristics for $R_r = 1{,}77\Omega; R_r = 5{,}3\Omega$

The changes made on the rotor resistance have a negative impact on the rotor flux characteristics, Fig.8, observing the instability of the system.

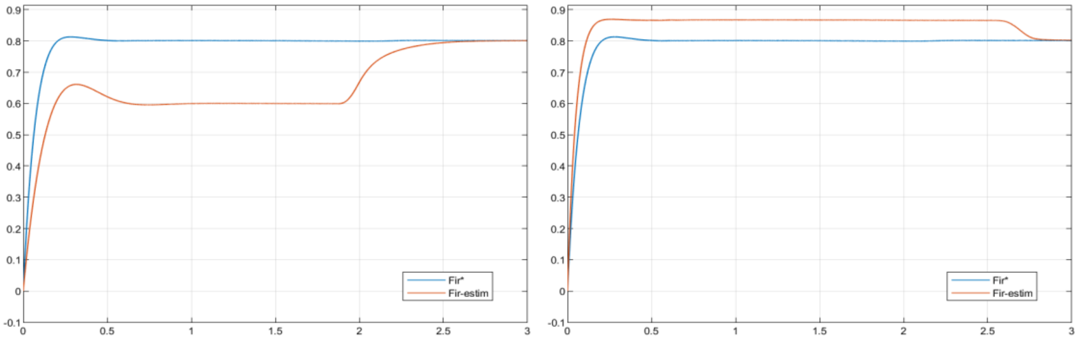


Figure 8. Rotor flux characteristics for $R_r = 1{,}77\Omega; R_r = 5{,}3\Omega$

Another parameter for which the system stability is checked is the frequency, ($f = 10Hz; f = 60Hz$), respectively the reference applied to the rotor speed, ($\omega_r^* = 2\pi f\ rad/s$). From Fig. 9 it can be seen that the rotor speed characteristics converge towards the estimated value, the faster the lower the frequency. The influence of frequency on the characteristic of the estimated rotor flux can be seen in Fig. 10, where the instability of the system is visible for high frequencies. At low frequencies the graph remains unchanged, Fig. 6.

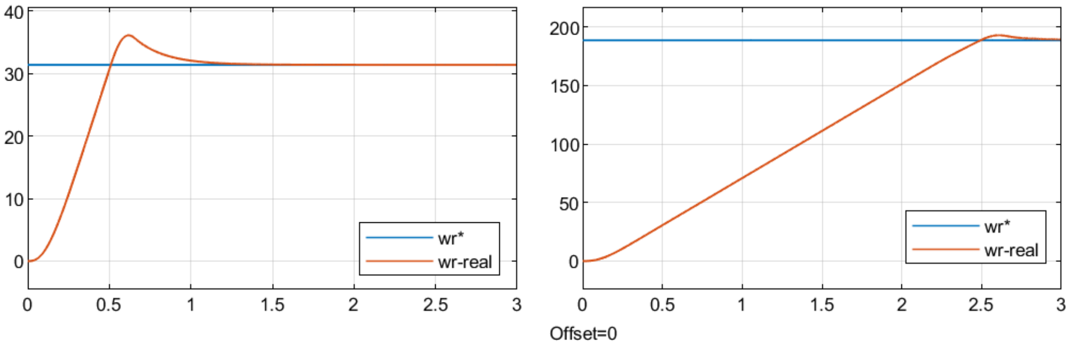


Figure 9. Rotor speed characteristics for $f = 10Hz; f = 60Hz$

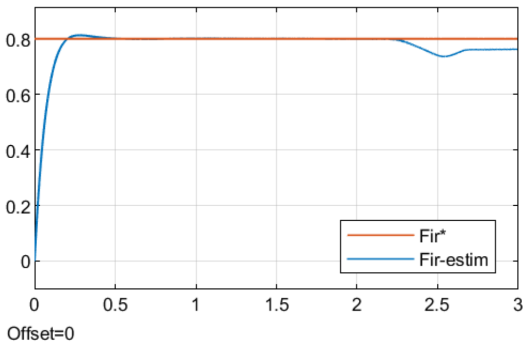


Figure 10. Rotor rotor flux characteristics for $f = 10Hz; f = 60Hz$

Conclusions

The paper highlighted the advantages and challenges of implementing indirect vector control of sensorless induction motors with emphasis on parameter variation, the use of a rotor flux estimator and overall system performance.

The variation of the parameters, especially the rotor resistance, can affect the accuracy of the flux and torque separation. It has been observed that significant parameter variations reduce the performance of the IFOC algorithm, highlighting the need for a robust model or real-time adaptation method.

The rotor flux estimator based on the volt-ampere (VI) integration technique has demonstrated the ability to provide accurate estimates under normal operating conditions. It is also relatively simple to implement and is used in applications where flux sensors are not available, but indirect control of the induction motor is required.

IFOC has proven to be an effective solution, providing accurate motor control even in the absence of sensors. Simulations have shown that the method provides good dynamic performance, but practical implementation requires additional compensations for the effects of parameter variation.

By integrating closed-loop current control, fast and robust system regulation was achieved, demonstrating the applicability of this solution in demanding industrial environments with high performance and reliability requirements.

References

[1] P. C. Krause, *Analysis of Electric Machinery*. IEEE Press, ISBN 0-471-14326-X, 1995.

[2] W. Leonhard, *Control of Electrical Drives*, 3rd ed. Berlin, Germany: Springer, 2001.

[3] I. Boldea and S. A. Nasar, *Electric Drives*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2005.

[4] R. Krishnan, *Electric Motor Drives: Modeling, Analysis, and Control*, Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[5] G. S. Buja, D. Casadei, and G. Serra, *Direct torque control of induction machines*, IEEE Transactions on Industrial Electronics, vol. 51, no. 4, pp. 744-757, Aug. 2004.

[6] L. Harnefors, *Design and analysis of adaptive observers for sensorless control of induction motors*, IEEE Transactions on Industrial Electronics, vol. 48, no. 2, pp. 127-137, Apr. 2001.

[7] B. K. Bose, *Modern power electronics and AC drives*, IEEE Press, 2001.

DOI: 10.2478/ijasitels-2024-0003

Reflow oven with Wi-Fi connectivity

Barb Ovidiu-Andrei[1]

[1]Computer Science and Electrical Engineering Department, "Lucian Blaga" University of Sibiu, Romania, E-Mail: ovidiu.barb@ulbsibiu.ro

Abstract

This paper will address a theme that highlights the role of an electrical equipment in the electrical industry, focusing on the process of soldering components onto the surface of printed circuit boards.

The choice of the theme "Reflow Oven with Wi-Fi Connectivity" was driven by the desire to create a system with practical and applicative outcomes. The aim is to develop equipment that facilitates the soldering process of Surface Mounted Device (SMD) components. Such an automated system is valuable in the production of small batches of equipment, such as electronic modules for prototypes or various applications.

The equipment or electronic module aims to depict a Printed Circuit Board (PCB) on which both Surface Mounted Devices (SMD) and Through Hole Technology (THT) components are mounted. Throughout the paper, the implementation of an automated system capable of carrying out the soldering process for SMD components will be discussed.

Keywords: IoT, reflow oven, Wi-Fi, soldering, PCB

1 Introduction

Humanity has been experiencing a powerful technological innovation in the development of electrical and electronic components and equipment, aimed at assisting people by integrating them into daily life. The automation and intelligent control aspect seems to increasingly take over both simple and complex operations, replacing human personnel, with a focus on complete automation of industrial processes and beyond.

One of the most important categories of electronic devices consists of circuit elements that provide connectivity between devices or, specifically, to the internet. Connected devices are becoming more prevalent in people's purchasing desires because of their intuitive, simple, and user-friendly interaction and usage.

From this perspective, Internet of Things (IoT) devices are increasingly present in people's homes and in the industrial environment due to their convenient interaction and numerous benefits. Through such connected devices, the user experience becomes much friendlier, as they can be controlled from devices users already own, such as mobile phones.

In terms of construction, simplicity is the key, as IoT devices are largely uncomplicated, with few control or display elements. The entire interaction with these devices takes place through the internet using mobile phones, web pages, or other interconnecting devices.

For electrical devices to achieve remarkable performance, it is necessary for the components used to be energetically efficient. Therefore, the continuous search for new devices that propose energy improvements is essential, as electrical energy, the "raw material" of this industry, is crucial and cannot be wasted carelessly. Simultaneously, the improvement of current devices is pursued to approach the ideal characteristics, including minimal energy dissipation, low current consumption, fault-free operation over time, or the ability to withstand increasing command or data processing speeds.

In the case of producing small series of equipment, industrial reflow ovens cannot be used efficiently due to factors such as the low number of printed circuit boards passed through the oven or the small number of components that need to be soldered onto the printed circuit board. Industrial reflow processes are most efficient when applied to the production of large batches of similar equipment. Therefore, acquiring an industrial reflow oven for small-scale needs is often impractical. Hence, the concept of designing a smaller device for such purposes is considered. This involves creating a compact reflow oven that closely approximates the characteristics of an industrial reflow oven.

To benefit from this technology in small equipment series, the design of a much smaller system, compared to an industrial device, is proposed to carry out the reflow process in an easy and automated manner. Therefore, the chosen theme proposes the implementation of an automated system that serves as a reflow oven, which can be wirelessly controlled from various mobile or fixed devices by connecting them directly to the designed system.

The designed system is intended for the production of prototypes, small or medium series of electronic equipment concerning soldering components onto the surface of the printed circuit board for interconnection purposes.

Thus, the oven is defined as a well-insulated enclosure where the temperature will be controlled according to a certain characteristic initially set. The main idea is to create a system that monitors a temperature characteristic to solder Surface Mounted Device (SMD) components onto the printed circuit board using solder paste intended for this process.

In this regard, a food preparation oven will be used, which will be entirely adapted from an electrical and mechanical point of view to meet the thematic requirements. Essentially, the significance of the theme lies in creating equipment that can monitor a temperature characteristic established according to the requirements of the components and solder paste.

It is imperative that the communication with the entire device be wireless, functioning as an Access Point (AP). Devices with wireless connectivity can connect to it to establish communication following the Transmission Control Protocol/Internet Protocol (TCP/IP) model. The need for communication with the system arises from the desire for the reflow process to be monitored and controlled through a web page accessible from any device's browser, initiating the connection and communication with the terminal. From this perspective, the implementation method should not involve additional components to control the terminal; instead, they should be replaced by a web page interface, significantly reducing the use of electronic components. The advantage of software implementation is that any command button or display can easily be transposed into a graphical interface by replacing them with virtual ones. The interaction with these virtual components can be done using devices that are already part of our technology era.

Constraints related to implementation are accentuated by the fact that, from the beginning, a small-sized oven is desired for use in small equipment series. Thus, for the production of multiple electronic modules exceeding the oven's capacity, multiple soldering cycles will be required, repeated until the desired number of boards with soldered components is achieved.

Another technical challenge would be the rapid cooling process inside the oven to meet the imposed reflow characteristic. In this regard, various methods for implementing the cooling system will be analyzed.

The imposed objectives are related to the implementation of an IoT system, which is ultimately intended to be functional and productive. Requirements can be outlined starting from the main idea of implementing the described system based on a microcontroller with the greatest possible integration of peripherals required for the design, followed by the need to implement a temperature regulator and a web interface for oven control and monitoring.

The device should be safe in terms of handling during operation and electromagnetic compatibility with other devices. It should also be environmentally friendly in relation to the surrounding environment or the power grid from which it will be powered. Among the requirements, the possibility of selecting the reflow curve through the graphical interface, as well as monitoring the process through real-time data monitoring, can be mentioned.

2 Stages of Soldering through Reflow Procedure

Soldering through reflow involves a series of well-defined individual stages in the process. The following presentation provides an overview of these stages as they constitute an important topic within the scope of the work.

The stages through which the assembly requiring soldering must pass are described by the temperature variation curve of the reflow process, as depicted in Figure 1.

The necessity for such a procedure arises from the fact that temperature represents the greatest adversary to electronic components. As soldering involves a thermal process on these components, it is crucial to control the thermal exposure.

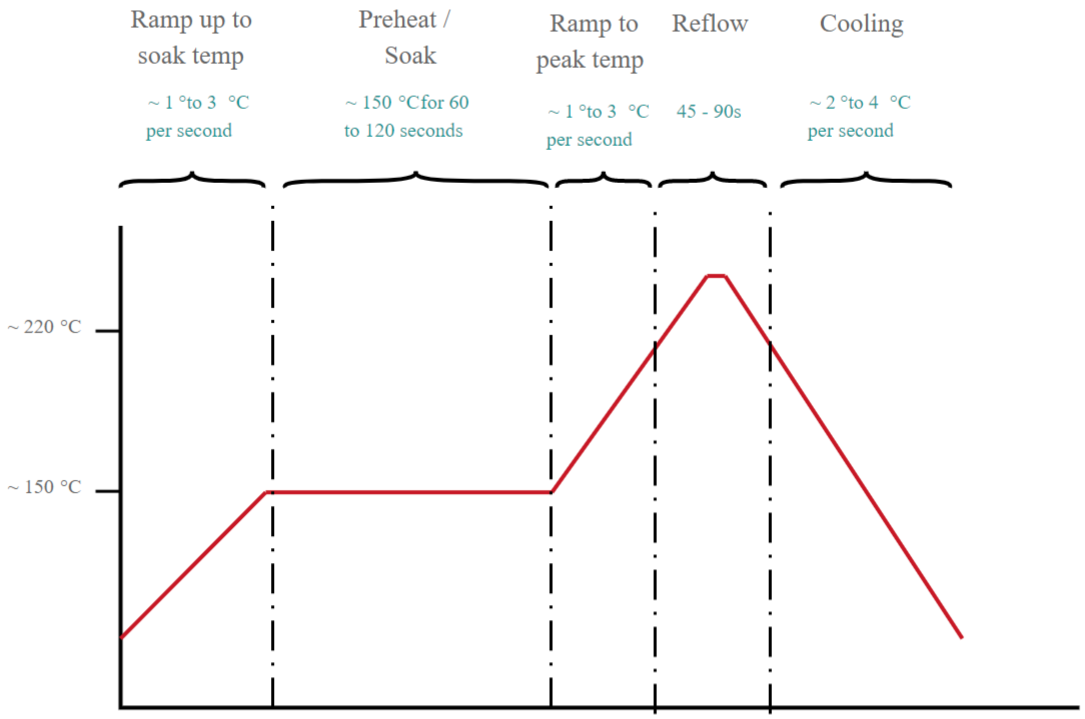


Figure 1: Reflow Process Temperature Profile

Achieving the correct temperature profile during the process ensures the quality of the resulting solder joints and is crucial for preventing component damage.

The following are brief descriptions of the stages that must be followed for the assembly of electronic equipment, with reference to the aforementioned diagram:

Preheating (Ramp up to soak temperature): This stage initiates when the assembly, consisting of the printed circuit board with pre-applied paste and components, is introduced into the oven. The goal of the first segment in the diagram is to gradually bring the assembly to a constant temperature. The gradual temperature increase avoids thermal shocks to the components, preventing potential damage. A sudden temperature increase could cause certain areas not to reach the required temperature due to their thermal mass. The temperature ramp-up rate typically ranges between approximately 1 and 3°C.

Soaking (Soak): After reaching the preheating temperature, the assembly enters the "soak" segment, where it is maintained at a constant temperature for an extended period. This stage ensures that the PCB and components reach the required temperature uniformly, considering possible thermal shadowing areas caused by larger components. Additionally, it allows for the removal of volatile substances from the solder paste and provides time for the activation of the flux.

Ramp to Peak Temperature (Ramp to peak temperature): This segment involves a continuous temperature increase of the assembly to prepare it for the next stage.

Reflow: This is the stage where the soldering process takes place. It reaches the highest temperature in the entire process. Soldering occurs as the solder paste melts, creating the necessary electrical connection. Microscopically, this stage is analysed from the perspective of the solder paste. The flux in the solder paste reduces surface tensions at the meeting point of the two metals (component pad and PCB pad), facilitating the formation of a metallurgical bond by allowing individual solder balls to combine through melting.

Cooling: The temperature decreases for already soldered assemblies is carried out in a way that avoids potential mechanical or thermal stress on the components. Proper cooling inhibits the formation of excess intermetallic compounds (cold solder) or thermal shock. Cooling temperatures typically range between 30 and 100°C. The cooling rate is relatively fast, chosen to create a solder joint with a fine structure and a solid bond, ensuring mechanical strength through the solidification of the liquid metal. The described stages of the reflow process will serve as a foundation for designing the oven within the project.

In the industry, large-sized ovens are used, and the entire mentioned process takes place horizontally. The structure of such an oven can be observed in Figure 2.

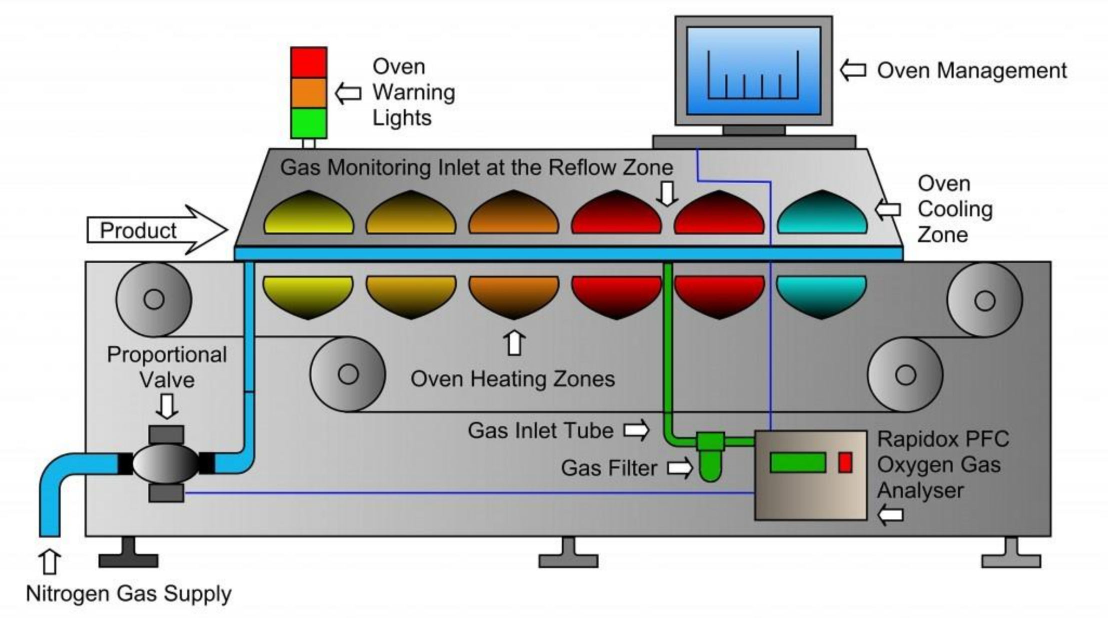


Figure 2. Internal Structure of an Industrial Reflow Oven

In addition to the previously described stages, it should be mentioned that an industrial oven will perform the entire soldering process in a gas environment, with nitrogen, which contributes to the quality of the solder joints. The absence of oxygen in the environment halts the oxidative process, positively impacting the quality of the solder joints.

The oven is composed of several horizontally arranged segments, defining the temperature zones within the reflow curve.

3 Approach

For the design of the entire system, requirements and ways in which they could be implemented will be specified by finding a theoretical solution that can be realized practically with the necessary materials. The final assembly needs to impose a safe operating regimen, considering the integrability of multiple components into the same system.

The technical design of the system will be done in stages, dividing the entire assembly into subsystems that will eventually create the whole system. It is crucial that each block element to be implemented in the end constitutes a whole by combining all the blocks and imposing the condition of coexistence between them. In other words, each implemented block must take into account the other components of the system so that, when combined, the final system can be functional, stable, and safe.

The main idea of a reflow oven is to have a controlled, insulated, and heated chamber where the reflow or "reflow soldering" process of SMD components takes place. Such ovens are used in the industry for mass production of electronic equipment. Before reaching the soldering process, one or more individual PCBs go through the process where solder paste is applied using a stencil, and then automated or semi-automated component placement takes place. Once the PCB reaches the equipped stage, soldering of the components is necessary.

The reflow oven comes into play at this stage, making the soldering of components possible. It is important that soldering is done in a controlled manner, adhering to the data-sheet of the components, as well as the soldering details imposed by the solder paste in terms of soldering phases and temperatures.

This paper aims to address this topic with the purpose of implementing such equipment following the stages of analysis and design. In the end, the goal is to obtain equipment that can produce small and medium-sized batches, enabling the soldering of multiple SMD components simultaneously while adhering to the stages of the reflow process. The oven is intended to be controlled and monitored wirelessly from a device.

4 Implementation

The starting point for designing the system involves traversing the reflow curve and establishing the necessary elements to enable temperature variations using the raw material described earlier.

In the first stage of the reflow process, a gradual temperature increase occurs, involving the two heating elements and the fan for temperature homogenization within the chamber. Solid State Relay (SSR) modules will be used to control these three elements, connecting the DC control part to the AC power part.

The decision to use SSR modules is influenced by their galvanic isolation between circuits and their ability to provide faster command execution compared to traditional relays with coils and moving armature.

Another important factor in choosing SSR command modules is their capability to switch at regular and well-defined intervals, ensuring a smoother current absorption from the power grid while respecting its sinusoidal shape as much as possible. Another element that requires power from the mains is the light bulb for illuminating the oven chamber, and a classic relay can be used for its control since it operates in only two states.

For temperature monitoring, a 100Ω RTD sensor will be used along with a conversion element to be interpreted by the microcontroller.

Regarding the cooling stage of the chamber, a fan will be used to introduce air into the oven, facilitating the cooling process. This stage involves the use of two elements to open and close the ventilation openings.

Safety elements include the addition of a fuse across the entire oven structure to prevent unforeseen incidents, as well as a sensor to determine the state of the door during operation.

Since audible warnings are among the most easily perceived, such an element will be added to emit sounds at certain stages of the process or in case of warnings.

All these elements, of course, require a central control module to execute the entire soldering process.

Finally, it is worth mentioning that the power supply needs of all modules within the equipment are another important consideration in its design.

To provide a clear overview of the equipment's needs, a block diagram has been created, as shown in Figure 3.

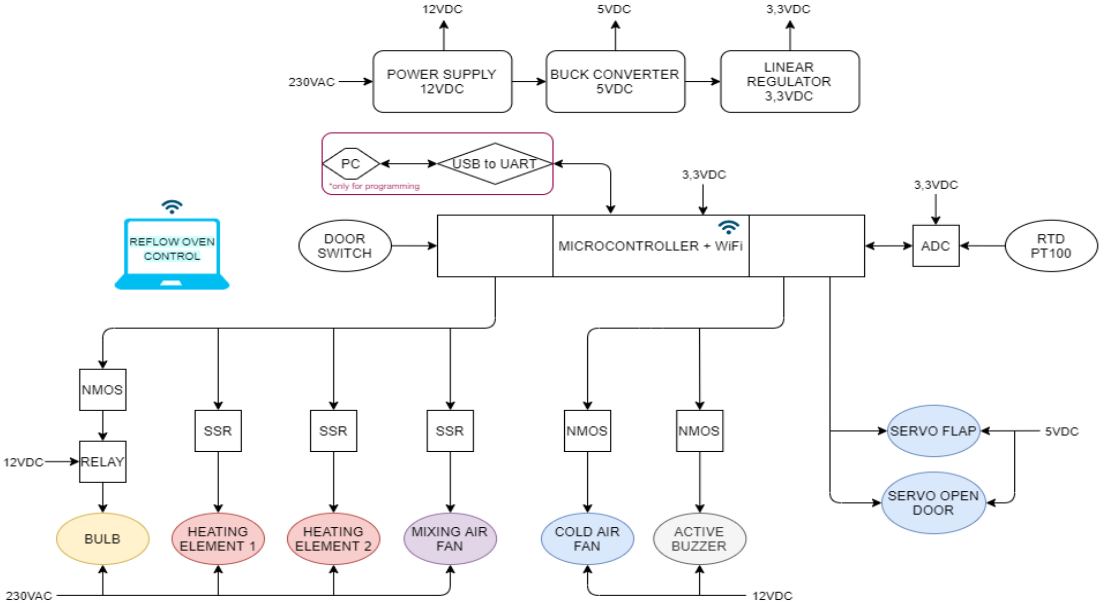


Figure 3: Block Diagram of the System

5 Technical solution

To enable the interconnection of all modules designed throughout the work, the design of the printed circuit board (PCB – Figure 4) was necessary. Thus, based on the schematic that was created, each functional group of the schematic was translated into the PCB layout.

Several essential considerations were taken into account for the PCB design, especially in this case when dealing with mains voltage, which will be present on the same PCB along with low voltages for supply and control, which are direct current.

The PCB component layout aimed to create an imaginary line of separation between the section working with mains voltage and the section handling low voltages.

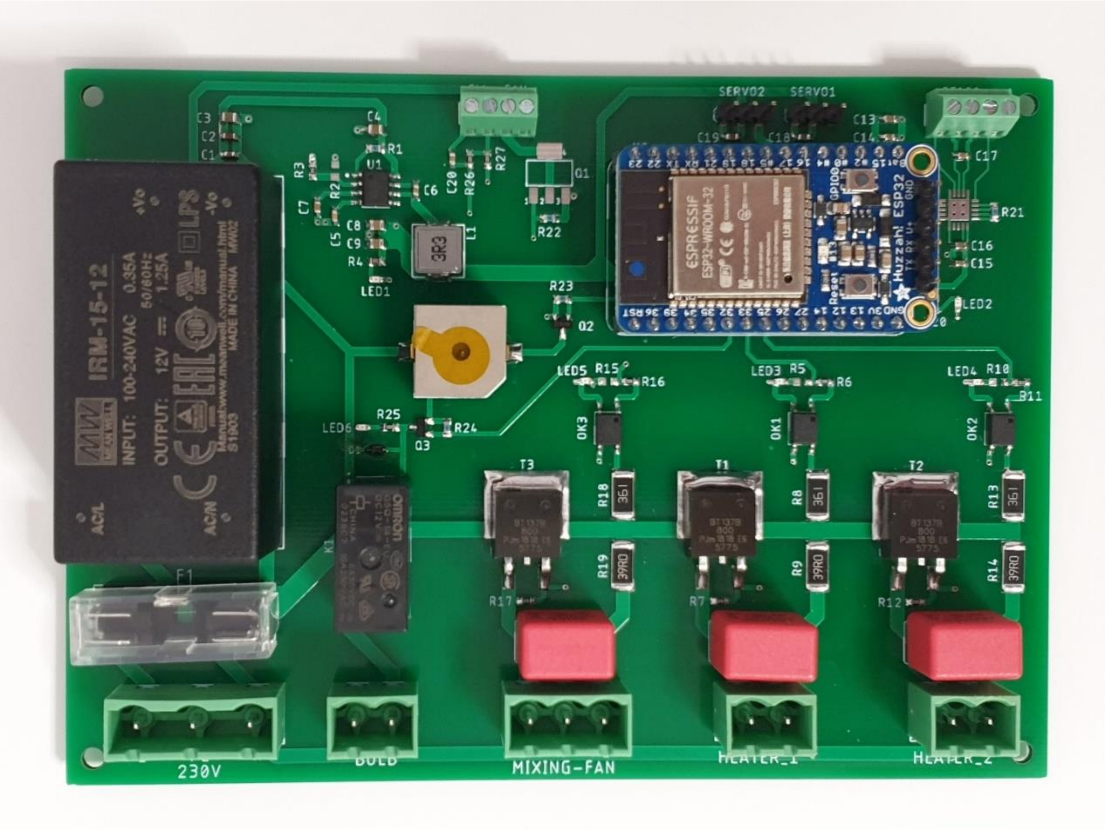


Figure 4: Assembled PCB (Motherboard of reflow oven)

6 Conclusion

The entire conceptual mobilization discussed throughout the work originated from the desire to create equipment through which the soldering procedure of SMD components could be carried out as efficiently as possible. The adoption of this concept was derived from the industrial environment, where such equipment is capable of producing electronic modules on a large scale. However, these devices are intended for mass production of electronic modules, and the process for producing prototypes or small series of equipment is costly and cumbersome.

In this regard, studies were conducted on the soldering process and methods through which it could be performed on a much smaller scale. The objective of the study involved seeking solutions for implementing a reflow oven with functionalities as close as possible to those of an industrial oven. To achieve this goal, a small electric oven was used as the starting point for the research.

It should be noted that heat is the greatest enemy of electronic components, and soldering SMD components onto the printed circuit board must be done in accordance with the manufacturer's specifications in terms of the reflow characteristic for specific components and solder paste. Although this process requires time, it is compensated by the correct operation of the components, preventing damage due to the heat involved in the soldering process.

The constructive elements of the oven were analysed from both electrical and mechanical perspectives. Measurements were taken to determine the electrical parameters of the components, which was necessary for the implementation of control modules. Among the most important parameters were the supply voltage and current consumption.

The next step involved the step-by-step creation of electronic modules, starting from elements that perform input functions to those that perform output functions. Input elements included temperature sensors and micro-switch sensors for door state detection. An industrial implementation method, using an integrated circuit, was employed to interpret the temperature sensor readings.

Output elements included heating resistors, the light bulb, and the fan for air circulation. Control modules using Solid State Relays (SSRs) directly on the circuit board were created for these elements. Component choices were made within all functional blocks following detailed calculations and considering the objectives they were expected to achieve. Various simulations were conducted, along with practical tests, to ensure the implementation was both safe and functional.

In the design process, an initiative was taken to consider the cooling process of the soldered module to fully comply with the reflow process characteristic. Additional elements were added to create an airflow to assist in cooling the oven enclosure.

The project's scope is oriented towards the production of electronic modules, particularly rapid prototyping or creating small series of electronic equipment. In this context, the oven benefits the user by reducing the practical soldering time.

Future development of the project will focus more on user interaction and functionality, aspects that form the basis of the software structure. There is a desire to enhance how the oven operates, closely following the reflow process.

Another idea for further development is to create an Android application for easy user interaction with the equipment. The emphasis of the work has been more on hardware design.

In conclusion, such equipment is useful due to the advantages it offers users in terms of aiding the soldering of SMD components, which is not always easy to accomplish manually and is a time-consuming and energy-consuming process.

Additionally, the advantage of such an IoT device lies in the fact that control can be done from devices we already possess. The reflow oven does not have peripheral elements that allow physical interaction with the user, such as buttons or various types of displays, resulting in material savings. Control is achieved solely through wireless means.

It is worth mentioning that throughout the system's design, significant attention was given to the area of electrical safety, ensuring that the user is not exposed to potential hazards during the operation of the equipment.

References

[1] https://www.researchgate.net/figure/A-close-up-view-400X-of-63Sn-37Pb-solder-paste-with-particle-size-smaller-than-20_fig17_3967598
[2] https://en.wikipedia.org/wiki/Solder_paste
[3] https://www.pcbdirectory.com/community/what-is-a-pcb-stencil
[4] https://www.cambridge-sensotec.co.uk/blog/2017/08/22/detecting-residual-oxygen-within-solder-reflow/
[5] https://www.pcbgogo.com/Blog/Solder_Paste___Something_You_Need_to_Know.html
[6] https://liteplacer.com/
[7] https://www.electronics-notes.com/articles/constructional_techniques/soldering/reflow-soldering.php
[8] https://en.wikipedia.org/wiki/Reflow_soldering
[9] https://en.wikipedia.org/wiki/Reflow_soldering
[10] https://www.electronicshub.org/wireless-communication-introduction-types-applications/
[11] https://www.toyo-chem.com/en/products/electronics/column/emishield.html
[12] https://ro.wikipedia.org/wiki/Wi-Fi
[13] https://components101.com/articles/buck-converter-basics-working-design-and-operation
[14] http://www.airsupplylab.com/embedded/61-pulse-width-modulation-pwm.html
[15] https://www.youtube.com/watch?v=W6NOV6b8kxs&t=1s&ab_channel=Explorer
[16] http://offers.kd2.org/en/ie/lidl/peclN/
[17] https://en.wikipedia.org/wiki/Reflow_soldering
[18] https://www.ia.omron.com/support/guide/18/further_information.html
[19] https://ro.mouser.com/datasheet/2/408/TLP268J_datasheet_en_20200217-1915950.pdf
[20] https://learnabout-electronics.org/Semiconductors/thyristors_66.php
[21] https://www.mouser.com/datasheet/2/302/BT137B_SERIES-110367.pdf
[22] https://www.tme.eu/Document/cfdb8f8c8e63fe897cce310c42942d6f/MOC3063SR2VM.pdf
[23] https://www.electroschematics.com/do-it-yourself-solid-state-relay/
[24] https://www.youtube.com/watch?v=rLpFtSayZ3Q&t=2s&ab_channel=Robert%27sSmorgasbord
[25] https://www.homemade-circuits.com/efficient-electronic-relay-ssr-circuit/
[26] https://www.st.com/resource/en/application_note/cd00004096-rc-snubber-circuit-design-for-triacs-stmicroelectronics.pdf
[27] https://www.thierry-lequeu.fr/data/AN437.pdf
[28] https://www.st.com/resource/en/application_note/dm00451014-controlling-a-triac-with-a-phototriac-stmicroelectronics.pdf
[29] https://www.youtube.com/watch?v=wgNMepGIrTk&t=593s&ab_channel=w2aeww2aew
[30] https://www.youtube.com/watch?v=-hSbgJhAmgA&ab_channel=Chris%27sWorkbench
[31] https://www.tme.eu/Document/9e42ba54d36a9bd3dfa2901986ac022d/rezystor-smd.pdf
[32] https://www.youtube.com/watch?v=2UtL2uAYCUA&ab_channel=GreatScott%21GreatScott%21
[33] https://www.electronics-tutorials.ws/blog/relay-switch-circuit.html

[34] https://ro.mouser.com/datasheet/2/307/en_g5q-22333.pdf
[35] https://www.onsemi.com/pdf/datasheet/bss138-d.pdf
[36] https://www.tme.eu/Document/4bd5d86455f1a302d8c4826bfeca5ba1/BAS16HT1-D.pdf
[37] https://www.eejournal.com/article/ultimate-guide-to-switch-debounce-part-3/
[38] https://www.tme.eu/Document/5cfedf3b6b75da7c93f650e2057f2381/SMT-1205X.pdf
[39] https://www.tme.eu/Document/d320f91e8cbcfe0e779b6c2898dd7a87/YJM04N10A.pdf
[40] https://ardushop.ro/ro/electronica/93-servomotor-sg90.html
[41] https://www.basstor.com/thermocouple/4-wire-pt100-temperature-sensor.html
[42] https://www.frigotehnie.ro/blog/senzor-temperatura-pt100-vs-pt1000/
[43] https://www.meerstetter.ch/products/systems-software-accessories/sensors/pt100-sensor-suggestions
[44] https://datasheets.maximintegrated.com/en/ds/MAX31865.pdf
[45] https://www.digikey.ro/product-detail/en/espressif-systems/ESP32-WROOM-32/1904-1010-1-ND/8544305
[46] https://www.espressif.com/sites/default/files/documentation/esp32_technical_reference_manual_en.pdf
[47] https://randomnerdtutorials.com/esp32-pinout-reference-gpios/
[48] http://esp32.net/
[49] https://en.wikipedia.org/wiki/ESP32
[50] https://lastminuteengineers.com/esp32-sleep-modes-power-consumption/
[51] https://www.espressif.com/sites/default/files/documentation/esp32_datasheet_en.pdf
[52] https://www.espressif.com/sites/default/files/documentation/esp32-wroom-32_datasheet_en.pdf
[53] https://www.espressif.com/sites/default/files/documentation/esp32_hardware_design_guidelines_en.pdf
[54] https://learn.adafruit.com/huzzah32-esp32-breakout-board
[55] https://www.diodes.com/assets/Datasheets/AP2112.pdf
[56] https://ro.mouser.com/datasheet/2/115/AP65550-335805.pdf
[57] https://www.tme.eu/en/details/hpi0640-3r3/smd-power-inductors/ferrocore/
[58] https://www.meanwell-web.com/content/files/pdfs/productPdfs/MW/IRM-15-spec.pdf
[59] https://www.digikey.ro/ro/resources/conversion-calculators/conversion-calculator-pcb-trace-width
[60] https://ro.wikipedia.org/wiki/Interfa%C8%9Ba_serial%C4%83_SPI
[61] https://randomnerdtutorials.com/esp32-access-point-ap-web-server/
[62] https://learn.adafruit.com/adafruit-max31865-rtd-pt100-amplifier/arduino-code
[63] https://danstrother.com/2011/01/15/reflow-oven-controller/
[64] https://ro.wikipedia.org/wiki/Legea_lui_Ohm
[65] https://www.ti.com/lit/an/spra906/spra906.pdf?ts=1624119147213&ref_url=https%253A%252F%252Fwww.google.com%252F
[66] Rui Santos & Sara Santos - ESP32 Web Server with Arduino IDE, eBook
[67] Neil Kolban – Kolban’s Book on ESP32, September 2018
[68] https://ftdichip.com/wp-content/uploads/2020/08/DS_FT230X.pdf

DOI: 10.2478/ijasitels-2024-0016

Single Phase Induction Motor Electrical Performances

Alina Cristina Viorel[1]*, Gabriela Crăciunaș*[2]

[1]*Computer Science and Electrical Engineering Department, "Lucian Blaga" University of Sibiu, Romania, E-Mail: alina.viorel@ulbsibiu.ro gabriela.craciunas@ulbsibiu.ro.*

Abstract

A **single-phase induction motor (SPIM)** is one of the most widely used types of electric motors for applications that require small to medium power even his features don't look very efficient energy parameters point of view. This motor runs on single-phase AC power, a real advantage for residential and light industrial applications and coupled with a DC permanent machine can emphasize its behavior in some specific situations.

Starting and running performance of single-phase induction motor depend on running torque requirements of the load. Some statements about power factor and efficiency will be done, also.

Keywords: SPIM, running torque, power factor, efficiency.

1. Introduction

1.1 Brief description of the SPIM

Most single-phase induction motor are two-phase motors with unsymmetrical windings and their axes have perpendicular positions. One of the stator windings, the main one, is in 2/3 slots distributed and is supplied straight on the source. The other one, considered secondary one, can be disconnected from the source after starting process or can remain connected through a serial impedance. The role of secondary windings is to create a create a phase shift to start the motor.
The rotor construction remains in conventional direction and the winding is a squirrel one which consists of conductive bars shorted at both ends.[1]
The construction of a single-phase induction motor is done to realize operating features more efficiently when a single-phase power supply is available.

1.2 Working Principle

To produce a starting torque a phase shift is necessary to create between the sinusoidal space distribution of magnetomotive force (mmf). Without shift phase the stator windings will produce an equal forward and backward- rotating mmf waves. By symmetry, such a motor inherently will produce no starting torque since at standstill, it will produce equal torque in both directions (figure 1).

The role of secondary stator winding is more than obvious to determine an interaction between the magnetic field produced by the stator and the magnetic field created by the induced current in the rotor. The torque created by this action causes the rotor to start.[2]

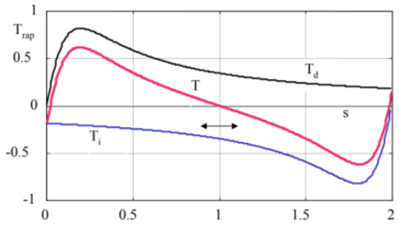


Figure 1. Torque-slip characteristic of a single-phase induction motor on the basis of constant forward and backward flux waves

The resultant torque-slip characteristic, which is the algebraic sum of the two components curves (direct torque Td and inverse torque Ti), shows if the motor were started by auxiliary means, it would produce torque in whatever direction it was started.

The type o impedance implemented could be resistive or capacity one. The expression of stator current from auxiliary winding is:

$$\underline{I}_{aux} = \frac{\underline{U}}{\underline{Z}_{sc\ aux} + \underline{Z}} \tag{1}$$

Where: $\underline{I}_{aux}$ - the value of auxiliary stator winding

$\underline{U}$ - the supply voltage

$\underline{Z}_{sc\_aux}$ - the value of impedance placed on auxiliary winding

$\underline{Z}$ - the impedance of main winding

The maximum of starting torque is depending on the shift phase between phases (φ_{sc} - φ), due to the interaction of fields created when the currents pass the windings placed in 90-degree angle. The φ is the argument of $\underline{Z}$ impedance and φ_{sc} is the argument of main short circuit winding.

The expression of the maximum starting torque $T_{p\ max}$ is:

$$T_{p\ max} = K_T I_{pr} I_{aux} \tan\left(\frac{\varphi_{sc} - \varphi}{2}\right) \tag{2}$$

Where: K_T – motor constant torque.

φ - the argument of main winding

φ_{sc} – the argument of main short circuit winding (for identical windings).[3]

The difference (φ_{sc} -φ) represents the shift phase between windings if the phases are identical. If $\underline{Z}$ is resistance, then φ =0. In this case the starting torque is not so significative.

The value of starting torque is significative if the impedance contains a capacitor. The problem is for considered value of capacitors is that the shape of rotational magnetic field is circular for a certain load. If will be another motor load the capacitance of capacitor should be changed.

1.3 Starting and running performance of single-phase induction motor.

Classification of single-phase induction motor depends on the starting methods. Selection of the appropriate motor is based on the starting and running-torque requirements of the load, the duty cycle of the load, and the limitations on starting and running current from the supply line for the motor. The cost of single-phase motors increases with their rating and with their performance characteristics such as starting torque to current ratio. Typically, to minimize cost, an application engineer will select the motor with the lowest rating and performance that can meet the specifications on application.

This paper presents a typical single-phase induction motor, connected to dc permanents generator used as a load. More, at the end to armature of dc generator winding a resistance is connected. This assembly is typical laboratory essay in according to study the features of the SPIM. [4]

2 Simulation of a SPIM connected with a dc permanent magnet generator

The electric drive is realised with a single-phase induction motor as electric motor and a dc permanent magnet generator coupled to the shaft of the induction motor. As the motor turns, it drives the generator to rotate within its magnetic field. This rotation induces a DC voltage in the generator’s stator windings.

The electrical output from the generator is used to power a load, a resistance in this case.[4]

2.1 Performance of SPIM coupled with a dc permanent magnet generator

2.1.1 Description of the model

This simulation model realized in MATLAB using *Power Systems Blocks* shows the operation of a single-phase asynchronous motor with auxiliary phase operation modes.

This model uses single-phase asynchronous motor, a block existing in Matlab library, to compare their performance characteristics, such as torque, efficiency and power factor in different situations. The motor has 186.5 W, 220 V, 50 Hz, 1500 rpm. The system is fed by a 220V single phase power supply. It has identical stator windings (main and auxiliary) and rotor squirrel cage.

The working machine is a dc permanent magnet machine with no pre-set data.

The motor is first started at no load, at t=0. Then at t = 0.5 sec, a 0.3 Nm. torque is suddenly applied on the shaft (torque signal build with a signal generator). Also, the load machine (dc permanent magnet generator) is coupled with the second breaker.

At t= 1 sec. the first breaker is disconnected, and the system works with the main single-phase induction motor feeds by supply.

The entire process of simulation ended at t = 2 sec. when the system remains stationary.

2.1.2 The simulation model realised in Matlab/Simulink application

In the figure 2 the simulation model is presented. The model was realised in Matlab 2024 version. [5]

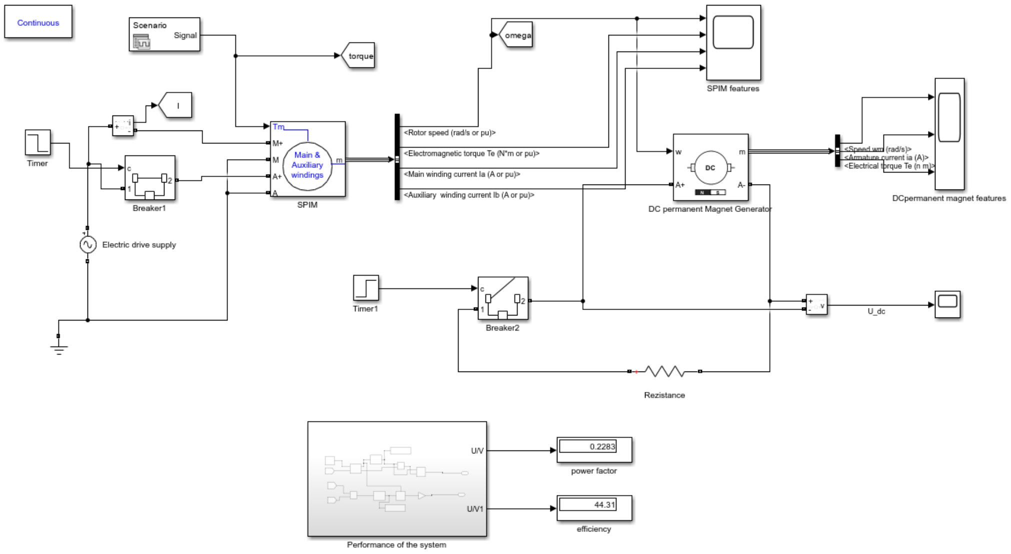

Continuous
Scenario
Signal
torque
omega
I
Timer
Breaker1
Electric drive supply
Tm
M+
M
A+
A
Main & Auxiliary windings
m
SPIM
<Rotor speed (rad/s or pu)>
<Electromagnetic torque Te (N*m or pu)>
<Main winding current Ia (A or pu)>
<Auxiliary winding current Ib (A or pu)>
SPIM features
w
m
DC
A+
A-
DC permanent Magnet Generator
<Speed wm (rad/s)>
<Armature current ia (A)>
<Electrical torque Te (n m)>
DCpermanent magnet features
Timer1
Breaker2
U_dc
Rezistance
U/V
U/V1
0.2283
power factor
44.31
efficiency
Performance of the system

Figure 2. Electric drive simulation system

The model contains a subsystem for performance calculus: the power factor and the efficiency of the SPIM. To obtain these two parameters the usual equations were used. The equation for power factor is:[6]

$$\cos\varphi = \frac{P}{S} = \frac{P}{\sqrt{P^2 + Q^2}} \tag{3}$$

Where: P - active power
Q -reactive power
S- apparent power

For the calculus of efficiency, the equations used is:

$$\eta = \frac{P_{out}}{P_{in}} \tag{4}$$

Where: P_{out} –mechanical power
P_{in} – the power measured at the beginning of the system

The values are obtained using a specific block from library named *power* which compute active and reactive powers of voltage-current pair at fundamental frequency. In the subsystem a multimeter is used for making the connections easier.

2.1.3 Results obtained from simulations.

The Scope block displays the following signals for the model regarding SPIM: rotor speed (green trace) electromagnetic torque (blue trace), main winding current (magenta trace), auxiliary winding current, (brown trace), The mechanical power, power factor and efficiency of motor are computed inside the *Performance of the system* subsystem. Figure 3 describes the behavior of the system in the work sequence created.

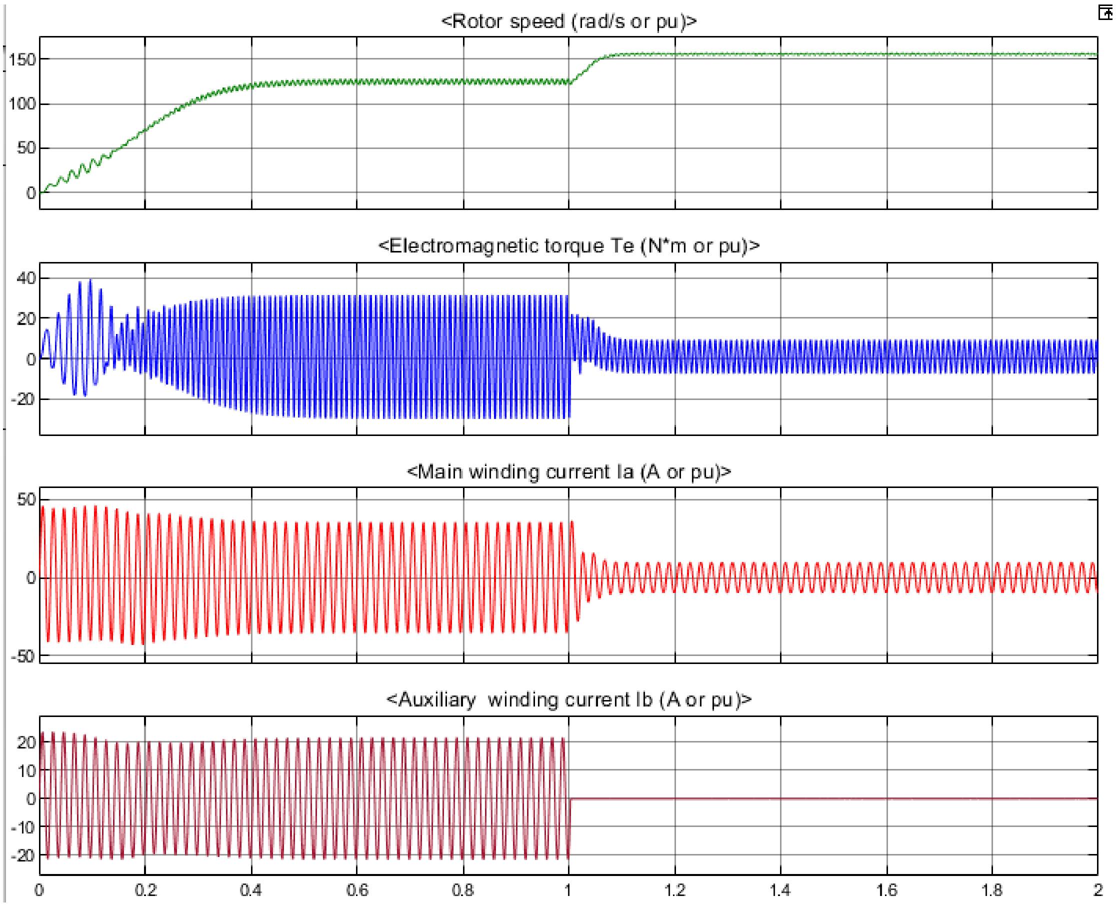


Figure 3. The behaviour of SPIM in total time considered.

All three moments of simulations are well represented on the features. The major aspect is when the auxiliar stator winding is disconnected and the SPIM continues to run but with other values. The generator doesn't have any major influence on the motor parameters.

2.1.4 DC permanent magnet generator

The parameters monitored on it are: speed wm[rpm], electromagnetic torque Te [Nm], armature current ia [A] and armature voltage U_dc [V]. The values obtained in the considered moments are presented in figure 4.

As it described before, the electric drive is running in no-load conditions till t = 0.5 sec. The moment when the resistance is coupled at the armature winding is very well emphasized on the simulations.

The system running at the same rotor speed at 150 [rad/s].
The output voltage of a DC permanent magnet generator is directly proportional to the speed of the rotor. At higher rotational speeds, more magnetic flux is cut by the stator windings, resulting in a higher voltage.
For small-scale applications, DC permanent magnet generators can be more cost-effective due to their simple construction.

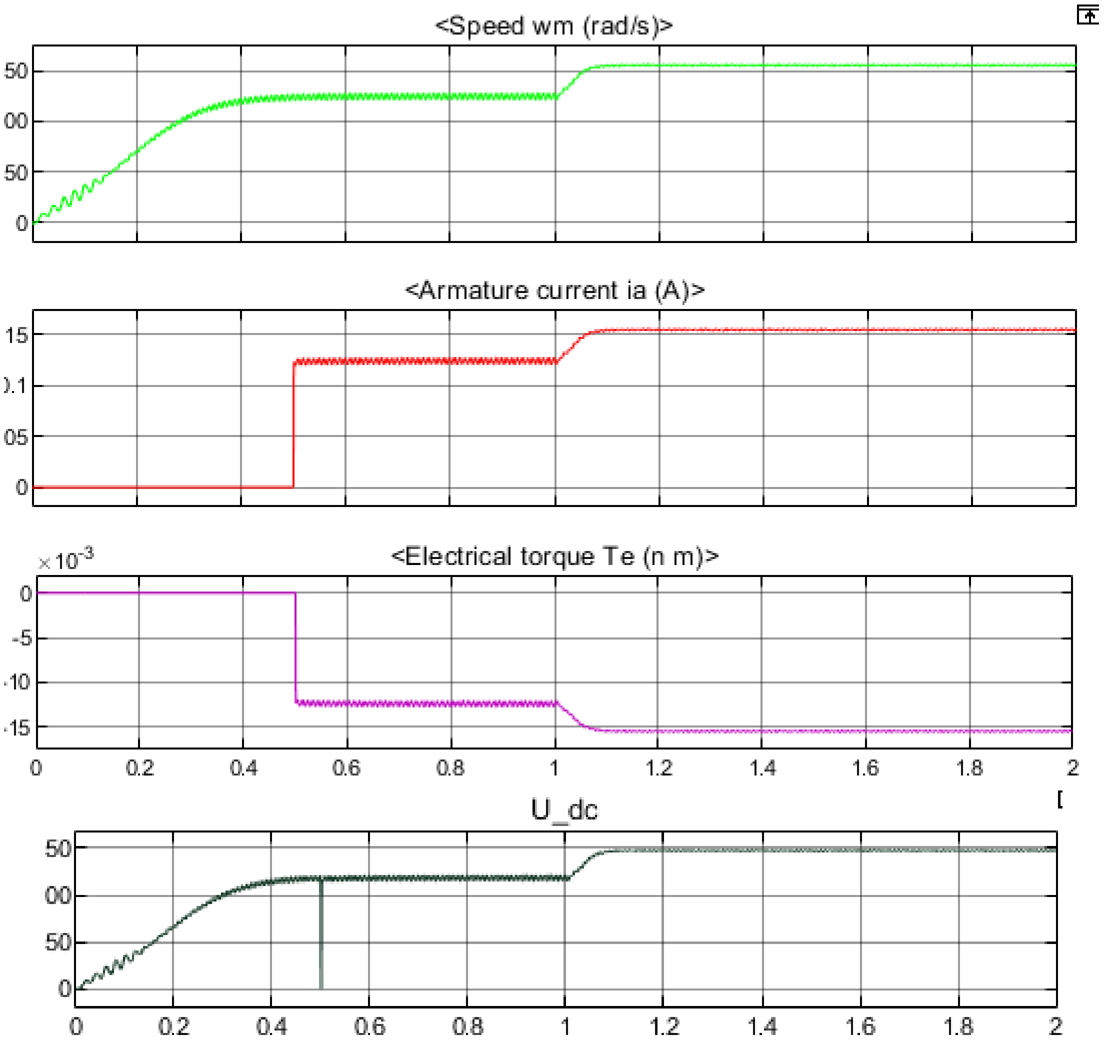


Figure 4 The dc permanent magnet output characteristics

The energetic parameters, power factors and efficiency are presented in Table 1.

Table 1. Table example

Moment of time	T_p [Nm]	Power factor	Efficiency [%]
t = 0.4 sec	20.5	0.96	2.21
t = 0.8 sec	21.5	0.96	2.4
t = 2 sec	6.05	0.23	43.33

If in the first two situations (before and after the connected DC permanent magnet generator) are quite similar, the third situation (when the stator auxiliar winding is disconnected) introduces a big difference between parameters. If the power factor and starting torque is reduced, the efficiency is increased.

3 Conclusions

The application could extend to power other electrical loads, charge batteries, or be fed back into the grid. The generated DC can also be converted to AC if necessary, using an inverter.

The advantages of this electric drive are simplicity and the low cost of single-phase induction motors. Using a structure with permanent magnet develops a process of reliability and low maintenance costs. The system emphasizes the possibility to generate DC power which is easier to store in batteries.

But the system is less efficient than a three-phase motors drive and it is possible to require additional regulation for stable power delivery. Moreover, the power factor and efficiency are not feasible for higher power applications.

Integrating such a system can be a practical solution for specific applications, especially in small-scale power generation and motor-driven systems. To increase efficiency a specific design and control strategies should be implemented.

References

[1] Biro K.A., Viorel I.A., Szabo L., Hennerberger G. *Mașini electrice speciale*, Editura Mediamira, Cluj Napoca, 2005.

[2] Fitzgerald A.E., Kingsley Ch., Jr, Umans S. *Electric Machinery, Sixth Edition,* McGraw-Hill Higher Education, ISBN 0-07-123010-6, Singapore, 2003.

[3] Viorel I.A., Iancu V., Biro K.A., *Determinarea impedanței optime de pornire în monofazata motorului de inducție,* Volumul celui al VI-a Simp.de inf. Și cond., Cluj, 1980, pp. 23-28.

[4] Nica C., Enache S., Enache Monica-Adela – *Mașini Electrice Speciale – Îndrumar de laborator,* Reprografia Universității din Craiova, 2001.

[5] www.mathworks.com

[6] www.mathworks.com - H. Ouquelle and Louis-A.Dessaint (Ecole de technologie superieure, Montreal)

DOI: 10.2478/ijasitels-2024-0005

Smart plants care system using Bluetooth technology

Giorgiu Alexandra-Maria

Abstract

This paper discusses the replacement of manual processes and the way in which the possibility of human errors in plant care is eliminated. The benefits that the intelligent system brings are the reduction of human errors, the reduction of time for plant care such as watering the flowers, checking the humidity of the plants, moving them to a brighter place.

Keywords: system, errors, efficiency, Bluetooth, smart

1 Introduction

Nowadays, the automation and efficiency of intelligent systems is increasingly pursued. This work in turn represents a replacement of manual plant care processes. The realization of this automatic process is done by taking and processing the data by a light detection sensor, a humidity sensor and a water level sensor. The data, once transmitted to a microcontroller, are analyzed in order to make decisions for the constant maintenance of brightness and soil moisture according to certain specified parameters. At the same time, the monitoring of the water level in the tank is done by the level sensor which, in case of emptying the container, will notify the user through a message sent by the microcontroller.

After processing, these data will be displayed to the user through an Android application that can be installed on any smartphone starting from version Android 9.0 or later and that uses Bluetooth technology. The application was created from the desire to facilitate the maintenance of plants, reduce costs by optimizing the consumption of resources, reduce human errors (for example adding too much or not enough water), save time and precisely monitor external factors, thus also having the facility to receive status through a notification to the user with the status of the entire system in real time.

For an easier visualization and for the ease of physical design of the components that serve the work, a 3D prototype was made, with the appearance and approximate dimensions of the smart Plant Care System through Bluetooth technology (Figure 1.0.).

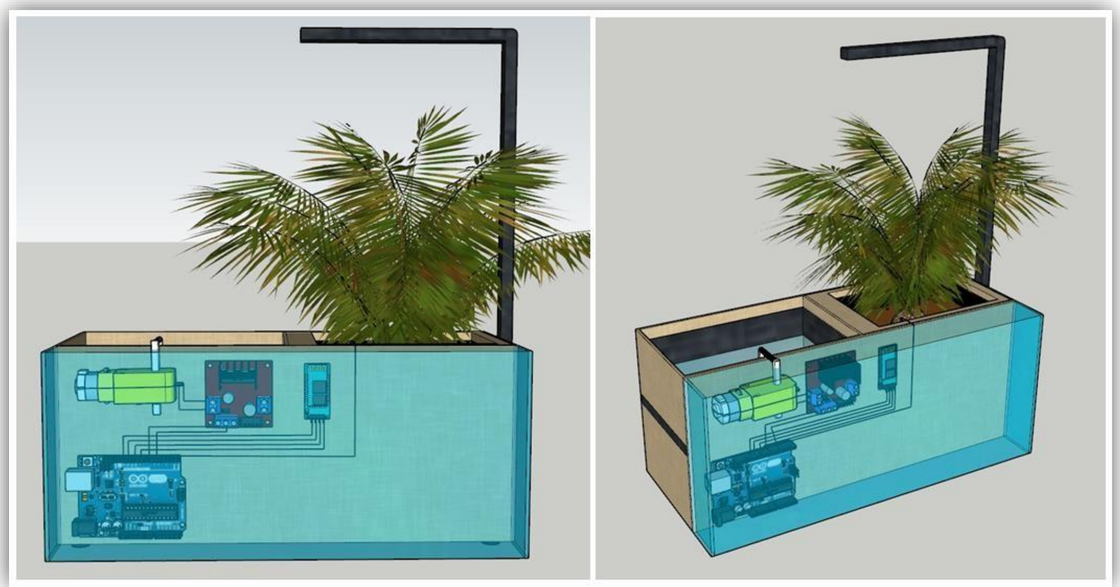


Figure 1.0. Smart plant care system through Bluetooth technology - Prototype in SketchUp

2 Introductory notes

2.1 The .NET platform

.NET is an open-source development platform, a free cross-platform, for making various applications. For this, multiple programming languages, editors and libraries can be used to create web applications, phone applications, desktop applications, games and IoT (applications that use the Internet).

The main solutions offered by the .NET platform are the following:

- Interoperability with existing code
- Full language integration (inheritance support, exception throwing and debugging for multiple programming languages)
- base class library
- Code implementation in a simpler way [1].

Regardless of the programming language supported by this platform, the code can run on any compatible operating system. Different components that help the implementation:

- Cross-platform .NET – for websites, console applications on Windows, servers, macOS, Linux
- NET Framework – supports websites, desktop applications, services
- Xamarin – a .NET implementation for building and running applications on all known mobile operating systems [2].

2.2 Control system with negative feedback loop

In a closed-loop control system the input signal is affected by the output signal. Using this information from the output to modify the system's input signal in some way, a feedback signal is applied to that system. Many times, the applied signal from the system output is compared with a reference input signal, and the result of their difference being used as the input signal on the control system. Often the input reference signal is closely related to the desired output signal, and this being a constant value with respect to time is called the input set point [3].
Such a closed-loop system of a soil moisture control system is shown in figure 1.3.1., where the input data is the optimum moisture, the error is the difference between the optimum soil moisture and the actual recorded value, and the response is given by the loop negative response.

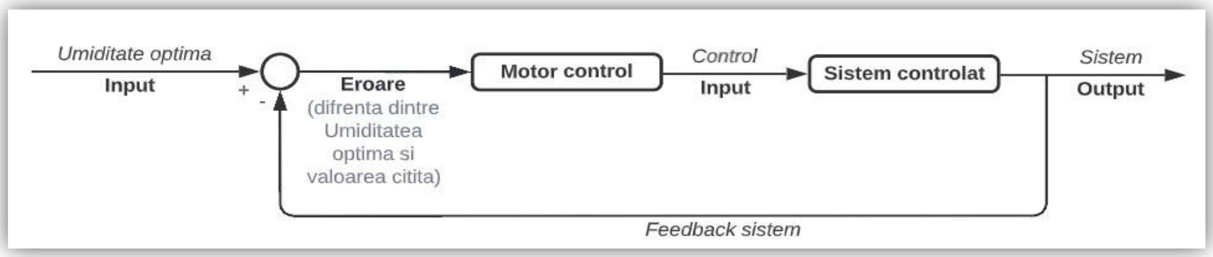


Figure 1.3.1. Closed loop control system

Due to the negative feedback (negative because the output signal is subtracted from the input signal) the accuracy of the system with respect to the desired value is greatly improved compared to the response of an open-loop system, this is because the goal of the control system is to minimize the error between the output and the reference value input.
A disadvantage is that in a closed-loop system oscillations can occur at the output that would not occur in an open-loop system. These oscillations can occur in an attempt to minimize the error as much as possible.

3 System architecture

As part of the project, the system is made up of four major components. The control system part is composed of three modules:

- sensor part (light sensor, water level sensor, soil moisture sensor)
- actuator part (water pump, LEDs)
- the communication part (Bluetooth)

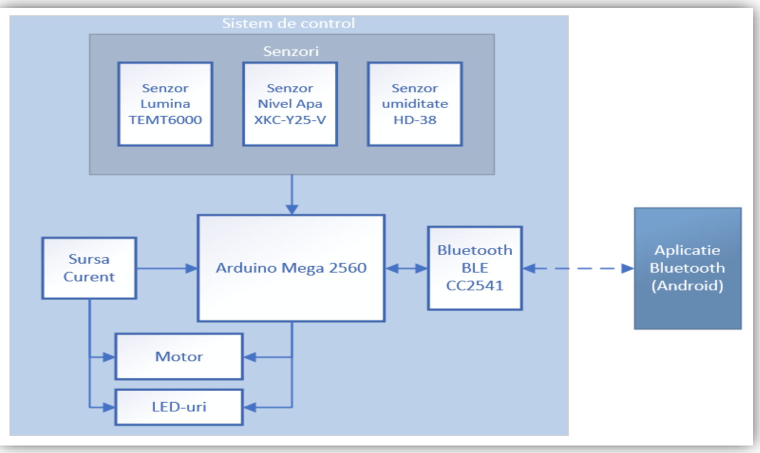


Figure 4.1. System block diagram

The fourth component of the entire system is the mobile application that uses the Android operating system.
The functionality of the whole system has as its first step the reading of data through the sensors, the data is analyzed and processed according to the program with which it is programmed on the Arduino development board and it will make decisions according to certain parameters.

The criteria by which the system will act:

- If the brightness reaches a specified minimum threshold, the LEDs will be on, and when the maximum threshold is reached, they will be off.
- If the soil moisture level falls below the specified limit, the motor will be turned on to pump water for the plant, otherwise, if the moisture level exceeds the set maximum threshold, the motor will be turned off.
- If the water level in the tank is below the set level of the sensor, a message will be generated for the user to add water, otherwise, a status will be displayed to inform the user that the tank is still full.

All this data is checked every 60 seconds to be updated in real time. The control system will send the data to update it within the mobile application as well.

4 Software component

4.1 Algorithm Implementation for Arduino Mega 2560

The code implementation for the Arduino Mega 2560 development board was done in the Arduino IDE 1.8.19. The first part of the program includes the Arduino.h libraries, <WS2812FX.h> [19] and which serve to recognize the internal functions of the Arduino language, namely the LED library.
Also, in the first part of the program, you can find the definition of some variables for the set pins and variables with predefined values that will later be used when comparing with certain values.

To be able to check values for the variable MOTOR_START_COUNTING_5s which will be a flag that will signal the moment when 5 seconds have passed, a boolean type variable (this one having only the value of 0 or 1 - High or Low) will be declared similarly for variables that signal engine that has been running for 60 seconds, LED status, LED pin status. The declaration of such a flag for the example given above is done as follows (similarly for the other flags used):

boolean MOTOR_START_COUNTING_5s LOW;

4.2 Code implementation and structure of the greenApp mobile application

To view the data and how to manage the smart plant care system, a mobile application has been created that can be installed on any smartphone device, starting from an Android version 9.0 or later.

The system is implemented to access data without the need for physical connections, the application using Bluetooth technology for short distance control, where the communication protocol used to transmit and receive data is called Bluetooth 4.0.

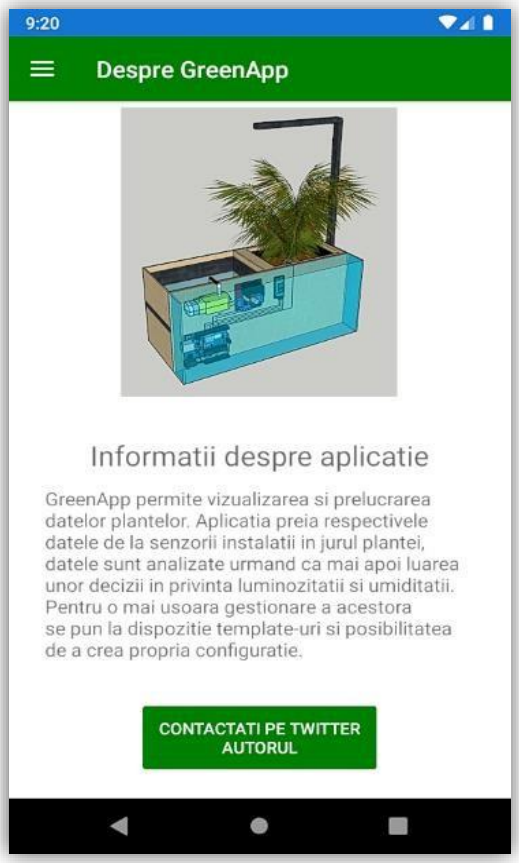
The greenApp application was implemented with the aim of facilitating the maintenance of plants that have specific moisture and light requirements. For example, for a cactus, the brightness level must be increased, and the humidity level must be very low, whereas for an orchid, the brightness level can be average, but the soil humidity level must be very high.

Another advantage that greenApp brings is the reduction of costs by optimizing the use of resources and reducing human errors. An error could be that when we water the plants, we can add much more than needed or not enough water, this problem is solved by the control system that constantly monitors the plant through sensors and makes intelligent decisions so that the level of brightness and humidity will always be the most optimal.

A user warning system has been implemented for the situation where during plant monitoring the water level in the tank is detected as almost empty. The user will be alerted by a message in the water level status section.

The system status is monitored in real time; thus, a data update is performed every 60 seconds.

Open page from the first button shows us the name of the accessed page in the top bar, a suggestive title in the middle of the screen, the text with the description of the application and an image of the device. (Figure 5.1.)



Informatii despre aplicatie

GreenApp permite vizualizarea si prelucrarea datelor plantelor. Aplicatia preia respectivele datele de la senzorii instalatii in jurul plantei, datele sunt analizate urmand ca mai apoi luarea unor decizii in privinta luminozitatii si umiditatii. Pentru o mai usoara gestionare a acestora se pun la dispozitie template-uri si posibilitatea de a crea propria configuratie.

Figure 5.1. Application description page

5 Conclusions

In the framework of this work, we created a smart plant care system through Bluetooth technology with the aim of finding a way to replace manual processes and eliminate human errors. The fact that we forget to do certain things, like watering the plants is perfectly normal, but why not find a way in which we can save both time and resources.

The benefits that the smart system brings can be: reducing or even eliminating the time lost to do repetitive operations such as watering flowers, reducing costs by using the smart system that "knows" how to use the necessary amount of water for each individual plant, increasing the quality of the results by finding the most optimal solution and eliminating the possibility of "drowning" of plants or their drying due to lack of water.

The functionality of the system is to take data such as soil moisture and brightness through sensors and process them in order to make decisions. The system also has a water level sensor that verifies in real time the state of the liquid in the container. If the water level drops to the set limit, the user will be warned through the application. These decisions are made based on certain parameters entered and ensure the most efficient use of resources.

Another advantage is that the system is independent of physical connections, real-time monitoring of the data is achieved through a mobile application that connects via Bluetooth to an Arduino Mega 2560 board. The communication protocol is Bluetooth 4.0, and the mobile application is compatible with smartphones s that have an Android 9.0 or newer operating system. For the future project, the next step is to implement a new communication technology such as Wi-Fi and store user data in a database to create graphics that can be generated based on user configurations and preferences. Another method of improving the whole system is to mount a camera for take daily pictures and add them to a gallery.

References

[1] Example of integrated systems, https://www.bartleby.com/subject/engineering/computer-science/concepts/embedded-software-development

[2] Michael J. Pont, *Embedded C*, Great Britain, 2002.

[3] Block diagram for timer, https://embedjournal.com/timer-modules-guide/.

[4] TUTORIAL ON SUCCESSIVE APPROXIMATION REGISTERS, https://www.maximintegrated.com/en/design/technical-documents/tutorials/1/1080.html

[5] Basics of PWM (Pulse Width Modulation), https://docs.arduino.cc/learn/microcontrollers/analog-output

[6] Kevin Warwick, *An introduction to control systems*, World Scientific Publishing Co. Pte.Ltd., 1996.

[7] Control system engineering-I, https://www.vssut.ac.in/lecture-notes.php?url=electrical-engineering

[8] I.J. Nagrath, *Control Systems Engineering*, New Age International Pvt, 2005.

[9] Documentation, https://www.arduino.cc/en/Guide/Introduction

[10] Arduino Uno SMD R3, https://en.wikipedia.org/wiki/Arduino#/media/File:Arduino_Uno_-_R3.jpg.

[11] Arduino, https://en.wikipedia.org/wiki/Arduino

[12] Bluetooth module, https://roboteh.ro/13-modul-bluetooth-40-ble-cc2540-cc2541-hm10-1000012.html

[13] Bluetooth module 4.0 BLE CC2541 HM10 Ibeacon, https://roboteh.ro/13-modul-bluetooth-40-ble-cc2540-cc2541-hm10-1000012.html

[14] Logic Level Translator Module, https://www.sigmanortec.ro/Modul-Translator-nivel-logic-I2C-IIC-bidirectional-4-canale-3-3V-p126421751

[15] Ambient light sensor TEMT6000, https://ardushop.ro/ro/home/1290-senzor-lumina-ambientala-temt6000.html

[16] Neopixel bar WS2812, https://ardushop.ro/ro/home/912-bara-de-neopixeli-ws2812-cu-8-led-uri.html

[17] Liquid level sensor, without contact, XKC-Y25-V, https://www.emag.ro/senzor-nivel-lichid-fara-contact-xkc-y25-v-ai248-s314/pd/DYHQ33MBM/#product-gallery

DOI: 10.2478/ijasitels-2024-0006

Vacation Planning Application using API

Gulean Paula-Florina

Computer Science and Electrical and Electronics Engineering Department,
Faculty of Engineering, "Lucian Blaga" University of Sibiu, Romania
paula.gulean@ulbsibiu.ro

Abstract

Choosing and booking a vacation can be difficult when the application where these things are done is not user-friendly. Therefore, the application that I have created will have to make the user experience much more pleasant. The purpose of the application is to be able to choose a holiday according to the desired filters, with the option to search for a flight in the same period. After booking it, the user who created an account will have stored in his account all the reservations made. Using the logic behind this application, any travel agency can create a new interface that they can customize according to their wishes, but the functionalities remain the same. They can also change what data they want to receive from the API, what data they want to store in their database and in addition modify the offers according to their wishes.

Keywords: application, API, MVC

1. Introduction

API or Application Programming Interface is a concept to develop applications. With its help, two or more applications can communicate bidirectionally. For example, using the logic behind this vacation planning application and changing only the interface, it can be used by several travel agencies. This greatly facilitates the work of programmers. The implemented API architecture was REST which was introduced in 2004 [1]. The calls that make up the API of the REST architecture are also known as endpoints. In the REST architecture, at an endpoint it is allowed to find only one resource (a user, a schedule list, a hashmap of bank accounts, etc.). An endpoint is characterized by a URI type identifier.[2]

For example : /api/users/ -- this endpoint descriptively specifies that I expect the resource provided by the API to be a list of users

/api/users/1 // -- fetch the user with id 1

Among the purposes of an API, there is also the hiding of the internal information of the way a system works.

Web APIs are a service accessed from the client to a web server via the optional Hypertext Transfer Protocol (HTTP). Clients send a request in the form of an HTTP request (REQUEST) and receive a response (RESPONSE), JavaScript Object Notation (JSON) or Extensible Markup Language (XML – no more practiced) format. Developers typically use APIs in querying a server for a specific set of data on that server.

The word API is often used in reference to web APIs, approving the connection between computers. There are also APIs for programming languages, software libraries, operating systems, and computer hardware. The origins of APIs date back to the 1940s, but the use of the term only appeared in the 1960s and 1970s. Recent developments in the use of APIs have led to the rise in popularity of microservices, which are ultimately loosely coupled services accessed through public APIs.[3]

For this project, I chose the Amadeus API for bringing information. Amadeus contains a selection of APIs for searching, booking and inspiring a trip. By calling endpoints from Amadeus, I was able to load hotel data into the database. To get a key and make an API call, you need to create an account. This API works via a key and secret pair. This data pair is regularly used in API application to obtain an application access token. This principle stops any bot (automated program) from "overloading" the API with unnecessary requests, because each access token must be requested again after a period of time (it has an expiration time usually 30-60 min).[4]

2 Theoretical considerations

2.1 Development environments

The application was developed in the Visual Studio 2019 programming environment using the C# programming language. This is an integrated programming environment that can be used to create new programs, web or mobile applications.

.Net Framework is a software development environment made by Microsoft. It contains a library of classes called the Framework Class Library (FCL) and provides increased flexibility because each language can use code written in another language.[5] The Framework Class libraries provide the interface part, data access, database connectivity, cryptography, building WEB applications, numerical algorithms and network communications.

When creating an application, a developer combines their source code with the one from the .Net Framework. Entity framework is a framework for Microsoft .NET applications. It allows developers to work with class-specific objects without focusing on the tables where the data is stored.

2.2 Programming languages

The programming language used, called C#, is a very well-known and used language by most programmers. It is based on the C language.

HTML or Hyper Text Markup Language is the standard language for creating and structuring web pages. This language comprises a string of elements that tell the browser how the content needs to be designed. The elements are set with the help of tags, which consist in writing them between „<" „>". The name of the element inside a tag is case sensitive, that is, it cannot be written in uppercase letters.[6]

CSS or Cascading Style Sheets describes how HTML elements will be displayed on screen. It saves a lot of time because once created, a CSS file can be used by several web pages at the same time. CSS describes page layout, design and screen size variations.

JSON, i.e. JAVASCRIPT OBJECT NOTATION, is a format that makes it much easier to represent data structures. Json is a much simpler variant than XML. It is based on a subset of the ECMA-262 JavaScript programming language standard. Json is a format that is language-independent, but uses conventions familiar to programmers in the C family of languages. An object is an unordered set of name/value pairs. An object starts with a left brace and ends with a right brace. Each name is followed by a "colon" then the value, and name/value pairs are separated by a comma.[7]

JavaScript (JS) is an object-oriented programming language based on the concept of prototypes. It is mainly used to introduce functionality into web pages, the JavaScript code in these pages being run by the browser. The language is well known for its use in building websites, but it is also used for accessing embedded objects in other applications. It was originally developed by Brendan Eich of Netscape Communications Corporation under the name Mocha, then LiveScript, and finally called JavaScript.[8]

2.3 Used during implementation

MVC (Model-View-Controller) is a software architecture model used to implement large-scale projects that can be scaled vertically (improving the current product) and horizontally (generating replicas with the same specifications). As its name suggests, the logic of the program is divided into three interconnected elements.[9]
User requests are directed to the controller. It must communicate with the model to execute and send responses to user requests. The controller orders to display on the view for the user the data he needs.
Components:

- Model- represents the logical scheme of a resource o Image of a table in the database
  - image of a response between client server etc.

- View- has the role of characterizing how the data should be displayed on the page (it doesn't care what the data is)
- Controller- is the one that makes the connection between the model and the view, also having the role of request validator.

Data Transfer Objects

DTOs or Data Transfer Objects are objects that transport data between processes in order to reduce the number of method calls. Martin Fowler is the one who implemented this kind of concept. He explained that the purpose of the pattern is to reduce the number of calls to the server by joining multiple parameters into a single call. Also, this principle was introduced to standardize the role of the model as a single SOT (Source of Truth) for the database, thus standardizing the fact that the model represents the outline of a table in the database.[10]

Readers

This principle is characterized not so much by design patterns as by the observance of SOLID principles[11]:

- The Single-responsibility principle: “There should never be more than one reason for a class to change.”

● The Open–closed principle : “Software entities ... should be open for extension, but closed for modification.”
● The Liskov substitution principle: "Functions that use pointers or references to base classes must be able to use objects of derived classes without knowing it."
● The Interface segregation principle: "Clients should not be forced to depend upon interfaces that they do not use."
● The Dependency inversion principle: "Depend upon abstractions, [not] concretes."

With the implementation of this principle, all the logic necessary for reading from files was encapsulated in a class, further using an own templating mechanism for generating offers for hotels

Services

Classes of this type have the role of providing an encapsulation layer of access to database operations. These classes are the only classes that can provide access to read\write operations thus allowing us to have a single source of truth (single SOT) for the data provided by our database. In Fig. 1 you can see the application of this principle, the actors involved in the communication with the database actually use a service to always get the data, without taking into account exactly the logic of communication with the database. This service actually uses inside the principle called Facade.

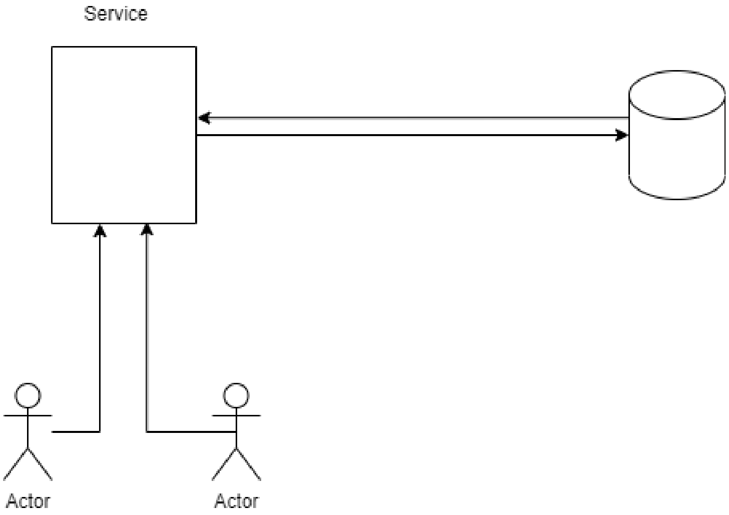


Figure 1 - Services

Templating

This is not necessarily a principle so much as a reference to an idea of structuring structural similarities and generating concrete examples based on them.
In the case of my application, I developed an algorithm that, based on some templates, will generate the offers dynamically. Illustrating this facility, I have generated JSON templates for the key features of an offer and JSON templates for the facilities of a location, these templates can be further extended for easy application development.

Mappers

Classes of this type allow the extraction of concrete attributes from a data structure with a very large variety of attributes. These classes are meant to allow simplistic implementation at a small development level, allowing the application implementation to be extended by adding new attribute support to the resulting class.
An example can be described by mapping objects in the database to objects of the type of interface that the user sees.

Security

To maintain security, I have implemented a security method by which the password is encrypted in the database with an irreversible algorithm (PassWord Bcrypt), using dynamic padding (move the entered password with n random characters to the left or right) and a single jump (a character string that is added to the password to ensure greater dynamism).
Thus, in order to check the integrity of the password saved in the database, the hashes of these passwords will be checked to match.

3 Implementation

For the implementation I will present the most important things that were implemented using the theoretical information from previous chapter.

3.1 Database

During the development of the application, I decided to use a Mysql database, because it facilitates the relationships between the tables, allows a quick access to the data and accurate filters. In the figure below you can see the structural diagram of the application’s database.

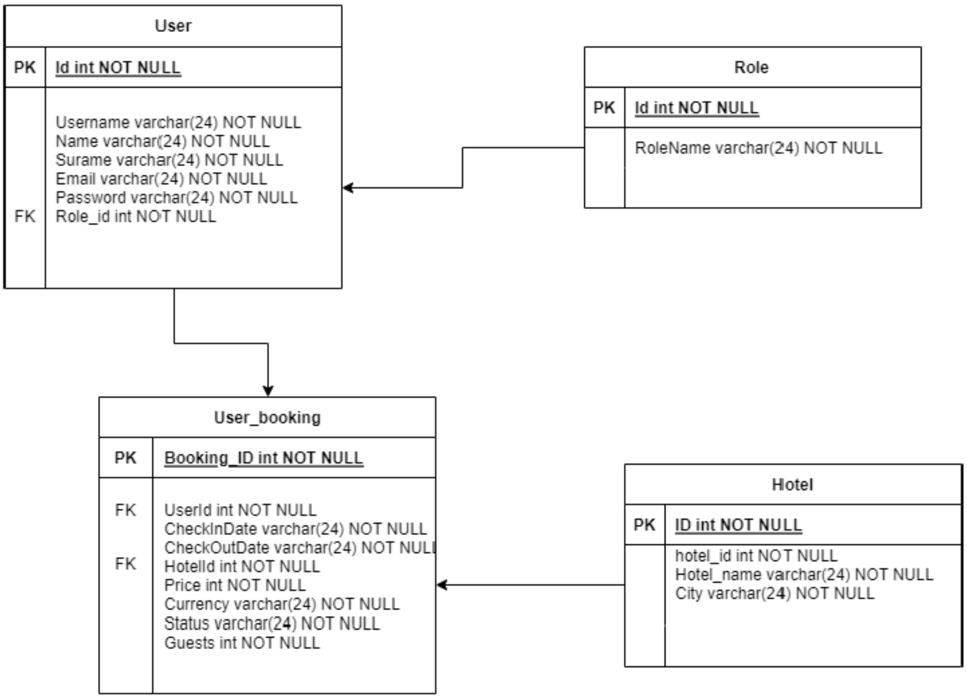

User
PK
Id int NOT NULL
Username varchar(24) NOT NULL
Name varchar(24) NOT NULL
Surame varchar(24) NOT NULL
Email varchar(24) NOT NULL
Password varchar(24) NOT NULL
FK
Role_id int NOT NULL
Role
PK
Id int NOT NULL
RoleName varchar(24) NOT NULL
User_booking
PK
Booking_ID int NOT NULL
FK
UserId int NOT NULL
CheckInDate varchar(24) NOT NULL
CheckOutDate varchar(24) NOT NULL
FK
HotelId int NOT NULL
Price int NOT NULL
Currency varchar(24) NOT NULL
Status varchar(24) NOT NULL
Guests int NOT NULL
Hotel
PK
ID int NOT NULL
hotel_id int NOT NULL
Hotel_name varchar(24) NOT NULL
City varchar(24) NOT NULL

Figure 2 - Database

3.2 Architecture of services

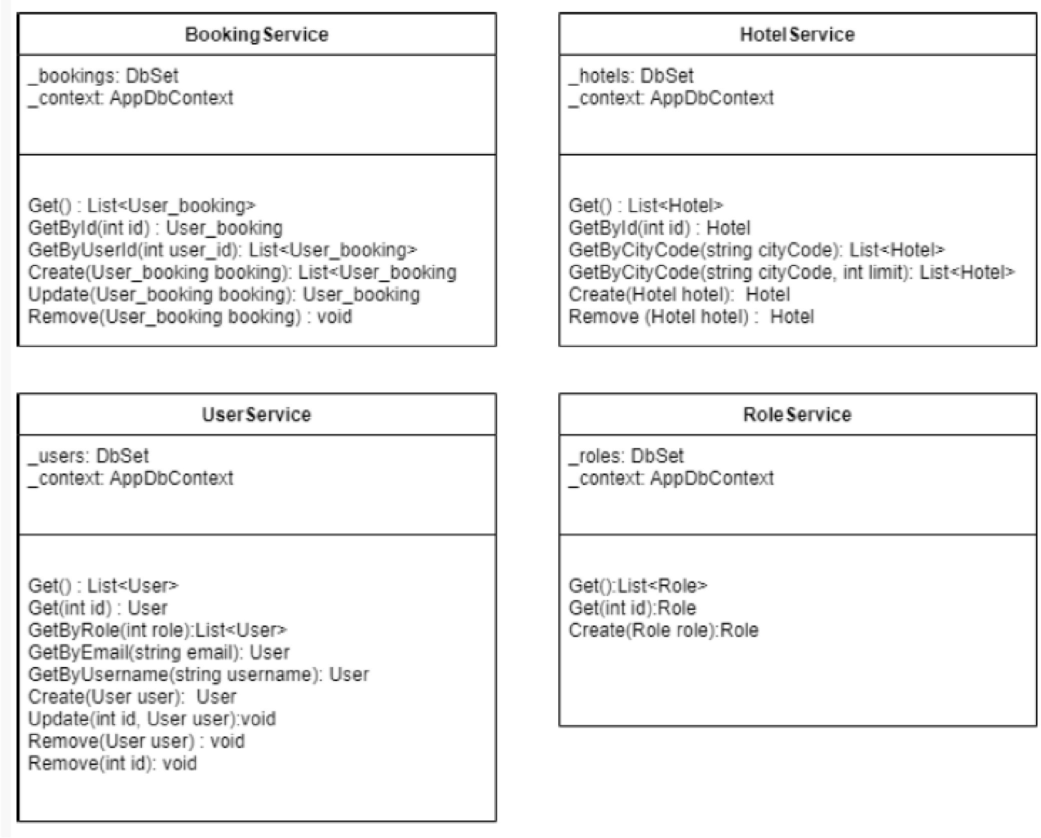
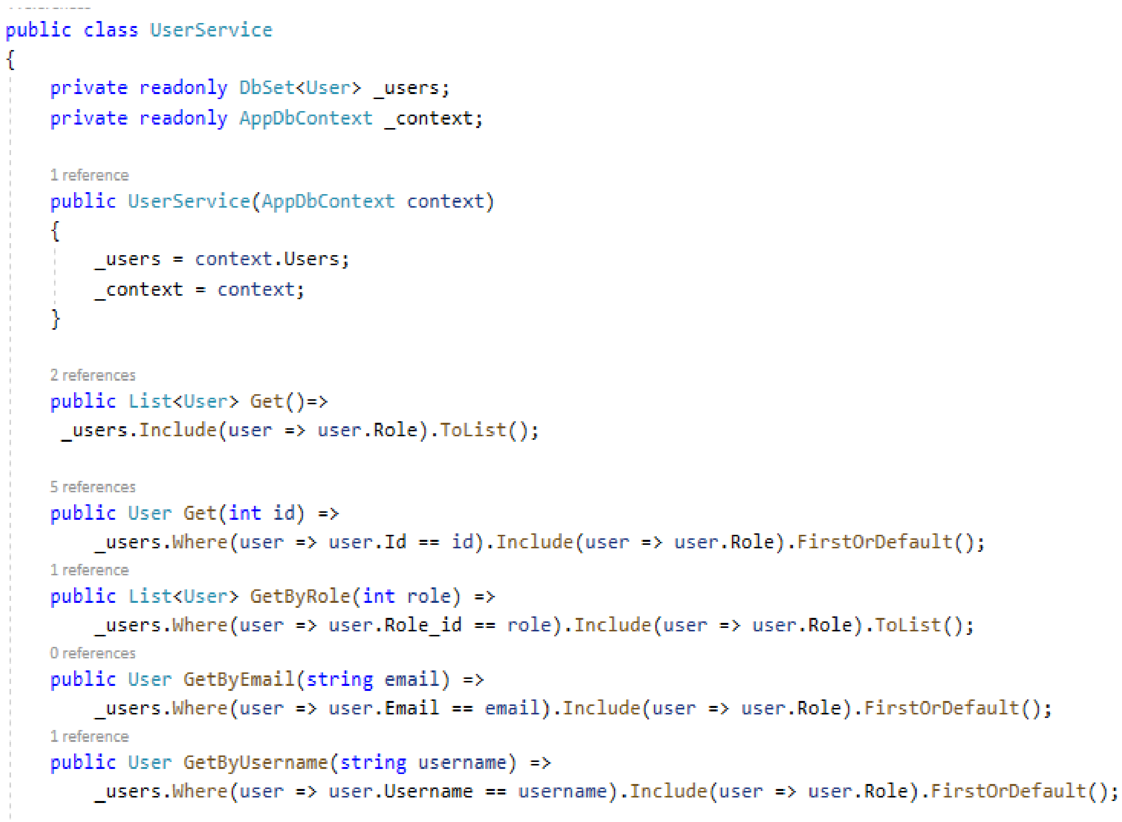

BookingService
_bookings: DbSet
_context: AppDbContext
Get() : List<User_booking>
GetById(int id) : User_booking
GetByUserId(int user_id): List<User_booking>
Create(User_booking booking): List<User_booking
Update(User_booking booking): User_booking
Remove(User_booking booking) : void
HotelService
_hotels: DbSet
_context: AppDbContext
Get() : List<Hotel>
GetById(int id) : Hotel
GetByCityCode(string cityCode): List<Hotel>
GetByCityCode(string cityCode, int limit): List<Hotel>
Create(Hotel hotel): Hotel
Remove (Hotel hotel) : Hotel
UserService
_users: DbSet
_context: AppDbContext
Get() : List<User>
Get(int id) : User
GetByRole(int role):List<User>
GetByEmail(string email): User
GetByUsername(string username): User
Create(User user): User
Update(int id, User user):void
Remove(User user) : void
Remove(int id): void
RoleService
_roles: DbSet
_context: AppDbContext
Get():List<Role>
Get(int id):Role
Create(Role role):Role

Figure 3 - Services

For example, in Fig. 4 is presented the UserService. It can be seen that it holds a member called users which actually represents the table rows in the database and a context, which is used to update the information in real time. This context will be saved for the safety of keeping the information at its most recent version.
The constructor injects the database context into the current service. Access to the users table will be taken from this context and put into the Users variable, and the context will be saved into the context variable to keep the data up to date.

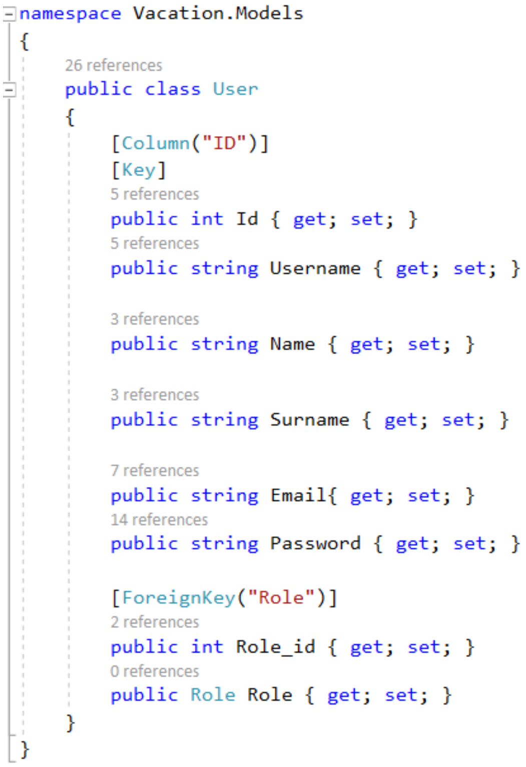


```
public class UserService
{
    private readonly DbSet<User> _users;
    private readonly AppDbContext _context;

    1 reference
    public UserService(AppDbContext context)
    {
        _users = context.Users;
        _context = context;
    }

    2 references
    public List<User> Get()=>
     _users.Include(user => user.Role).ToList();

    5 references
    public User Get(int id) =>
        _users.Where(user => user.Id == id).Include(user => user.Role).FirstOrDefault();
    1 reference
    public List<User> GetByRole(int role) =>
        _users.Where(user => user.Role_id == role).Include(user => user.Role).ToList();
    0 references
    public User GetByEmail(string email) =>
        _users.Where(user => user.Email == email).Include(user => user.Role).FirstOrDefault();
    1 reference
    public User GetByUsername(string username) =>
        _users.Where(user => user.Username == username).Include(user => user.Role).FirstOrDefault();
```

Figure 4 - UserService

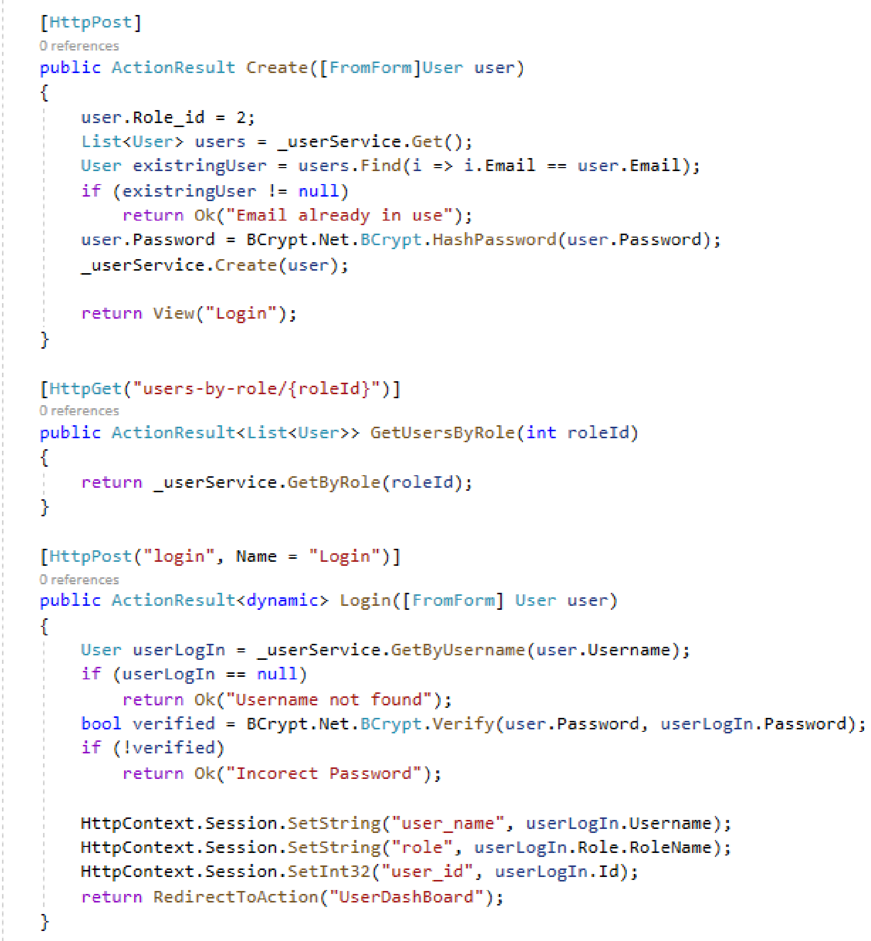
3.3 MVC implementation

In the following figures (Fig.5,6,7) I will present the Model, Controller and View for User. All others MVC were represented also using this logic behind.



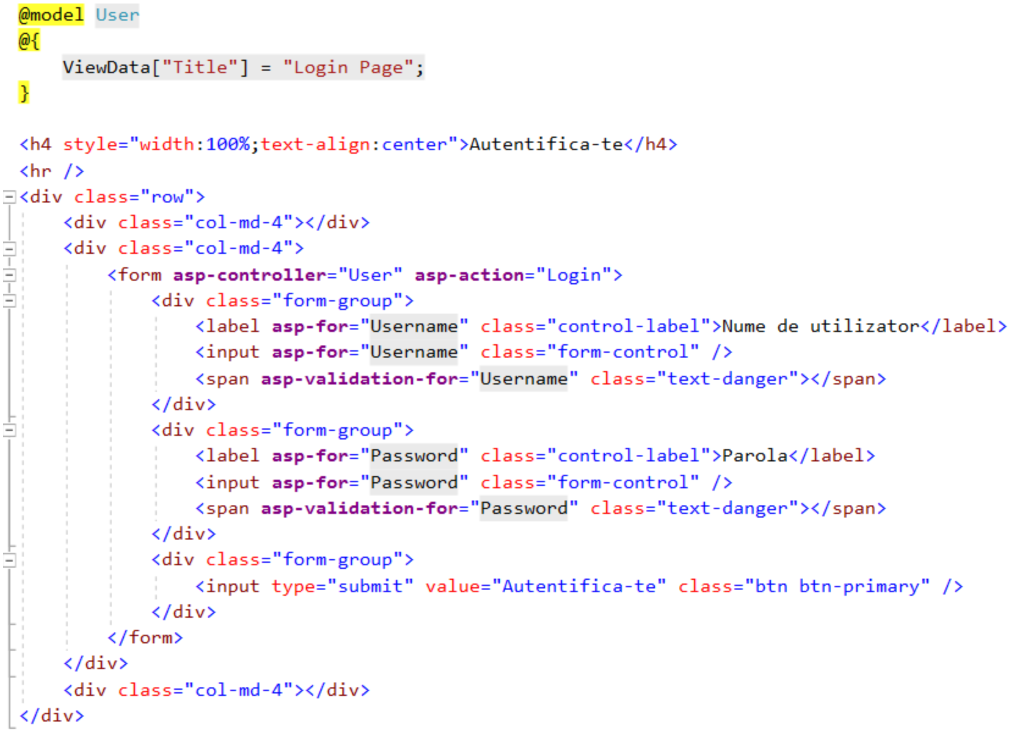
```
namespace Vacation.Models
{
    26 references
    public class User
    {
        [Column("ID")]
        [Key]
        5 references
        public int Id { get; set; }
        5 references
        public string Username { get; set; }

        3 references
        public string Name { get; set; }

        3 references
        public string Surname { get; set; }

        7 references
        public string Email{ get; set; }
        14 references
        public string Password { get; set; }

        [ForeignKey("Role")]
        2 references
        public int Role_id { get; set; }
        0 references
        public Role Role { get; set; }
    }
}
```

Figure 5 – Model



```
    [HttpPost]
    0 references
    public ActionResult Create([FromForm]User user)
    {
        user.Role_id = 2;
        List<User> users = _userService.Get();
        User existringUser = users.Find(i => i.Email == user.Email);
        if (existringUser != null)
            return Ok("Email already in use");
        user.Password = BCrypt.Net.BCrypt.HashPassword(user.Password);
        _userService.Create(user);

        return View("Login");
    }

    [HttpGet("users-by-role/{roleId}")]
    0 references
    public ActionResult<List<User>> GetUsersByRole(int roleId)
    {
        return _userService.GetByRole(roleId);
    }

    [HttpPost("login", Name = "Login")]
    0 references
    public ActionResult<dynamic> Login([FromForm] User user)
    {
        User userLogIn = _userService.GetByUsername(user.Username);
        if (userLogIn == null)
            return Ok("Username not found");
        bool verified = BCrypt.Net.BCrypt.Verify(user.Password, userLogIn.Password);
        if (!verified)
            return Ok("Incorect Password");

        HttpContext.Session.SetString("user_name", userLogIn.Username);
        HttpContext.Session.SetString("role", userLogIn.Role.RoleName);
        HttpContext.Session.SetInt32("user_id", userLogIn.Id);
        return RedirectToAction("UserDashBoard");
    }
```

Figure 6 - Controller

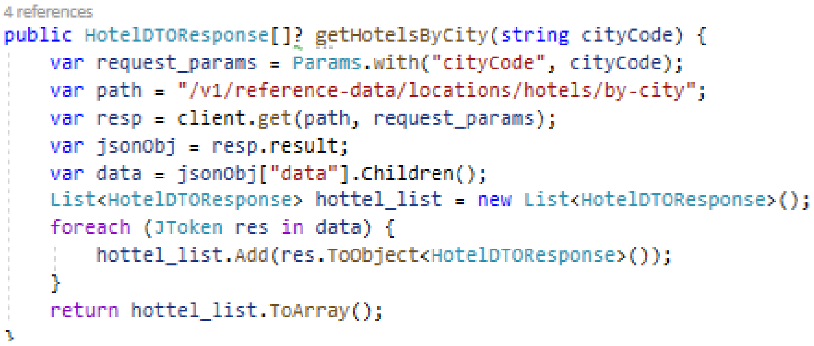


```
@model User
@{
    ViewData["Title"] = "Login Page";
}

<h4 style="width:100%;text-align:center">Autentifica-te</h4>
<hr />
<div class="row">
    <div class="col-md-4"></div>
    <div class="col-md-4">
        <form asp-controller="User" asp-action="Login">
            <div class="form-group">
                <label asp-for="Username" class="control-label">Nume de utilizator</label>
                <input asp-for="Username" class="form-control" />
                <span asp-validation-for="Username" class="text-danger"></span>
            </div>
            <div class="form-group">
                <label asp-for="Password" class="control-label">Parola</label>
                <input asp-for="Password" class="form-control" />
                <span asp-validation-for="Password" class="text-danger"></span>
            </div>
            <div class="form-group">
                <input type="submit" value="Autentifica-te" class="btn btn-primary" />
            </div>
        </form>
    </div>
    <div class="col-md-4"></div>
</div>
```

Figure 7 - View

3.4. API

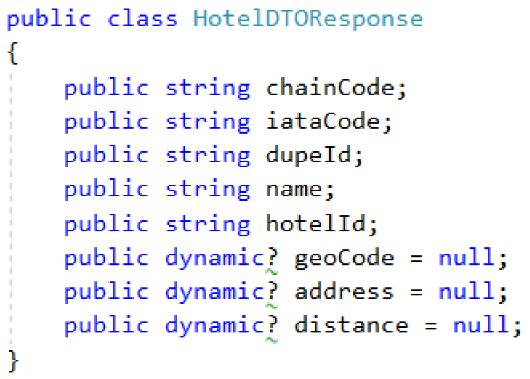
This class is used as a communication medium between the developed application and the API implemented by Amadeus.In the constructor of this class it is observed that with the help of the key and secret mechanism Amadeus.builder ("AoWlJ6f5AMkZYcbBAxlrFTSt54uGGheK", "S1apMroKeQ8Bgnml")), a connection to Amadeus will be made. This key and secret mechanism works in the same way as an email and a password, only when authenticating through the client with a key and secret, you will receive back a string of characters that represents the "password" to access Amadeus. This access password is also accompanied by an expiration time of ~30 min. That is, in other words, when it is desired to use Amadeus, the client variable checks if it has a valid Amadeus access password, if it does not have that password or if it has expired, a new one will be requested. Then allowing other methods from the Amadeus API to be called.



```
4 references
public HotelDTOResponse[]? getHotelsByCity(string cityCode) {
    var request_params = Params.with("cityCode", cityCode);
    var path = "/v1/reference-data/locations/hotels/by-city";
    var resp = client.get(path, request_params);
    var jsonObj = resp.result;
    var data = jsonObj["data"].Children();
    List<HotelDTOResponse> hottel_list = new List<HotelDTOResponse>();
    foreach (JToken res in data) {
        hottel_list.Add(res.ToObject<HotelDTOResponse>());
    }
    return hottel_list.ToArray();
}
```

Figure 8 - getHotelsByCity method

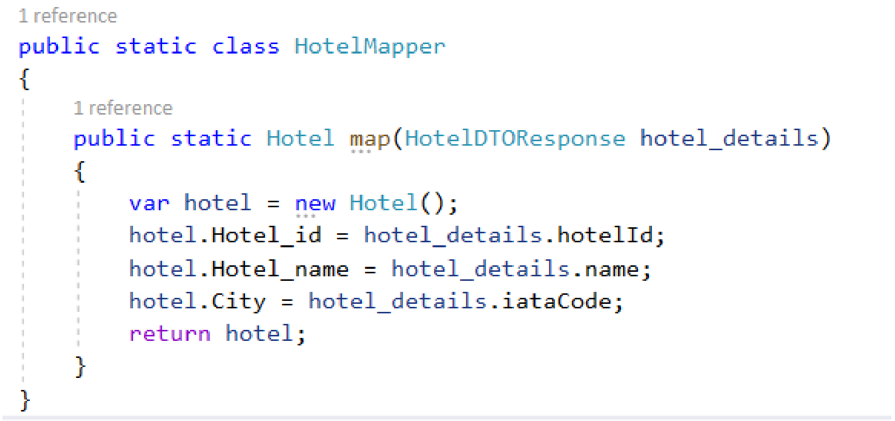
The getHotelsByCity method (Fig.8) has the role to fetch hotels via the Amadeus API based on a city. In this method, it is set in the request_params variable, in which city to search for hotels, then a request will be sent to "/v1/reference-data/locations/hotels/by-city" "to get this list.Once a response is received from Amadeus, it will take that response from the data variable and iterate over it to convert the data, which is passed in JSON format, into the format of the HotelDTOResponse class.



```
public class HotelDTOResponse
{
    public string chainCode;
    public string iataCode;
    public string dupeId;
    public string name;
    public string hotelId;
    public dynamic? geoCode = null;
    public dynamic? address = null;
    public dynamic? distance = null;
}
```

Figure 9 - HotelDTOResponse class

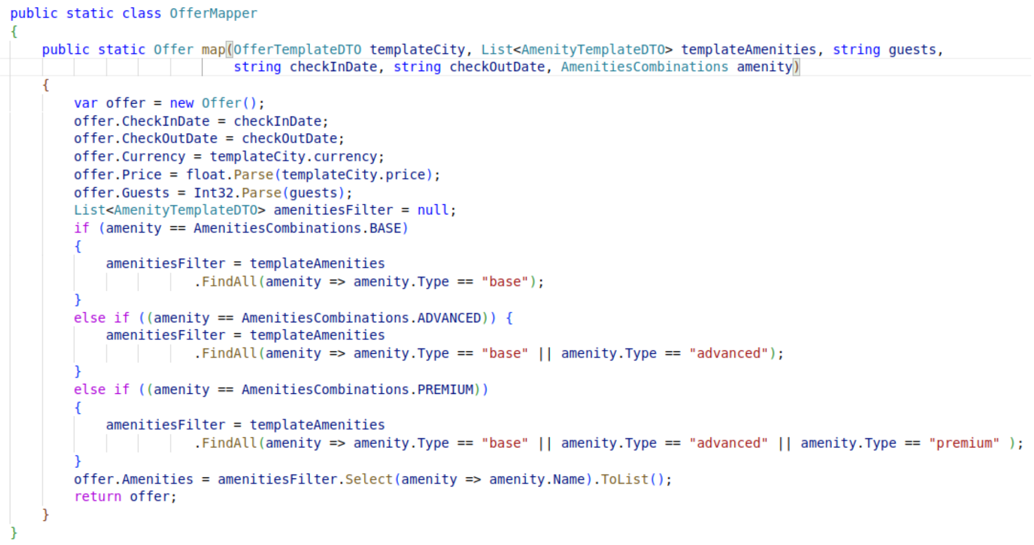
For my application, I needed the available hotels in a city. For this I called the endpoint "/v1/reference-data/locations/hotels/by-city" which, depending on the city code provided, brings the required information. These will bring more information about each hotel in the city I searched for. For the implemented application, I chose to use only a part of this information stored with the help of the API and with the help of a mapper, I kept only the desired information.



```
1 reference
public static class HotelMapper
{
    1 reference
    public static Hotel map(HotelDTOResponse hotel_details)
    {
        var hotel = new Hotel();
        hotel.Hotel_id = hotel_details.hotelId;
        hotel.Hotel_name = hotel_details.name;
        hotel.City = hotel_details.iataCode;
        return hotel;
    }
}
```

Figure 10 - HotelMapper class

In Fig. 10 you can see the implementation for the HotelMapper class. This class contains the map method that receives as a parameter an object of type HotelDTOResponse. In the body of this method, an object of the Hotel type will be instantiated and the necessary attributes will be set, and then that object will be returned.



```
public static class OfferMapper
{
    public static Offer map(OfferTemplateDTO templateCity, List<AmenityTemplateDTO> templateAmenities, string guests,
                            string checkInDate, string checkOutDate, AmenitiesCombinations amenity)
    {
        var offer = new Offer();
        offer.CheckInDate = checkInDate;
        offer.CheckOutDate = checkOutDate;
        offer.Currency = templateCity.currency;
        offer.Price = float.Parse(templateCity.price);
        offer.Guests = Int32.Parse(guests);
        List<AmenityTemplateDTO> amenitiesFilter = null;
        if (amenity == AmenitiesCombinations.BASE)
        {
            amenitiesFilter = templateAmenities
                        .FindAll(amenity => amenity.Type == "base");
        }
        else if ((amenity == AmenitiesCombinations.ADVANCED)) {
            amenitiesFilter = templateAmenities
                        .FindAll(amenity => amenity.Type == "base" || amenity.Type == "advanced");
        }
        else if ((amenity == AmenitiesCombinations.PREMIUM))
        {
            amenitiesFilter = templateAmenities
                        .FindAll(amenity => amenity.Type == "base" || amenity.Type == "advanced" || amenity.Type == "premium" );
        }
        offer.Amenities = amenitiesFilter.Select(amenity => amenity.Name).ToList();
        return offer;
    }
}
```

Figure 11 - OfferMapper class

In the figure illustrated above (Fig. 11) you can see the implementation of the OfferMapper class. This class contains the map method that receives as parameters the details necessary to generate an offer. In the body of the method, an object of type Offer will be instantiated and members of this instance will be set. At the end, the instance of the generated offer will be returned.

4. Conclusions

The application allows choosing and booking a holiday, letting the user choose their favorite destination, the desired period and the number of accompanying people. In addition to this, after logging in, he is able to see all his created bookings, from his account.

The administrator can see the bookings created by any user at any time.

Another useful thing is the fact that hotel data is brought via an API.

The application can be developed in the future by adding an option (a check-box) that by ticking will also bring offers for air transport to the desired destination. Thus, at the end the user will receive a complete offer, which he can reserve.

Another thing that could be implemented is the addition of a possibility to pay by card, so the administrator will receive a confirmation of the reservation created, by paying the customer. Also, a useful thing could be receiving a holiday confirmation on the user's email.

Because the information is fetched with the API, we can always change what we want to display depending on what information we can get through it.

References

[1] REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces, Mark Masse, O'Reilly Media, Inc., 2011
[2] https://en.wikipedia.org/wiki/Uniform_Resource_Identifier
[3] https://en.wikipedia.org/wiki/API
[4] https://developers.amadeus.com/
[5] https://en.wikipedia.org/wiki/.NET_Framework
[6] https://developer.mozilla.org/en-US/docs/Web/HTML

[7] https://www.json.org/json-ro.html
[8] https://en.wikipedia.org/wiki/JavaScript
[9] ASP.NET MVC Framework Unleashed, Stephen Walther,Sams Publishing, 2009
[10] https://www.baeldung.com/java-dto-pattern
[11] https://en.wikipedia.org/wiki/SOLID

DOI: 10.2478/ijasitels-2024-0004

AUTOMATIC PART-OF-SPEECH TAGGING

Adelina Manolea[1],

[1]*master student in the Embedded Systems program, Faculty of Engineering, "Lucian Blaga" University of Sibiu, Romania*

adelina.manolea@ulbsibiu.ro

Abstract

Natural language processing (NLP) is a key technique in Business Process Management (BPM). The performance of BPM methods, which are based on NLP, is limited by the accuracy of automatic part-of-speech tagging, a base subtask of NLP.[9] The automatic part-of-speech tagging is the process of assigning a tag to every word in a text or a document.[1] I have developed and presented in this paper an application that learns to correctly predict parts-of-speech for words within a sentence using a machine learning algorithm. For this I used a pre-labeled data set (Brown Corpus) and implemented, evaluated and compared several versions of the n-Gram algorithm with the aim of obtaining the best classification accuracy of the automatic part-of-speech tagging process.

Keywords: part-of-speech tagging, n-Gram language model, text normalization

1 Introduction

Natural language processing is a technique that allows computers to understand human language. A correctly done part-of-speech tagging of a word supplies linguistic signals about how it is used in a sentence, and therefore it is useful for distinguishing the meaning of a word. Often words are lexically ambiguous, meaning they can have several parts of speech and depending on them several meanings. Automatic part-of-speech tagging is a disambiguation problem, its purpose focusing on ambiguous words and their correct tagging in different contexts.[1] In this paper I will present how I decided to implement, evaluate and compare several versions of the n-Gram language model with the aim of obtaining the best possible classification accuracy of the process of automatic part-of-speech tagging.

2 Application Architecture

I started developing this application with the pre-labeled Brown corpus data set and divided it into a training data set and a test data set to later evaluate the performance of the algorithms on new data. Then I preprocessed the training data set and saved the parts of speech with which each word appears and the frequency with which they appear. On the test data set I evaluated the implemented predictors, namely: the non-adaptive predictor, the 1-Gram predictor, the 2-Gram predictor and the 3-Gram predictor.

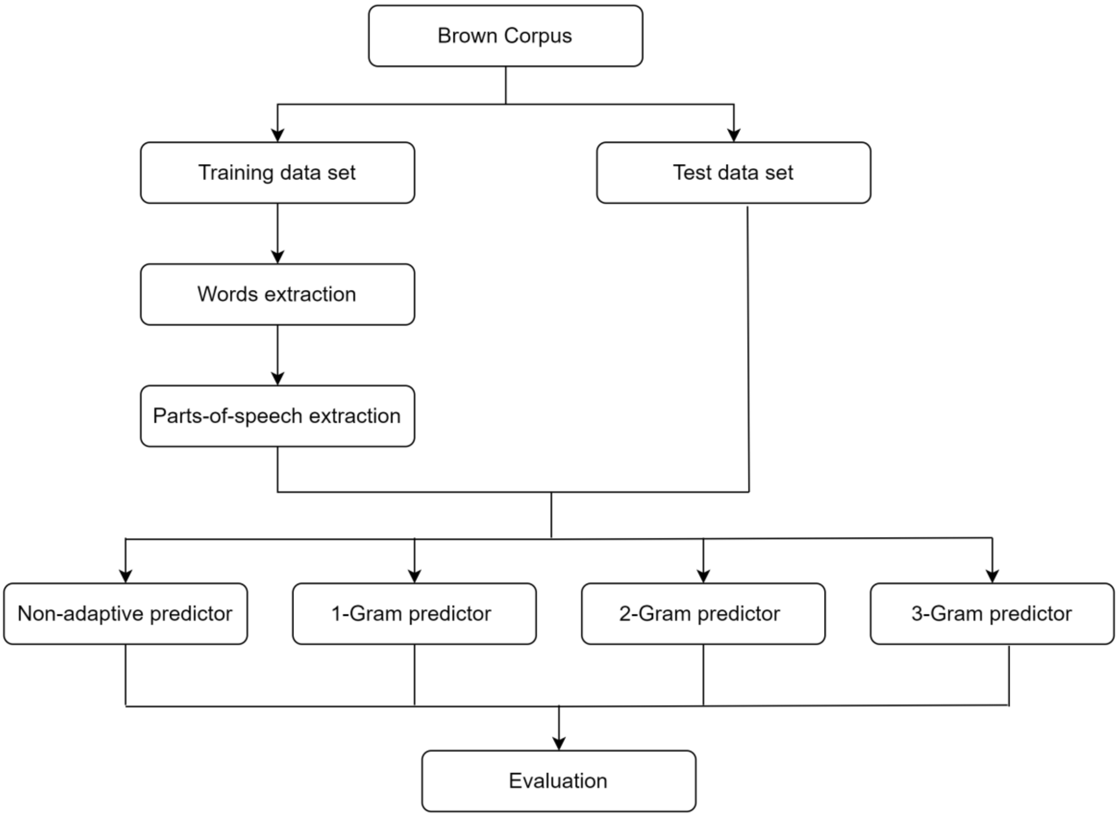


Figure 1. Application Architecture

2.1 Brown Corpus Processing

As input data for the application, I chose to use the C form of the Brown corpus, which is the grammatically tagged version. It consists of 500 files of approximately 2000 words each from 15 different domains. Each word is provided with a label that assigns it to a specific word class.
The words in the Brown corpus are of the "word/part of speech" form. I separated them by "/" and saved in a dictionary all the words, the parts of speech they appeared with and the frequency with which they appeared with those parts of speech.

Example:

Table 1. Sentence from Brown Corpus file

	0
text	The/at Fulton/np-tl County/nn-tl Grand/jj-tl Jury/nn-tl said/vbd

Table 2. Vector of words

	0	1	2	3	4	5
words	The/at	Fulton/np-tl	County/nn-tl	Grand/jj-tl	Jury/nn-tl	said/vb

Table 3. Words dictionary with the parts-of-speech and the frequency with which they were met with those parts of speech

	key	value	
		key	value
wordsList[0]	The	at	1
wordsList[1]	Fulton	np-tl	1
wordsList[2]	County	nn-tl	1
wordsList[3]	Grand	jj-tl	1
wordsList[4]	Jury	nn-tl	1
wordsList[5]	said	vbd	1

After this processing I obtained 64,735 individual words that occur with several parts of speech more than once.

Based on this dictionary, I also created a dictionary for the parts of speech with which to count the frequency with which each appears.

Example:

Table 4. Part-of-speech dictionary

	key	value
PoS[0]	at	1
PoS[1]	np-tl	1
PoS[2]	nn-tl	1
PoS[3]	jj-tl	1
PoS[4]	nn-tl	1
PoS[5]	vbd	1

I obtained 472 parts-of-speech.

To ease learning, I have reduced these parts-of-speech to 11 general parts-of-speech, namely: noun, verb, preposition, pronoun, article, adjective, conjunction, adverb, numeral, interjection and other.

After reducing the parts of speech, I obtained the following statistics:

Table 5. Parts-of-speech statistics on the entire data set after generalization

Curt. No.	PoS	Frequency of occurrence	Percentage of total words
1	noun	274336	27.30%
2	verb	197743	19.68%
3	preposition	122473	12.19%
4	pronoun	107717	10.72%
5	article	99077	9.86%
6	adjective	80741	8.03%
7	conjunction	60306	6.00%
8	adverb	48488	4.82%
9	numeral	7428	0.74%
10	other	6014	0.60%
11	interjection	627	0.06%

As expected, the noun is the most frequent part of speech with a frequency of 27.30%, followed by the verb with a frequency of 19.68%.

2.2 Splitting the data set

It is important to divide the data set into a training data set and a test data set to evaluate algorithms with new data that are part of the same domains. I chose to divide the data set into 70% training data set and 30% test data set. I did the same processing that I did on the whole data set on the training data set, and I obtained the following statistics:

Table 6. Part-of-speech statistics on the training dataset

Curt. No.	PoS	Frequency of occurrence	Percentage of training data set words
1	noun	182077	27.22%
2	verb	131981	19.73%
3	preposition	81632	12.20%
4	pronoun	71867	10.74%
5	article	65732	9.83%
6	adjective	53549	8.01%
7	conjunction	40102	5.99%
8	adverb	32635	4.88%
9	numeral	4922	0.74%
10	other	4000	0.60%
11	interjection	438	0.07%

The proportions of the general parts-of-speech are preserved.

2.3 Non-adaptive predictor

The non-adaptive predictor is the predictor that returns the most frequent part of speech every time. After evaluating the predictor on the test data set, it managed to predict 92,259 words correctly and 243,756 incorrectly, so it had an accuracy of 27.46%.

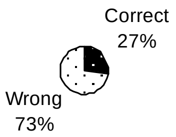


Figure 2. Non-adaptive predictor accuracy

2.4 1-Gram adaptive predictor

The 1-Gram adaptive predictor is the predictor that, if it finds the word, returns the most frequent part of speech with which it was met, and otherwise returns the most frequently met part of speech in general. After evaluating the predictor on the test data set, it managed to correctly predict 140,309 words and 195,706 incorrectly, so it had an accuracy of 41.76%.

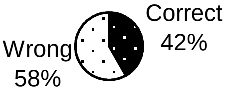
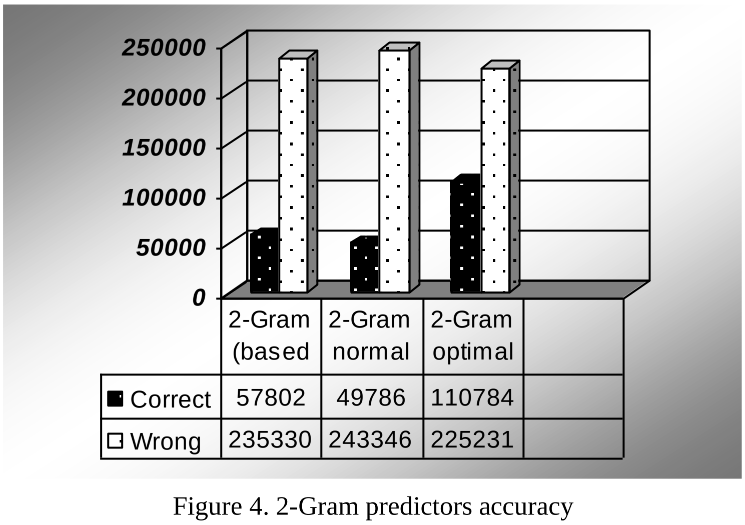


Figure 3. 1-Gram predictor accuracy

2.5 2-Gram adaptive predictor

The 2-Gram adaptive predictor is the predictor that returns the part of speech that occurs most frequently after the part of speech of the previous word. I have implemented 3 variants of this predictor:

- Implementation based on 1-Gram predictions. The part-of-speech prediction of the current word is done with the 1-Gram algorithm's prediction of the previous word. After evaluating the predictor on the test data set, it managed to correctly predict 57,802 words and 235,330 incorrectly, so it had an accuracy of 19.72%.
- Implementation based on his own predictions. The part-of-speech of the first word in a sentence is predicted with the 1-Gram algorithm, and the rest will be predicted based on the output of the 2-Gram predictor for the previous word. After evaluating the predictor on the test data set, it managed to correctly predict 49,786 words and 243,346 incorrectly, so it had an accuracy of 16.98%.
- Largest limit available. This implementation of the adaptive 2-Gram predictor aims to achieve the highest possible performance by starting from the correct part of speech. This assumes that instead of using the 1-Gram predictor, the correct part-of-speech of sentence-beginning words will be read directly from the test data set. This will eliminate the highly likely possibility of mislabeling an entire sentence just because the 1-Gram predictor mispredicted the first word in the sentence. After evaluating the predictor on the test data set, it managed to correctly predict 110,784 words and 225,231 incorrectly, so it had an accuracy of 32.97%



	2-Gram (based	2-Gram normal	2-Gram optimal	
■ Correct	57802	49786	110784	
□ Wrong	235330	243346	225231	

Figure 4. 2-Gram predictors accuracy

2.6 3-Gram adaptive predictor

The 3-Gram adaptive predictor is the predictor that returns a part of speech considering both the part of speech of the previous word and the part-of-speech of the posterior word. Thus, the predictor will evaluate sequences of 3 words and return the most frequent 3-word sequence that has as neighboring parts of speech the parts of speech of the preceding word and the following word. After evaluating the predictor on the test data set, it managed to correctly predict 110,784 words and 199,778 incorrectly, so it had an accuracy of 31.81%

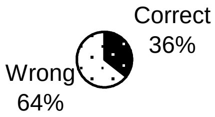


Figure 5. 3-Gram predictor accuracy

3 Experimental Results

After training the presented algorithms on the training data set, I obtained the following evaluation performance on the test data set:

Curt. No.	Algorithm	No. of correct predictions	No. of wrong predictions	Accuracy
1	Non-adaptive	92259	243756	27,46%
2	1-Gram	140309	195706	41,76%
3	2-Gram (based on 1-Gram)	57802	235330	19,72%
5	2-Gram normal	49786	243346	16,98%
6	2-Gram optimal	110784	225231	32,97%
7	3-Gram	93207	199778	31,81%

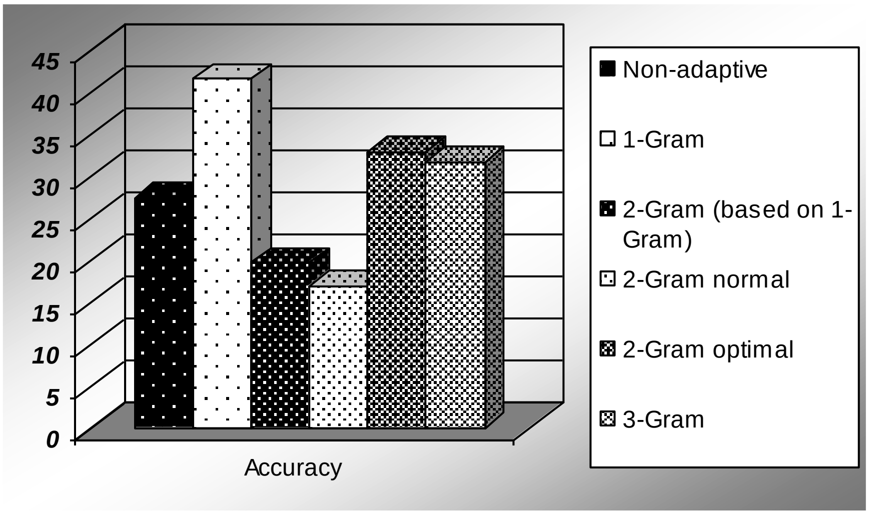


Figure 6. Accuracy of presented algorithms

4 Conclusions

The presented part-of-speech automatic tagging algorithm was based on the different forms of the n-Gram language model. Following the evaluations, I obtained the best performance in terms of classification accuracy using the 1-Gram adaptive algorithm. So, the maximum accuracy I was able to achieve was 41%.

Although I tried to get better performance by increasing the degree of the n-Gram algorithm, by evaluating more neighboring words to widen the context, I could not achieve better performance than the 1-Gram algorithm.

The n-Gram algorithm is not the best performing machine learning algorithm for part-of-speech automatic tagging, but it is a basic tool for understanding the fundamental concepts of language modeling.

References

[1] Dan Jurafsky, James H. Martin, *Speech and Language Processing*, available online at https://web.stanford.edu/~jurafsky/slp3/
[2] https://devopedia.org/n-gram-model , accessed in 05.2022
[3] https://www.ibm.com/cloud/learn/text-mining , accessed in 05.2022
[4] http://korpus.uib.no/icame/manuals/BROWN/INDEX.HTM#bc1 – Brown Corpus Manual, accessed in 05.2022
[5] https://en.wikipedia.org/wiki/Part_of_speech, accessed in 05.2022
[6] https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234 , accessed in 06.2022
[7] https://machinelearningmastery.com/how-to-evaluate-machine-learning-algorithms/ , accessed in 06.2022
[8] https://www.jeremyjordan.me/evaluating-a-machine-learning-model/ , accessed in 06.2022
[9] https://research.ibm.com/publications/a-novel-part-of-speech-tagging-framework-for-nlp-based-business-process-management?mhsrc=ibmsearch_a&mhq=part%20of%20speech%20tagging , accessed in 06.2022

DOI: 10.2478/ijasitels-2024-0008

Electronic voting system using Blockchain technology

Buliga Nicolae-Marian[1]

Abstract

This paper presents an application for a voting system that uses blockchain technology. As is known, blockchain technology has a wide range of applications. What makes it suitable for a voting system is that it can prevent fraud. Another advantage of using blockchain technology for a voting system is that it offers unlimited system availability: the system's operation is not affected by certain factors, such as power outages, updates or network connection.

Keywords: blockchain, electronic voting

1 Introduction

Blockchain is an innovative technology created by the group of people known as Satoshi Nakamoto [1].

Blockchain is basically a digital ledger of transactions, similar to a database, that is distributed to all nodes in the network. Each record in this register is called a block. Every time a new transaction takes place on the blockchain, a record of that transaction is added to the ledger of each node.

The components of a block are [2]:

- a public key;
- a list of transactions;
- the id for the current block;
- the id of previous block.

The main properties of blockchain technology are:

1. Decentralized. Blockchain system is a decentralized system witch means that the information is not stored by a single entity.
2. Transparency. This property assumes that anyone on that network can track the transaction history.
3. Immutability. The immutability property assumes that once a record has been created, it can no longer be changed.

The blockchain technology uses a special type of network named peer-to-peer network [1]. This network is represented by a collection of interconnected nodes. The nodes are individual computers that receive an input, apply an algorithm and generate a result.

In the following image is described how a blockchain system works:

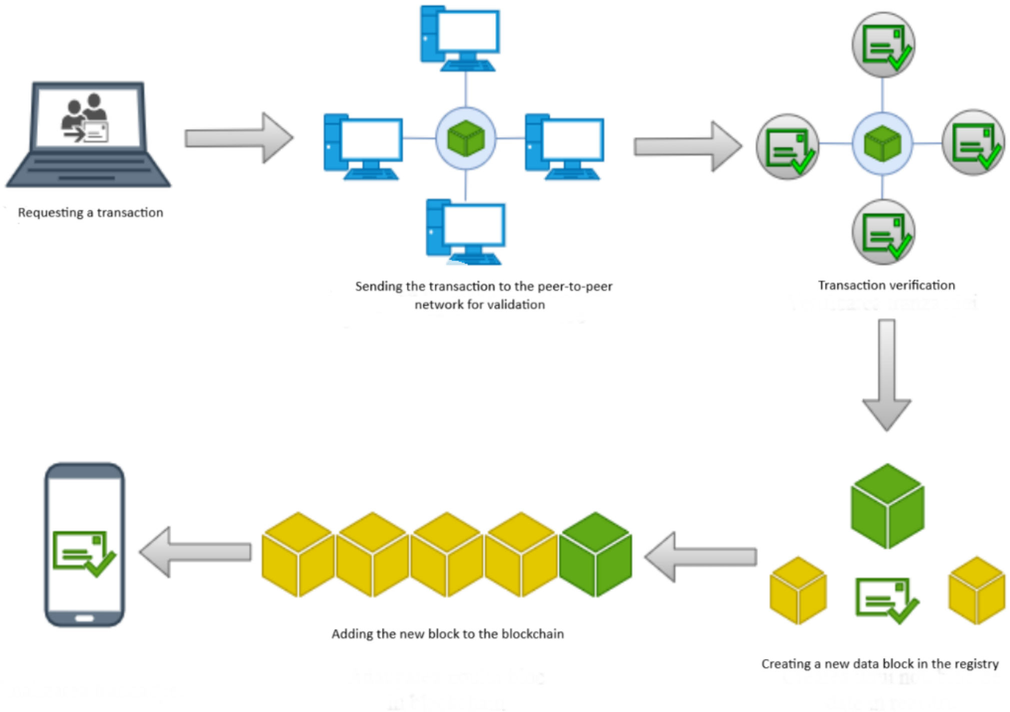

Requesting a transaction
Sending the transaction to the peer-to-peer network for validation
Transaction verification
Creating a new data block in the registry
Adding the new block to the blockchain

Figure 1. How a blockchain system works[1]

2 Related work

2.1 Requirements

In order to create a voting application, certain requirements must be respected. Bellow are presented the main requirements of a voting system and how the proposed application meets them:

a) Keeping the identity of the voter hidden. The proposed application uses the blockchain technology to store the votes. In the voting process, the personal data of the voter is not saved, but only an address.

b) Preventing multiple voting. Every time a vote is added, the block chain is checked (the users who have previously voted are taken and compared to the current user; if the current user is in the list obtained, the vote is prevented)

c) The possibility to generate the voting result.

2.2 Proposed system design

In order to create an e-voting application that uses blockchain technology, it was necessary to use an existing blockchain system. This system is represented by a program written in the C programming language that simulates the blockchain functionality.

In order to develop a voting application, it was necessary to create a new Windows Forms Application project that could communicate with this blockchain system. The graphical interface of the project includes a TabControl menu containing four windows (pages): the window for configuration, the window for voting system, the window for viewing the voting results and the window for displaying the specific information of each block on the blockchain.

2.3 Usage scenarios

The proposed application has four usage scenarios:

1. Configuration – the users can configure the blockchain system on the local machine.
2. Voting system - the users can choose a voting option.
3. Voting result - the users can view the result of the vote.
4. Block explorer – the users can generate the block chain that was formed after the vote.

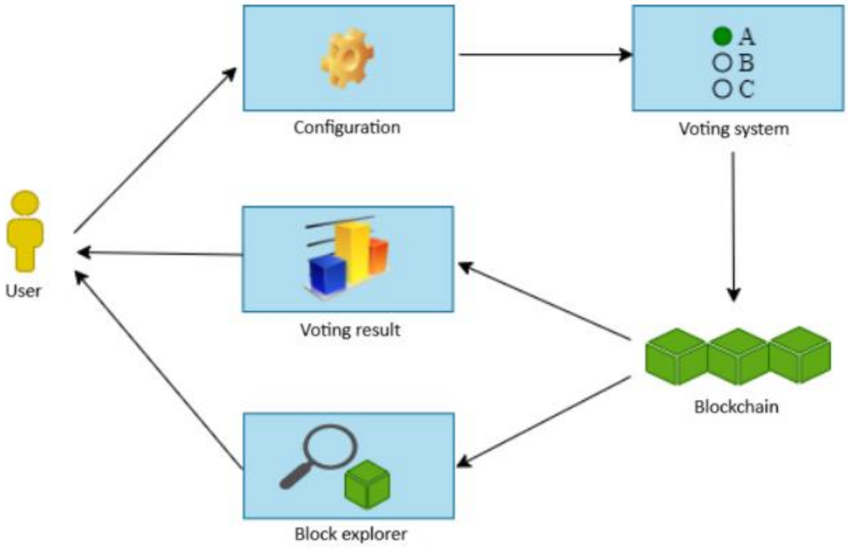
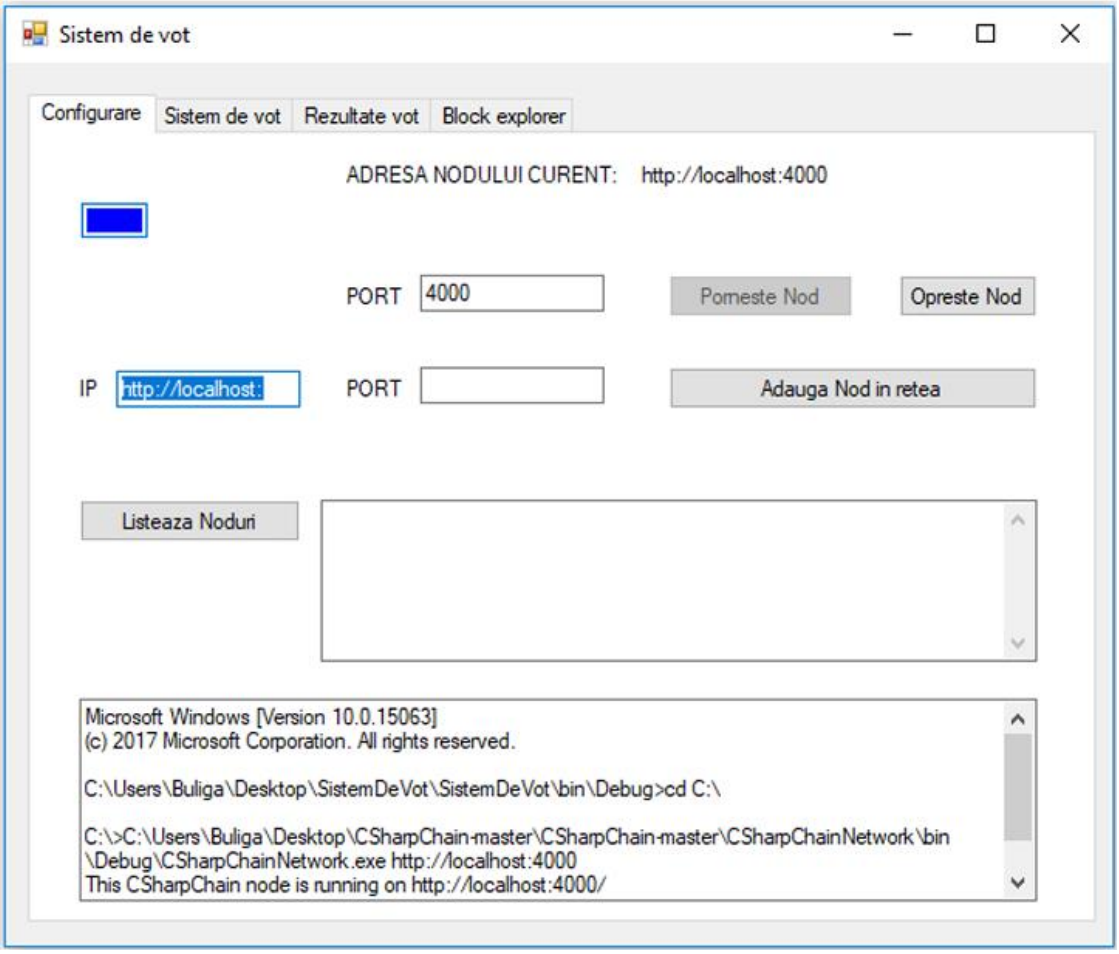


Figure 2. Usage scenarios

2.4 The voting process

Acording to the previous description, the application is developed in the C# programming language and consist of a TabControl component with four windows.

The first window represents the configuration window. As is shown in the Figure 3, this window contains four buttons through which users are allowed to start or stop a node on the local machine, add another node to the network and display the connected nodes. This window also allows to view all the operations that were done on the local machine.



ADRESA NODULUI CURENT: http://localhost:4000

PORT 4000

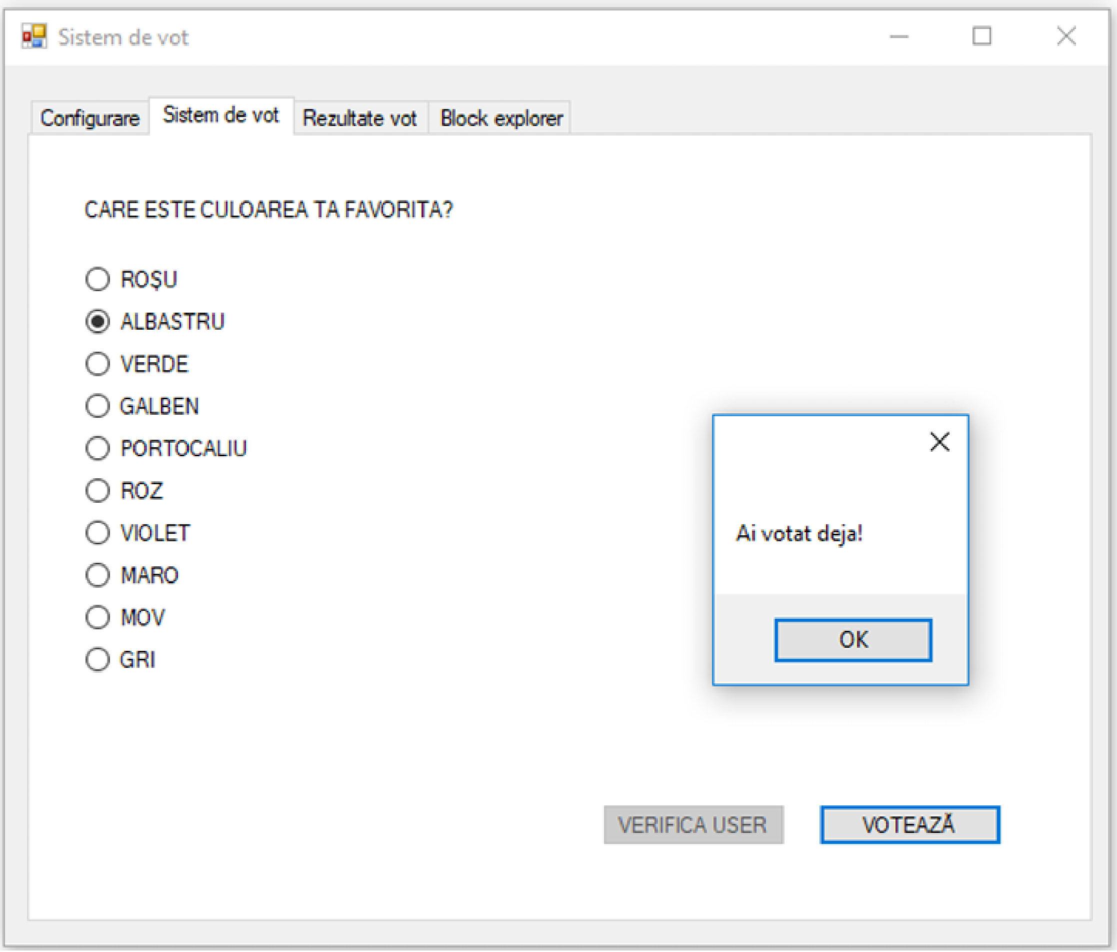
IP http://localhost:

PORT

```
Microsoft Windows [Version 10.0.15063]
(c) 2017 Microsoft Corporation. All rights reserved.

C:\Users\Buliga\Desktop\SistemDeVot\SistemDeVot\bin\Debug>cd C:\

C:\>C:\Users\Buliga\Desktop\CSharpChain-master\CSharpChain-master\CSharpChainNetwork\bin
\Debug\CSharpChainNetwork.exe http://localhost:4000
This CSharpChain node is running on http://localhost:4000/
```

Figure 3. The configuration window

The second window is the Voting system window and allows users to choose a voting option and to register their vote. At this step, the user must first press the Verify user button to unlock the voting options and the vote button. At the press of the Verify user button, all blocks on the blockchain will be covered, all users who have previously voted will be taken and compared to the current user. If the current user is not in the list of users obtained by taking over all the users on the blockchain, the voting options will be activated together with the Vote button.



CARE ESTE CULOAREA TA FAVORITA?

○ ROȘU
● ALBASTRU
○ VERDE
○ GALBEN
○ PORTOCALIU
○ ROZ
○ VIOLET
○ MARO
○ MOV
○ GRI

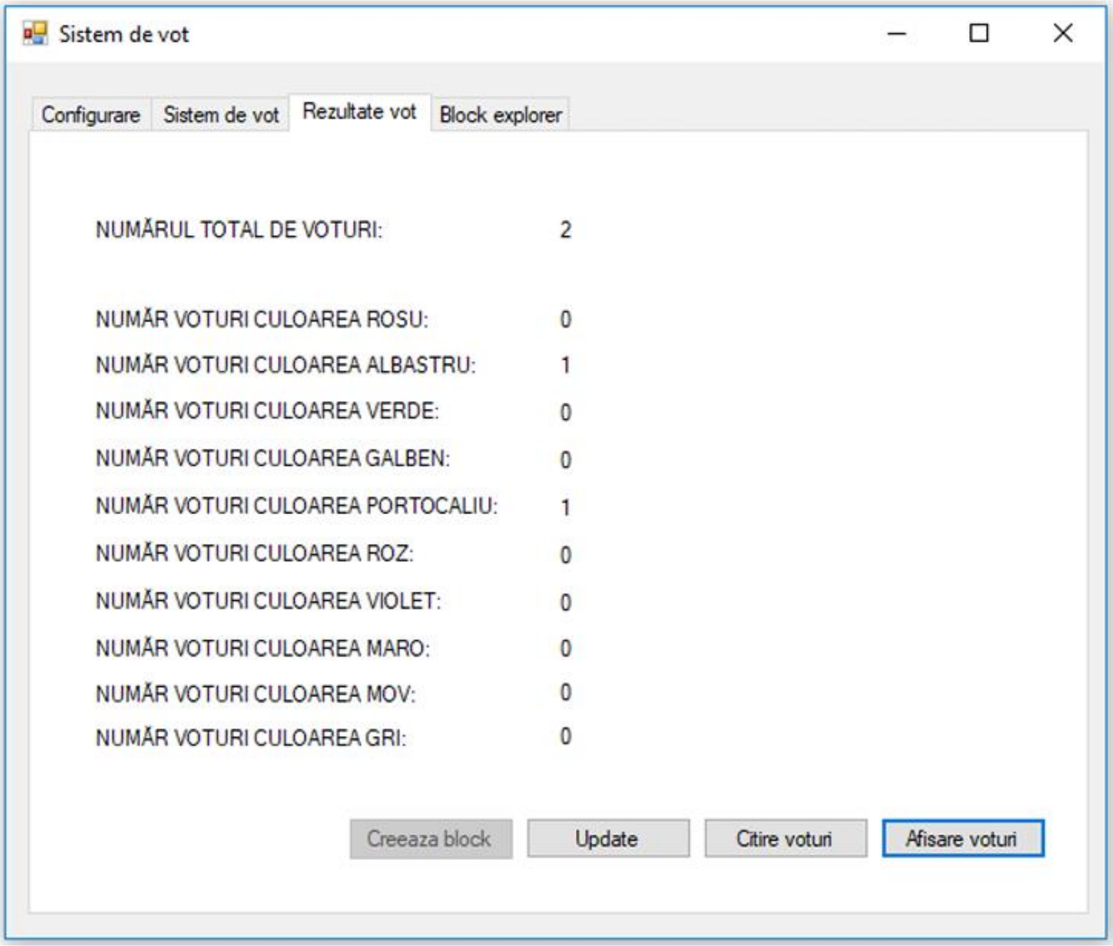
Ai votat deja!

Figure 4. Voting system window

The third window, the Voting results window is designed to display the number of votes for each record.

This window contains for buttons with the following functionality:

a) The Create block button – allows the addition of a new data block at a given time in blockchain, consisting of transaction that are pending;
b) The Update button – it is used to update the blockchain to the longest chain found in the network;
c) The Read vote button – goes to the blockchain and reads the votes;
d) The Display vote button – display the result of the vote.



NUMĂRUL TOTAL DE VOTURI: 2

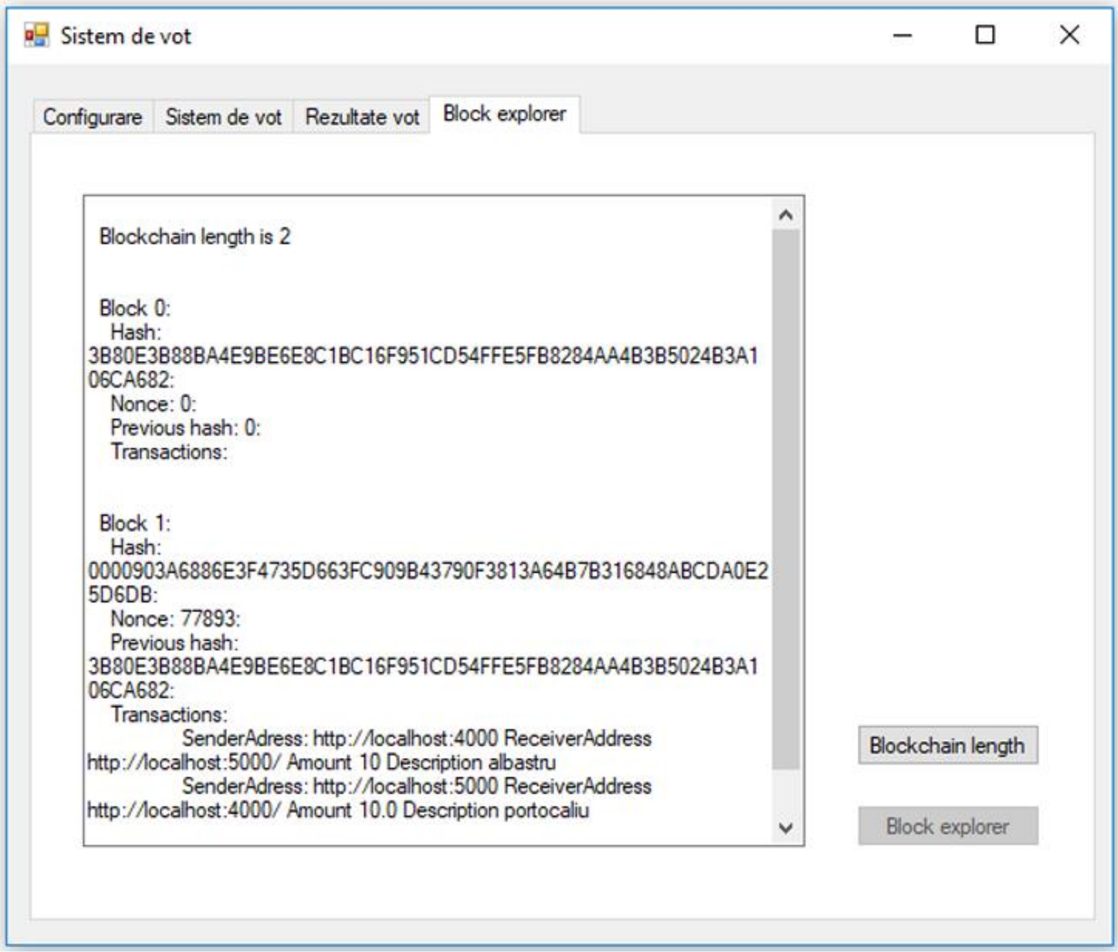
NUMĂR VOTURI CULOAREA ROSU: 0
NUMĂR VOTURI CULOAREA ALBASTRU: 1
NUMĂR VOTURI CULOAREA VERDE: 0
NUMĂR VOTURI CULOAREA GALBEN: 0
NUMĂR VOTURI CULOAREA PORTOCALIU: 1
NUMĂR VOTURI CULOAREA ROZ: 0
NUMĂR VOTURI CULOAREA VIOLET: 0
NUMĂR VOTURI CULOAREA MARO: 0
NUMĂR VOTURI CULOAREA MOV: 0
NUMĂR VOTURI CULOAREA GRI: 0

Figure 5. Voting results window

The forth window, the Block explorer window, is designed to display information about data blocks on the network. This window contains two buttons:

a) The Blockchain length button – used to display the total number of blocks from network;

b) The Block explorer button – used to display information about each block of data in the blockchain.



```
 Blockchain length is 2


 Block 0:
  Hash:
3B80E3B88BA4E9BE6E8C1BC16F951CD54FFE5FB8284AA4B3B5024B3A1
06CA682:
  Nonce: 0:
  Previous hash: 0:
  Transactions:


 Block 1:
  Hash:
0000903A6886E3F4735D663FC909B43790F3813A64B7B316848ABCDA0E2
5D6DB:
  Nonce: 77893:
  Previous hash:
3B80E3B88BA4E9BE6E8C1BC16F951CD54FFE5FB8284AA4B3B5024B3A1
06CA682:
  Transactions:
            SenderAdress: http://localhost:4000 ReceiverAddress
http://localhost:5000/ Amount 10 Description albastru
            SenderAdress: http://localhost:5000 ReceiverAddress
http://localhost:4000/ Amount 10.0 Description portocaliu
```

Figure 6. Block explorer window

3 Conclusions

The purpose of this paper was to present an electronic voting system that uses the blockchain technology.

The main advantages of using blockchain technology in a voting system include: keeping users anonymous, preventing fraud and ensuring unlimited system availability.

4 References

[1] Nick Darlington, Blockchain For Beginners: What Is Blockchain Technology? A Step-by-Step Guide, https://blockgeeks.com/guides/what-is-blockchain-technology/

[2] Alexandru Gavrilovici, Este blockchain-ul urmatorul internet?, https://smarternext.com/ro/este-blockchain-ul-urmatorul-internet/

DOI: 10.2478/ijasitels-2024-0019

The Emerging Technologies: The Drivers for Digital Transformation in Business and Education

Adrian Florea[1]

[1] Hasso-Plattner Knowledge Transfer Institute at Lucian Blaga University of Sibiu, adrian.florea@ulbsibiu.ro

Abstract

Emerging technologies, including Blockchain, Quantum Computing, Artificial Intelligence (AI), and Internet of Things (IoT), whether utilized individually or in combinations such as AIoT and IIoT, possess a disruptive nature that is poised to fundamentally reshape the functioning of specific industries, education, and society at large. These technologies act as catalysts for digital transformation, fostering the emergence of novel business models and opportunities for development. Launched as an initiative of the Cluj IT cluster and supported by the Hasso Plattner Institute in Potsdam, Germany, Sibiu Innovation Days (SID) represented a necessity for Sibiu, for the university, for companies in the area, in the idea of bringing together of all decision makers that acts around the concepts of digitalization, innovation and knowledge transfer to a smarter society and a better life for its members. Emerging technologies can exploit Romania's connectivity and broadband capabilities, where Romania is above the European average, bringing it into line with and keeping pace with developed countries in terms of the digitalization of industry, agriculture and society as a whole. In this regard, we believe that no effort is useless to increase Romania's innovation performance, and through the activities carried out at Sibiu Innovation Days 2024, solutions to some of these challenges were addressed and proposed.

Keywords: Emerging technologies, Innovation, Research, Digital Transformation

1 Introduction

1.1 Emerging technologies – catalysts for digital transformation

The disruptive nature of emerging technologies, such as blockchain, quantum computing, artificial intelligence (AI), and the internet of things (IoT), whether applied singly or in combination (AIoT – Artificial Intelligence of Things and IIoT – Industrial Internet of Things), has the potential to drastically alter how particular industries, education, and society as a whole operate. These technologies serve as drivers of digital transformation, encouraging the creation of innovative corporate strategies and growth prospects. Among a multitude of research initiatives, Blockchain technologies have undergone swift evolution, transitioning from applications like cryptocurrencies and decentralized ledgers to a decentralized, programmable, and secure infrastructure.

The rapid evolution of Blockchain technologies has led to the development of decentralized infrastructures, with diverse applications in compliance verification, data integrity, digital identity management and other areas. At the European Union (EU) level, the construction of a Blockchain services infrastructure is underway, known as the European Blockchain Partnership[1], with each country hosting a node. The applications stemming from Blockchain technologies encompass various domains, including the automation of compliance checks in time-sensitive processes, verification of data integrity, electronic medical records, European digital identity management, authentication of educational diplomas and credentials, dependable data sharing among different EU authorities, as well as applications in finance, insurance, and energy supply. At the national level, since 2023, the application for the digital assessment of the written papers of the Baccalaureate exam, administered by the Ministry of Education, integrates the Blockchain technology developed by the experts of the Special Telecommunications Service (STS), which is an active participant in the European Blockchain Services Infrastructure (EBSI).

Concerning generative artificial intelligence and tools embedded in technologies like ChatGPT, Bard, Bing Chat, etc., the year 2023 witnessed a surge in information and challenges. These developments culminated in the approval, on February 2nd, 2024, of the inaugural EU law[2] designed to regulate Artificial Intelligence (AI). This legislation establishes a classification framework based on the potential risks posed by applications utilizing artificial intelligence and the requirements for integrity in all areas of research in the context of current use of AI. Its primary objectives are to enhance the security of AI systems, ensure their compliance with human rights, and establish clear parameters for the use of AI in military, crime, security and other sensitive contexts. Thus, one challenge will be to develop best practice in ethical decisions using AI considering the current academic and research integrity concerns including transparency, informed consent, data privacy, authentic data collection and accuracy of source use.

Quantum represents the 21st century technology that will revolutionize complex computations, information security and medical diagnostics, and the EU is committed to becoming a global leader in quantum technology[3] by developing an ecosystem of excellence and innovation in this field. In the next few years, quantum technologies will make it possible to do things quickly that simply cannot be done today, from complex computational tasks such as modelling biomolecular and chemical reactions that the most powerful supercomputers cannot handle today to sending sensitive information securely anywhere or diagnosing diseases faster and more accurately just by looking inside cells. Intense research based on quantum mechanics has led to major technical advances in many different fields, including quantum computing, sensors, simulations, cryptography, and telecommunications. On 5 December 2023, the EU Council released a declaration[4] that EU Member States are signing to indicate that they recognize the strategic importance of quantum technologies for the EU’s scientific and industrial competitiveness and are committed to working together to develop a world-class quantum technology ecosystem across Europe, with the ultimate goal of making Europe the ‘quantum valley’ of the world, a leading global

[1] https://digital-strategy.ec.europa.eu/en/policies/blockchain-partnership
[2] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689#document1
[3] https://defence-industry-space.ec.europa.eu/eu-space/research-development-and-innovation/quantum-technologies_en
[4] https://digital-strategy.ec.europa.eu/en/library/european-declaration-quantum-technologies

region for quantum excellence and innovation. Through the RONAQCI (Romanian National Quantum Communication Infrastructure) project[5] and SID2024[6], Romania is aligning itself with these efforts.

In addition to the aforementioned themes, Education, Cybersecurity, and Digital Green Transition and Transformation in both industry and society were focal points at Sibiu Innovation Days 2024 conference. The emphasis, however, lied on approaching these subjects in a sustainable, resilient manner that is centered on human needs. The conference served as a valuable opportunity for all Lucian Blaga University of Sibiu (ULBS) partners to share the latest research and innovation outcomes from their ongoing projects.

1.2 About Sibiu Innovation Days (SID)

Celebrating its fifth edition, the Sibiu Innovation Days 2024 conference has become a significant milestone on Romania's innovation map, playing a crucial role in fostering collaborative networks among experts from industry, public administration, and the academic research community. Building upon the success of its four previous editions since 2020, the Sibiu Innovation Days 2024 conference continues to promote the growth of an innovation ecosystem in the Transylvania region, Romania. This ecosystem is centered around the Cluj IT and Sibiu IT Clusters, with a profound impact on shaping the economic and social vitality of the entire region.

The organizing team comprises the Lucian Blaga University of Sibiu (ULBS) through the Hasso-Plattner Knowledge Transfer Institute [1, 2] and the Cluj IT and Sibiu IT Clusters. The Sibiu Innovation Days 2024 conference was scheduled to take place on October 24th and 25th, 2024. As a novelty this year, within the framework of the European CoDEMO (Co-Creative Decision-Makers for 5.0 Organizations) project[7], ULBS was receiving support from Marquardt Schaltsysteme SCS to facilitate one of the workshops entitled “5.0 Value Co-Creation Mechanisms and Experimentations: Case Studies.”

The format of the event included four sections:

- Conference in the form of panels (S1)
- Hackathon for students (S2)
- Session of innovative projects developed by students (S3)
- Industrial Workshops – hands on experience for participants (WS)

1.2.1 European context of SID 2024

Technology is evolving faster than humans can adapt (AI, Blockchain, etc.) [3]. In the same time, the geopolitical and environmental pressures (e.g. global warming, climate changes) are intensifying. To all these challenges the European ecosystem of computing systems should react quickly and efficiently to improve its place within the

[5] https://www.ronaqci.upb.ro/
[6] https://events.ulbsibiu.ro/innovationdays/
[7] https://www.codemo-project.eu/

competition and propose solutions that conform to “European” ethics. The society faces now ‘next computing paradigm’ which combines aspects of the web, cyber-physical systems, digital twins, the cloud, the internet of things, the metaverse and artificial intelligence (AI) into a coherent continuum of computing, intertwining the real world with the cyberworld. Europe assumes leadership roles in promoting the “next computing paradigm” and developing the technologies that will make it happen. Among the objectives we mention: to make the European Union a strong player in AI and Blockchain, to develop innovative and efficient new hardware solutions, from architecture to technology for running new AI models. Making cybersecurity a major upfront concern in every computing system and making sustainability lifecycle assessment a requirement for all new computing systems are also of large interest. In order to fulfil these, there is a need to foster global thinking, to develop methods and tools relating to complex heterogeneous integrations and promote cross-domains and cross-topic collaborations between academic, industrial partners and legal authorities from multiple levels of governance (European, national and regional [1]).

1.2.2 Regional context of SID 2024

Originating as an initiative of the Cluj IT cluster and supported by the Hasso Plattner Knowledge Transfer Institute from ULBS (KTI HPI-ULBS) Sibiu Innovation Days aims creating a mindset for innovation and boost regional innovation by successful knowledge transfer and development of collaborative networks that replicate the best models of innovation from strong innovator countries to Sibiu, and Transylvania region, Romania. SID 2024 included actors from Quadruple Helix of Open Innovation and brought together at Sibiu partners from academia, businesses, municipalities and decision makers. This year participated 22 Academic Partners (new partners from countries like Spain, Germany, France or Romania), 27 Industrial Partners & Open Innovation Community and 1 representative of Municipalities and Decision Makers, bodies generating policies and rules at local or European level (European Commission).

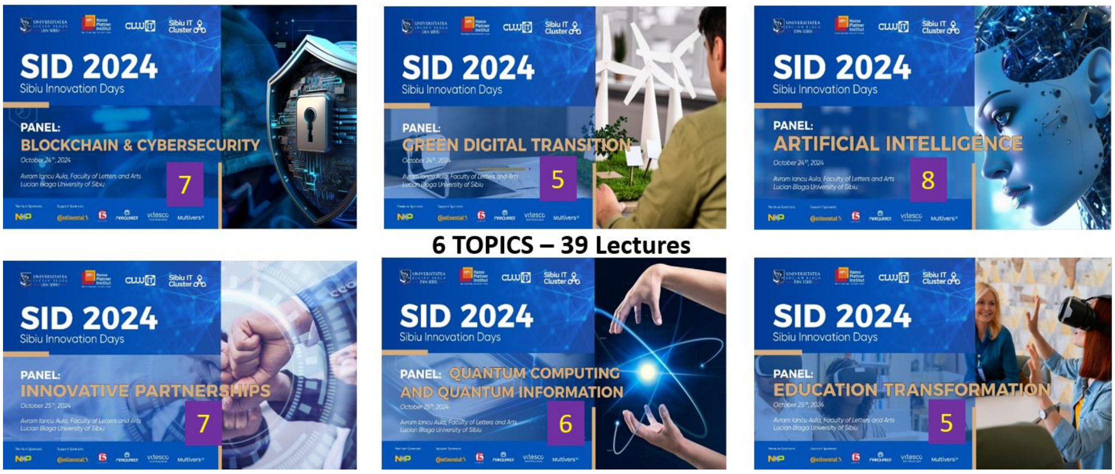


Figure 1. Panels and number of lectures at Sibiu Innovation Days 2024

For the fifth time, Sibiu Innovation Days reconfirmed its position on the regional and even national scene, positioning itself as a promoter of innovation and also a connector between private actors, representatives of public authorities, the local ecosystem and academia, research, bringing into attention current issues at a supranational and implicitly national level. SID conference series are already a reputable reference of the academic environment in Romania and played from its first edition in 2020 a fundamental role in the development of collaborative networks between specialists from industry, administration and research community from universities.

During the two days of the SID2024 conference it was explored, with the guidance and expertise of invited speakers, the process of digital transformation and how this can help our community of students, researchers, business and decision makers to add value to their projects, initiatives and day to day activities. The event was looking to strengthen the collaboration among all these stakeholders, while also improving the way in which we understand and use the innovative services available in Transylvania and the Sibiu region. The conference activities emphasized current research and development concerns of companies active in the area, as well as created new opportunities for collaboration among participants active in a wide array of industries and fields. While highlighting contributions from Hasso-Plattner Knowledge Transfer Institute, active in "Lucian Blaga" University of Sibiu, to the development of the local innovation ecosystem, the conference represented an important component in consolidating Sibiu City's position as an emerging innovation center.

With this 5th edition, SID offered a very rich and exciting program, focused on the emerging technologies seen as the drivers for digital transformation both in business and education. SID2024 topics and presentations aimed adapting industrial and administrative processes in the context of digitalization, as well as towards value creation through structural transformation and generating of new businesses, startups or spinoffs by innovative partnerships. At SID2024 the topics were very timely, focusing on the challenges of the ongoing 5th industrial revolution and the growing role of Artificial Intelligence, Blockchain, Quantum Computing and Communication, and Cybersecurity in systems development. Very interesting lecture presented the AI as a driver of innovative collaboration within and across 5.0 organizations.

The "*Artificial Intelligence*" panel included interesting lectures and debates regarding fairness, equity, ethics, and human-AI collaboration. The conditions of human-AI companionship were defined for ethically tempting jobs, especially in the case of top executive management. The speakers revealed that AI is transforming various fields by enabling personalized learning paths in education and facilitating multilingual journalism with automated news summaries. It enhances creativity in sustainable design hubs through co-creative tools like "Nature's Voice" and drives innovation by optimizing machine learning models for diverse use cases. Efforts to solve AI's black box problem aim to make its decision-making processes more transparent and meaningful, fostering trust. Additionally, artificial intelligence is revolutionizing business practices and workflows, reshaping how we work and innovate across industries.

Even if society in the digital age is dynamically developing, the changes must be sustainable and take into account the limited natural resources. In the "*Green Digitalization Transition*" panel, the speakers revealed limitations and synergies of twin-transition (green and digital), highlighting digital innovations as foundations for driving sustainability and empowering green enterprises and communities for the transition towards cleaner energy futures. As industries move towards a 5.0 transition, smart optimization approaches and digitalization play crucial roles in solving real-world challenges and enabling sustainable practices, ensuring the seamless transition of enterprises from the 4.0 paradigm, based on digitalization and technologies, to the 5.0 paradigm, focused on resilience, a “green” mindset, and a human-centric approach. These advancements create pathways for more resilient and efficient systems, aligning technology with environmental and societal goals.

In the panel entitled “*Blockchain & Cybersecurity*” many speakers from business environment mentioned these emerging technologies as drivers for innovation across diverse fields like automotive, energy sector, or finance. Blockchain is pioneering applications in sectors like energy, insurance, digital identity management, educational credits recognition and decentralized finance, offering secure, transparent solutions. Concurrently, advancements in cybersecurity, including IoT authentication, encryption, and automotive cybersecurity, are vital to protecting modern applications. Strategic planning and design of cybersecurity programs are essential to address emerging digital threats and ensure resilience in evolving technological landscapes.

The lecturers have presented in the “*Innovative Partnership*” panel research and innovation European initiatives, such as those under Horizon Europe that foster partnerships to drive innovation and support 5.0 organizations by enhancing skills and capabilities, and demonstrating how cutting-edge technologies can be used to address societal needs. Collaborations on different topics and research partnerships between academia and industry play a crucial role in advancing technology and using up-to-date tools and equipment in student education. The speakers also reflected on the challenges and successful recipes for building tech startups in Romania. Additionally, funding opportunities from the European level and national and regional ecosystem support are enabling deep-tech startups to scale and thrive, emphasizing the importance of technological transfer and boundaryless ideation for sustainable growth.

For the first time at Sibiu Innovation Days participated lecturers from academia and industry within the “*Quantum Computing and Quantum Information*” panel. Quantum computing and communications based on it represent 21st-century technologies, with advancements both in speeding up computing and with efforts for reducing data storage, with the development of its software stack and toolchains and metrology applications, driving innovation in many domains. Post-quantum cryptography is emerging as a critical area to secure systems against future threats. The density matrix formalism plays a vital role in understanding quantum systems, highlighting the growing importance of quantum mechanics in both theoretical and practical domains. The speakers emphasized the importance of developing education and research in the quantum technology industry, in quantum communications, because this is the first technology closest to the market and, in practice, has the greatest technological maturity.

The last panel of the Sibiu Innovation Days 2024 was dedicated to “*Education transformation: upskilling, reskilling, Lifelong Learning*”. Panelists analyzed digital education, showing how it contributes to transforming learning and skills development, mostly for adults who finished some (or many) years ago their academic study. They emphasized that Massive Open Online Courses (MOOCs) are becoming central to digital universities, and lifelong learning is empowering future-ready individuals. Integrating theories like self-determination in cybersecurity education highlights the importance of motivation in digital training. In the Industry 5.0 era, corporate amnesia underscores the need for continuous reskilling, while training programs play a vital role in supporting industry and organizational transformation. Innovative approaches, such as those at industrial companies from the Sibiu region, showcased how learning can drive adaptation and growth in a rapidly evolving digital world.

In collaboration with ULBS partners – the industrial companies MultiversX Sibiu and Marquardt Schaltsysteme Sibiu, there were organized two industrial demonstrative workshops as part of Sibiu Innovation Days 2024 conference program.

On 24.10.2024 took place the first industrial workshop entitled "*WS1: Introduction to Interacting with the MultiversX Network*" focused on blockchain education of students. The students learnt about challenges in blockchain scalability through parallel execution of transactions and how to detect and predict transactions conflicts (shared objects). In the end of the workshop the conclusions were drawn and open discussions took place and the participants were required to provide feedback. Judging by feedbacks and also that the time allocated for workshop of the 90 minutes was overtook with more than half of hour, important conclusions were that the young students and participants appreciated the technical quality of the presenter and of the workshop in general and are eager for such demonstrative events in future.

On 25.10.2024 took place the second industrial workshop entitled "*WS2: 5.0 Value Co-Creation mechanisms and experimentations, Case studies*" organized within the context of CoDEMO project. The goal of the workshop was to present how could be used a Digital Innovation Environment (established at Sibiu with the help of OMiLAB[8] - ULBS partner in CoDEMO project) to optimize a system test process within Marquardt Company. During product system test phase, it is necessary to determine the passed/failed status of tests via network signals analysis. A second layer of verification is introduced by using a camera taking pictures triggered by those signals. Using Design Thinking concept, Scene2Model and BeeUp tools the students learnt to develop a conceptual model for optimizing the system test process.

2 The Vision of SID2024 Organizers

"*Digitalization is the irreversible path to progress, and the dynamics of technology-driven change are constantly increasing. Blockchain technologies have evolved rapidly, moving from applications such as cryptocurrencies and decentralized ledgers to a decentralized, programmable and secure infrastructure. At the European Union (EU) level, the construction of a blockchain service infrastructure is underway, known as the European Blockchain Partnership, with each country hosting a node. Applications stemming from Blockchain technologies span diverse areas, including the automation of compliance checks in time-sensitive processes, data integrity verification, electronic health records, European digital identity management, diploma and accreditation authentication, secure data exchange between different EU authorities, as well as applications in finance, insurance and energy supply. At the national level, since 2023, the digital evaluation application for written works for the Baccalaureate exam, administered by the Ministry of Education, integrates Blockchain technology developed by experts from the Special Telecommunications*

[8] https://www.omilab.org/omilab_nodes/ulbs/

Service (STS), which is an active participant in the European Blockchain Services Infrastructure (EBSI).

In 2023, there was a surge in knowledge and issues with regard to generative Artificial Intelligence (AI) and the tools integrated into technologies like ChatGPT, Bard, Bing Chat, etc. As a result of these achievements, the EU approved the first law regulating AI on February, 2024. This law creates a classification scheme according to the possible dangers of artificial intelligence applications. Its primary goals are to make AI systems more secure, guarantee that they respect human rights, and establish precise guidelines for the application of AI in security, criminal, military and other sensitive contexts.

In the coming years, quantum technologies will make it possible to rapidly accomplish things that simply cannot be done today, from complex computational tasks such as modelling biomolecular and chemical reactions that the most powerful supercomputers cannot handle today, to sending sensitive information securely anywhere or diagnosing diseases faster and more accurately just by looking inside human cells. Intense research based on quantum mechanics has led to major technical advances in many different fields, including the development of quantum computers, sensors, cryptography, and telecommunications. ULBS as member of RONAQCI project and the Sibiu Innovation Days 2024 conference, makes efforts to align Romania with European context.

The European Union claim that innovation is the only way to maintain a strong, sustainable and competitive economy. However, there are currently significant gaps in the degree of innovation between EU Member States. Statistics from 2024 classify Romania as the least innovative country in the EU, with a performance of 34% of the union average. One solution to mitigate this disadvantage is to develop collaborative networks that reproduce the best innovation models from advanced countries and adapt them to developing countries in Europe. Emerging technologies can exploit Romania's capability in connectivity and broadband internet, where we are above the European average, making Romania align and keep pace with developed countries in terms of the digitalization of industry, agriculture and the entire society. In this sense, we believe that no effort is useless to increase Romania's innovation performance and, through the activities carried out at Sibiu Innovation Days 2024, we try to address and offer solutions to some of these challenges.", mentioned Prof. Dr. Eng. Adrian Florea, director of the Hasso Plattner Knowledge Transfer Institute at the "Lucian Blaga" University of Sibiu and vice-president of Cluj IT Cluster.

"*In the face of the accelerated pace of technological development, the essential thing is not to adopt innovations, but to integrate them in a way that brings real value to the economy and society. At Cluj IT Cluster, we believe that Romania can become an innovation pole in Europe if we focus on developing advanced digital skills and creating collaborative ecosystems between industry, academia and local authorities. Sibiu Innovation Days is the place where these visions can take shape. What I want to emphasize is that, in the process of digital transformation, the key to success lies in how we manage to adapt technologies to our local needs and transform them into long-term competitive advantages. Romania must not only be a consumer of technology, but also an active creator in this digital revolution*," said Prof. Dr. Eng. Stelian Brad, President of Cluj IT Cluster.

"*This year too, Sibiu IT Cluster is actively involved and supports impactful events, such as Sibiu Innovation Days. Digital transformation, sustainable development, community building and the creation of an ecosystem that supports innovation are priorities for us, and Sibiu Innovation Days does an excellent job of highlighting these topics and bringing real value to a diverse audience.*

We believe in the power of collaboration (after all, Sibiu IT Cluster is a collaboration between 35 important local actors) between the private, academic and public sectors, for the benefit of the community. We happily support such initiatives and are actively concerned about what else we can do together for Sibiu, but also at the national level, to strengthen Romania's position on the technological map of Europe," highlighted Mr. Eng. Cătălin Mihacea, President of Sibiu IT Cluster.

3 Conclusions

In the opinion of organizers this fifth edition of the Sibiu Innovation Days event was a real success. More than 400 participants registered for the event, and almost 1000 people were following online stream with 6 technical panels, 3 keynote presentations and 2 workshops involving more than 51 experts from industry and academic environment and delegates from local, national and international authorities and organizations, European Commission experts, and businessmen (from multinationals and established international brands to local startups) from Romania, Germany, France, USA, Italy, Norway, Austria, Poland, Spain. At the same time, they presented points of view, challenges and solutions related to using emerging technologies in business sectors, academia and public services.

At the same time, we are convinced that the collaboration relations between organizers and all participants will not only continue but will flourish in the interest of the ecosystem we represent.

Acknowledgment

Project financed by Lucian Blaga University of Sibiu through the Donation of Hasso Plattner Foundation according with ADDENDUM to the Donation of Funds Agreement dated 18th of July 2018, signed at 22 February 2023.

References

[1] Florea, A., Meinel, C. (2021). Successful Knowledge Transfer – A Boost for Regional Innovation. In: Camarinha-Matos, L.M., Boucher, X., Afsarmanesh, H. (eds) Smart and Sustainable Collaborative Networks 4.0. PRO-VE 2021. IFIP Advances in Information and Communication Technology, vol 629. Springer, Cham. https://doi.org/10.1007/978-3-030-85969-5_50

[2] Nicolaescu, S. S., Palade, H. C., Kifor, C. V., & Florea, A. (2017). Collaborative platform for transferring knowledge from university to industry-a bridge grant case study. In Proceedings of the 4th IETEC Conference, Hanoi, Vietnam (pp. 475-488).

[3] Duranton M., Carpenter P., De Bosschere K., Hoberg T., Robinson C., Vardanega T., Zendra O., (2024). The HIPEAC (High Performance, Edge And Cloud computing) Vision, 2024, https://cloud.hipeac.net/index.php/s/JJGGBK9RyC9m2K4, accessed 28 November 2024.

Interpretation of Student Responses in Teacher Evaluations: A Comparative Cluster Analysis Approach

Crețulescu Radu George[1]*, Pitic Antoniu Gabriel*[1]

[1]*Lucian Blaga University of Sibiu, Romania,*

{radu.kretzulescu, antoniu.pitic}@ulbsibiu.ro

Abstract

This study interprets student responses regarding teacher evaluations using advanced cluster analysis techniques. The responses were clustered using the K-Means and HDBSCAN algorithm from the Data Science GPT [Large language model]. Fifteen main features influencing teacher evaluations were identified, and their relationships were visualized using bar charts and heatmaps to illustrate cluster overlaps. The analysis compares traditional K-Means clustering with Hierarchical Density-Based Spatial Clustering (HDBSCAN), highlighting the benefits of density-based clustering in capturing nuanced insights. These findings provide actionable recommendations for enhancing teaching quality and student satisfaction in higher education.

Keywords: clustering, evaluation student feedback

1 Introduction

Teacher evaluations are crucial in assessing the quality of instruction and guiding improvements in educational practices. Student feedback serves as a fundamental source for understanding what factors contribute to effective teaching and how they influence the learning experience. Traditionally, clustering techniques like K-Means have been used to categorize and analyse such feedback. However, newer algorithms such as Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) offer the ability to identify more nuanced clusters, especially in complex datasets.

The goal of this paper is twofold: first, to interpret the student responses to teacher evaluations by identifying the primary features that contribute to the perception of teaching quality; second, to compare the effectiveness of K-Means clustering against HDBSCAN for categorizing and understanding the underlying patterns in the data. Fifteen key features have been extracted from the dataset, which represent the core aspects of teaching effectiveness as perceived by students.

2 Methods

2.1 Data Collection

Data was collected from student evaluations about the teaching activity across multiple courses in the Lucian Blaga University of Sibiu. The evaluations contained qualitative responses where students highlighted aspects they appreciated regarding the teaching style, course content, and overall classroom environment. Also, the aspects that should be improved regarding the teaching and working materials a practical support for the students in the labs. The dataset included more than 500 responses, each containing rich feedback ranging from course materials to the teaching and communication skills of the tutors and professors. The answers given by the students were collected into one single file and separated by a blank line. The answers are in Romanian.

2.2 Feature Extraction

2.2.1 TF-IDF

From the qualitative feedback, 15 main features were identified as recurring themes. These features included factors like "Teaching Clarity," "Course Organization," "Instructor Engagement," and "Application of Real-World Examples." To analyze the data, a text mining approach was employed using a Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer, which transformed the text into numerical representations. The data were then processed to extract key themes based on frequency and relevance.

The Term Frequency-Inverse Document Frequency (TF-IDF) method is used to evaluate the importance of a term within a document relative to a corpus. It is calculated as follows [1]:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t) \qquad (1)$$

Where:

- $TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$ represents term frequency.
- $IDF(t) = log\left(\frac{N}{n_t+1}\right)$ calculates the inverse document frequency.

Term Frequency – The number of times a term appears in a document, normalized by the total number of terms in the document. In our case it represents the raw count of term in document. Inverse Document Frequency – Measures how much information the term provides across the corpus. The addition of 1 prevents division by zero when a term does not appear in any document. Thus, the TF-IDF score increases proportionally to the number of times a term appears in a document (TF) but is offset by how frequently the term appears across all documents (IDF).The TF-IDF method effectively highlights terms that are important within a specific document while reducing the weight of commonly occurring terms across the corpus, making it ideal for text analysis tasks such as clustering.

2.2.2 Identified Features

The following fifteen features emerged as central aspects of teaching quality, as indicated by student responses:

1. *Claritatea predării* (Teaching Clarity): Clarity and understandability of lectures.
2. *Implicarea instructorului* (Instructor Engagement): Ability of the instructor to actively engage students during the course.
3. *Calitatea feedback-ului* (Feedback Quality): Quality and constructiveness of feedback on student assignments.
4. *Organizarea cursului* (Course Organization): Logical structuring and scheduling of course content.
5. *Aplicații practice în lumea reală* (Real-World Applications): Use of practical examples that connect theory with real-world scenarios.
6. *Disponibilitatea instructorului* (Instructor Availability): The teacher's accessibility for consultations outside class.
7. *Evaluare obiectivă* (Fair Assessment): Perceived fairness of grading and evaluation.
8. *Punctualitate* (Punctuality): The consistency and timeliness of the instructor regarding classes and assignments.
9. *Utilizarea tehnologiei* (Use of Technology): Integration of technological tools and resources in teaching.
10. *Incluziunea studenților* (Inclusivity): Promotion of an inclusive classroom that values diverse student perspectives.
11. *Interacțiunea în clasă* (Classroom Interaction): Encouragement of questions, discussions, and interactive learning.
12. *Materialele de curs* (Course Materials): Quality and availability of the resources provided for learning.
13. *Entuziasmul instructorului* (Instructor Enthusiasm): Enthusiasm demonstrated by the teacher for the subject matter.
14. *Respect față de studenți* (Respect for Students): Respectful treatment of students and encouragement of their participation.
15. *Exerciții practice* (Practical Exercises): Integration of hands-on activities and exercises to reinforce understanding.

3 Clustering Algorithms Overview

3.1 K-Means Clustering

K-Means is a widely used clustering technique that assigns data points into k clusters based on their similarity, measured using the Euclidean distance [2]. The centroids of each cluster are iteratively recalculated until convergence is reached. In this analysis, K-Means identified three major clusters from the data:

- High Satisfaction Cluster: Representing students who provided overall positive feedback, highlighting teaching clarity and instructor enthusiasm.

- Moderate Satisfaction Cluster: Characterized by students with a mix of positive and neutral responses, often pointing out areas for improvement such as instructor engagement or course organization.
- Low Satisfaction Cluster: Students in this cluster generally expressed dissatisfaction, citing issues with punctuality, fair assessment, and course organization.

However, the clusters were relatively broad, lacking granularity in understanding specific subgroups within each level of satisfaction.

3.2 HDBSCAN Clustering

HDBSCAN [2], a density-based clustering algorithm, was used as an alternative to K-Means to capture the complexity of student feedback. Unlike K-Means, HDBSCAN does not require a predefined number of clusters. Instead, it identifies clusters based on the density of data points, making it effective for data with varying cluster shapes and densities.

HDBSCAN identified five distinct clusters, providing a more nuanced understanding of student feedback:

- High Engagement and Clarity: This cluster included students who highly valued teaching clarity and instructor engagement.
- Focus on Real-World Applications: Students in this cluster appreciated the use of real-world examples and practical exercises that made the content relatable.
- Fairness and Inclusivity Emphasis: This cluster included students who particularly appreciated fair grading practices and the promotion of an inclusive learning environment.
- Instructor Availability and Support: A separate cluster was formed by students who valued the instructor's availability outside class and the support provided for their learning.
- Critical of Punctuality and Course Organization: This cluster consisted of students with concerns about punctuality and course structure

4 Visualizations

4.1 Bar Chart of Feature Importance

The bar chart (Figure 1) represents the frequency with which each of the fifteen features was mentioned in student responses. Features such as "Claritatea predării (Teaching Clarity)" and "Implicarea instructorului (Instructor Engagement)" were cited most frequently, indicating their critical importance for student satisfaction.

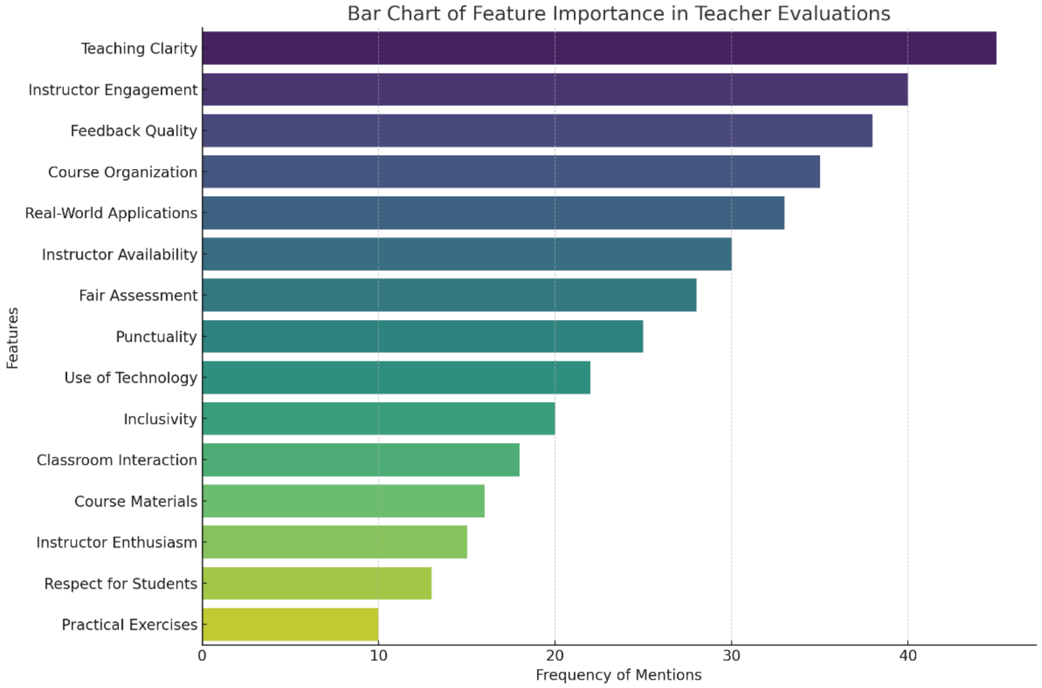


Figure 1: Bar Chart of Feature Importance

4.2 Heatmap of Cluster Overlaps

The heatmap (Figure 2) illustrates the overlaps between different clusters, highlighting the interconnected nature of certain features. Notably, features like "Aplicații practice în lumea reală (Real-World Applications)" and "Implicarea instructorului (Instructor Engagement)" had significant overlaps, suggesting that students who appreciated practical examples were also more engaged during lectures.

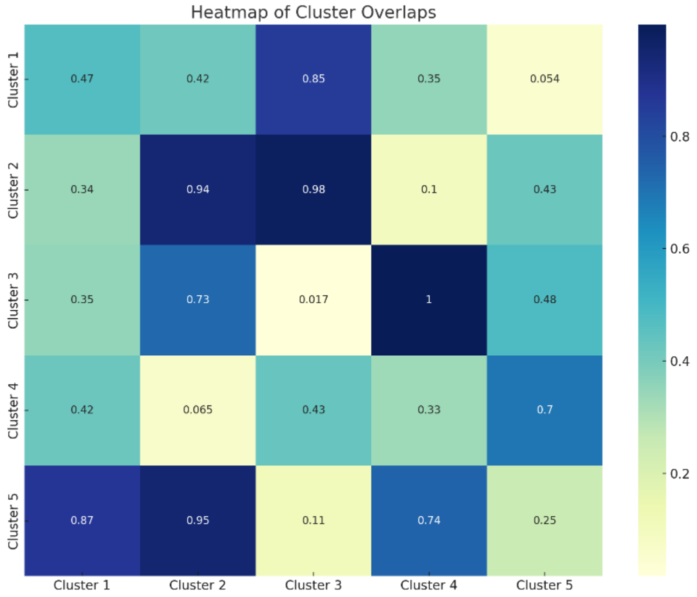


Figure 2: Heatmap of Cluster Overlaps

5 Comparison of Clustering Algorithms

5.1 K-Means vs. HDBSCAN

- **Granularity**: K-Means provided three broad clusters, while HDBSCAN identified five nuanced clusters. HDBSCAN was better able to differentiate between students who had specific preferences, such as valuing inclusivity versus those focused on practical application.
- **Predefined Clusters**: K-Means requires the number of clusters to be specified in advance, which can limit its flexibility. HDBSCAN, on the other hand, determines the number of clusters based on data density, making it more adaptive.
- **Handling Noise**: HDBSCAN has the inherent capability to label outliers as noise, ensuring that clusters are formed only from meaningful data points. K-Means tends to force all data points into clusters, which can result in misleading categorizations.
- **Cluster Shapes**: K-Means works well with spherical clusters, while HDBSCAN can identify clusters of varying shapes and sizes, which is crucial for complex datasets like student feedback.

5.2 Performance Metrics

For evaluating the performance of the two algorithms we have used the Silhouette score and the David Bouldin Index.
The Silhouette Score evaluates the quality of clustering by measuring how similar a data point is to its own cluster compared to other clusters. It is calculated as follows

$$S(i) = \frac{b(i)-a(i)}{\max\left(a(i),b(i)\right)} \qquad (2)$$

Where:

$a(i)$ is the average distance between i and all other points in its cluster.
$b(i)$) is the smallest average distance between i and the points in the nearest cluster.

The Silhouette Score ranges between -1 and 1:

- A score close to 1 indicates well-defined clusters.
- A score close to 0 indicates overlapping clusters.
- A score close to -1 indicates incorrect clustering.

The Davies-Bouldin Index evaluates clustering by measuring the average similarity between each cluster and its most similar cluster. It is calculated as:

$$DB = \frac{1}{N}\sum_{i=1}^{N} \quad \max_{j\neq i}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right) \qquad (3)$$

Where:
- (N): The total number of clusters.
- (σ_i):The average distance between points in cluster (i) and the centroid (c_i).
- ($d(c_i, c_j)$):The distance between the centroids of clusters (i) and (j).

The lower the Davies-Bouldin Index, the better the clustering quality, as it indicates less similarity between clusters.

Table 1. Comparison between metrics

	Silhouette Score	Davies-Bouldin Index
K-Means	0.45	1.85
HDBSCAN	0.60	1.20

HDBSCAN outperformed K-Means in both silhouette score and Davies-Bouldin index, indicating better-defined and more cohesive clusters. This demonstrates HDBSCAN's ability to adapt to the natural density of data and create clusters that are more reflective of actual patterns in student feedback.

6 AI-Based Algorithm Comparison with Classical Clustering [3]

6.1 K-Means Clustering (Classical Approach):

K-Means clustering follows a series of steps that involve initializing centroids, assigning data points to the nearest cluster, recalculating centroids, and iterating this process until convergence. It requires the number of clusters to be predefined and relies on distance metrics like Euclidean distance, making it more rigid in its ability to identify naturally occurring groups, particularly in complex datasets.

6.2 AI-Based Clustering Using HDBSCAN (OpenAI Approach):

HDBSCAN, powered by AI-driven advancements, employs a density-based approach that does not require the number of clusters to be predefined. Instead, it forms clusters based on natural data density and can detect varying cluster shapes and sizes. The AI-based algorithm is more adaptive, better handling the nuances and variability in qualitative data. It also excels in noise handling, identifying outliers and distinguishing them from meaningful data points, something that K-Means does not inherently manage well.

6.3 Advantages of Using AI-Based Clustering

1. **Flexibility and Adaptivity**: Unlike K-Means, which requires a fixed number of clusters, AI-based clustering (HDBSCAN) adapts to the dataset's complexity. This adaptability allows for a more accurate representation of nuanced student feedback.
2. **Handling Complexity**: AI-based clustering can effectively manage non-linear relationships and overlapping clusters, as illustrated in the heatmap. This allows for a more comprehensive understanding of the factors affecting student evaluations.
3. **Noise and Outlier Detection**: HDBSCAN's ability to identify and exclude noise provides cleaner and more insightful clusters. Student feedback often contains diverse perspectives, and removing noise ensures that the insights are focused on genuine patterns rather than anomalies.

4. **Improved Interpretability**: The AI-driven algorithm produced more interpretable clusters, allowing us to separate factors like "Instructor Availability" and "Inclusivity," which were otherwise grouped broadly in K-Means.

6.4 Insights and Recommendations

- **Focus on Teaching Clarity and Engagement**: These features emerged as the most important factors for student satisfaction. Institutions should provide training to enhance instructors' clarity in delivery and engagement strategies.
- **Real-World Applications**: The use of practical examples significantly impacts student satisfaction. Teachers are encouraged to incorporate more real-world scenarios and examples in their lectures.
- **Inclusivity and Fair Assessment**: Students value inclusivity and fairness. Institutions should ensure that instructors receive training on inclusive teaching practices and fair assessment strategies.
- **Instructor Availability**: Availability outside of class hours is appreciated by students. Institutions might consider incentivizing office hours or other forms of support to enhance instructor availability.
- **AI-Based Clustering Advantages**: HDBSCAN, powered by AI-driven advancements, offered more nuanced clustering and better handled the complexity of student feedback compared to the classical K-Means approach. Educational institutions should consider leveraging AI-based clustering techniques to derive more actionable insights from student feedback data.

7 Conclusion

This study demonstrates the value of advanced clustering techniques like HDBSCAN in interpreting student evaluations of teachers. Using the Data Science GPT is suitable for getting some insights into the data used and some tendencies. The most important aspect is getting the results very fast. Then classical algorithms can be applied to confirm the results.

By identifying fifteen critical features and comparing the clustering results from K-Means and HDBSCAN, the analysis highlighted the importance of using algorithms that account for data complexity. HDBSCAN's density-based approach provided more meaningful clusters, offering insights that could directly inform educational improvements.

These findings suggest that educational institutions should prioritize clarity, engagement, and constructive feedback as key drivers of student satisfaction. Future research could involve longitudinal studies to assess the impact of improvements in these areas and explore other machine learning techniques for even deeper insights. This detailed cluster analysis highlights the value of using advanced clustering algorithms, such as HDBSCAN, for interpreting student evaluations. The insights derived from the fifteen identified features provide a roadmap for educational institutions to improve teaching quality and student satisfaction.

The comparison of K-Means and HDBSCAN underlines the importance of using adaptive, AI-driven methods for analyzing complex qualitative data.

Future studies could explore additional machine learning methods to further refine the understanding of student feedback, and longitudinal analyses could be conducted to evaluate the effectiveness of interventions based on these findings.

References

[1] C. D. Manning, P. Raghavan and H. Schütze, (2008). Introduction to Information Retrieval, Cambridge University Press, 2008.

[2] D. Arthur and S. Vassilvitskii, "K-means++: The Advantages of Careful Seeding," in Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 2007.

[3] OpenAI, "Data Science GPT [Large language model]," https://openai.com, 2024.

[4] L. McInnes, J. Healy and S. Astels, "HDBSCAN: Hierarchical Density Based Clustering," Journal of Open Source Software, 2(11), 20, vol. https://doi.org/10.21105/joss.00205, 2017..

International Journal of Advanced Statistics and IT&C
For Economics and Life Sciences

VOLUME 14 **Number 1** **CONTENTS** **2024**

ISSN 2067 – 354X
eISSN 2067 - 354X