Paradigm
reference-global.com

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

Volume 15
Number 1
December 2025

ISSN 2559 - 365X
eISSN 2067 - 354X

DOI: 10.2478/ijasitels-2025-0001

Ground fault location analysis on transmission lines using MATLAB/Simulink environment

Ioan-Dorel Păştina[1], Claudiu Şolea[1], Maria Vinţan[1]
[1]Computer Science and Electrical Engineering Department, "Lucian Blaga" University of Sibiu, Romania, E-Mail: ioandorel.pastina@ulbsibiu.ro; claudiu.solea@ulbsibiu.ro; maria.vintan@ulbsibiu.ro

Abstract
With the continued expansion of the power systems, the accuracy of fault diagnosis on power transmission lines is crucial to ensure their safe and stable operation. A single-phase-to-ground fault can occur anywhere along an overhead transmission line. The amplitude of the fault current and its variation over time depend on the distance between the transformer substation and the fault location. This paper introduces a MATLAB/Simulink programmatic modelling framework for a faster simulation of fault scenarios in a given power system. It also presents the single-phase-to-ground fault current signal features for various fault locations. Both the amplitude of the current and the voltage decreases when the distance from the busbars of the station increases, with the current declining steeper than the voltage.

Keywords: ground fault, overhead transmission lines, fault location, Matlab, Simulink

1. Introduction

The synergy between power systems theory with the tools provided by MATLAB-Simulink environment continues to push the boundaries of power transmission networks analysis. MATLAB and Simulink are both software tools developed by MathWorks [1]. Using the MATLAB/Simulink environment, many models have been implemented for the design of a power transmission network, the assessment of different power quality problems, or for the analysis of faults in power systems. Its extensive simulation capabilities allow extensive modelling and in-depth analysis of different types of faults. For example, the „Single-Pole Reclosing of a Three-Phase Line" [2] model simulates a 735 kV line experiencing a phase-to-ground fault and demonstrates how selective protection isolates only the faulted phase, allows the arc to extinguish, and then automatically recloses it, with shunt reactors stabilizing voltage and preventing restrikes. Similarly, the „Three-Phase Series Compensated Network" [3] represents a 2100 MVA power plant delivering energy through 600 km of 735 kV lines segmented into 300 km sections. In this model, series capacitors reduce line reactance to increase transmission capacity, while shunt reactors maintain voltage stability during normal operation and faults. The „IEEE 13 Node Test Feeder" [4] illustrates an unbalanced distribution network composed of 13 nodes, allowing users to modify phase voltages, amplitudes, and phase angles, as well as inject active or reactive power, to study the impact of unbalanced loads. The „Unified Power Flow Controller

(UPFC) Phasor Model" [5] shows how a UPFC with two 100 MW IGBT converters can balance power flow in a looped transmission system and eliminate transformer overloads, demonstrating flexible control under low- to medium-flow conditions. The „Initializing a 5-Bus Network with the Load Flow Tool of Powergui" [6] example models a 125 kV system supplied by a 9 MW wind farm and a 150 MW plant. The wind farm connects through a 25 kV distribution line with two transformers, a dynamic load, and grounding, while the larger plant feeds the high-voltage grid via a 13.8/125 kV transformer. Using Powergui's Load Flow tool, users can adjust block parameters and observe power flow and node voltages across the five-bus network. Larger networks are represented in the „29-Bus, 7-Power Plant Network" [7], combining seven hydro turbines and a 9 MW wind farm feeding 29 buses, where users can introduce faults and calculate load flow for each bus to analyze system responses. For cable systems, the „Four Coupled 66-kV Cables / Power Cable Parameters" [8] model calculates R, L, and C parameters for four parallel underground cables feeding two 66 kV loads over 30 km, comparing PI-section and frequency-dependent line models to show how cable placement and modeling choices affect induced voltages. The „STATCOM Phasor Model" [9] simulates a device STATCOM that regulates voltage by controlling active and reactive power in a 500 kV transmission line. Installed at the midpoint of a 600 km line, it can vary its output by ±100 MW, significantly improving voltage stability compared to when it is off. A STATCOM unit is also used in the „Flickermeter on a Distribution STATCOM" [10] model, where a flickermeter—a standardized instrument based on the lamp-eye-brain response—measures voltage fluctuations that cause light flicker. The „Performance of Three PSS for Interarea Oscillations" [11] model represents a four-generator, two-area power system where the areas are interconnected by two 230 kV transmission lines of about 220 km each. It is designed to study low-frequency electromechanical oscillations in large networks and to compare the damping provided by three different Power System Stabilizers (PSS). By applying the PSS variants to this benchmark system, the simulation highlights how each controller improves the stability of inter-area oscillations.

A single-phase-to-ground fault can occur anywhere along an overhead transmission line. The amplitude of the fault current and its time variation depend on the fault location, i.e., the distance from the substation to the fault location. The accurate estimation of ground fault currents is essential for the design of grounding installations in electrical transmission and distribution networks, therefore this topic has been the subject of extensive research, with numerous analytical methods published [12], [13], [14], [15], [16], [17], [18], [19]. When a single-phase-to-ground fault occurs on an overhead transmission line in a three-phase power network with grounded neutral, the fault current returns to the grounded neutral through the tower structures, ground return path and ground wires, respectively. In our previous works, analytical methods have been presented to estimate the distribution of ground-fault currents in effectively grounded power networks for a ground fault located along the transmission line. It was shown that the ground fault splits the transmission line into two sections, each spanning from the fault to one end of the line. Depending on the number of towers between the faulted tower and the stations, and the span length, these two line sections can be considered either infinite, in which case the distribution of the ground-fault current is independent of the network end; or as finite, in which case the distribution of the ground-fault current may be highly dependent on the network end [13], [14], [15].

In this paper, the facilities offered by the MATLAB-Simulink environment are exploited. Thus, the Simulink environment serves as a visual platform for developing and testing models. Using the *Simscape Electrical toolbox*, it is possible to simulate various faults, including single-phase-to-ground faults, in an electrical network, by using the *Three-Phase Fault block* [20]. To analyse the influence of the fault location on the values of voltages and currents it is necessary to move the fault block in Simulink after each quadripolar electrical schematic modelling the overhead transmission line. For a complex network structure, this method is time-consuming. In this case, in order to reduce the required manual workload, a MATLAB program will be created to autmatically move the fault block to the targeted distance from the end of the power line. In this way, it is possible to estimate the distribution of the single-phase-to-ground-fault currents, by considering the location of the fault at, practically, any tower of the transmission line. Then, using the MATLAB-Simulink environment (version R2024b), the influence of the fault location on the values of voltages and currents is assessed. The analysis of the influence of the fault location provides a deeper understanding of the behaviour of three-phase power network during faults and facilitates the design of protection systems.

2. The Simulink model of the considered power system

To study the impact of ground fault location on the single-phase-to-ground fault current using the blocks from *Simscape Electrical toolbox*, a Simulink model is proposed. To build a valid test feeder, a power system available in the literature [21] was modelled. The single line diagram of the power system is presented in Figure 1 and includes the following components:

- the power plant G1 equipped with three generators, each with a power of 100 MW;
- the power plant G2 equipped with four generators, each with a power of 100 MW;
- the transformer stations TR1, TR2, TR3 and TR4;
- the double-circuit overhead power lines L1 and L4;
- the single-circuit overhead power lines L2 and L3;
- the consumer C1, considered the main consumer;
- the consumer C2, representing a medium-sized city;
- the consumer C3, representing an industrial zone and a large town.

The power plant G1 ensures the supply of electricity via the 220 kV busbars of the network. The power plant G2 supplies the local consumer C2 through the 10 kV busbar, directly connected to its terminals, and, through the transformer station Tr3, the system on the 220 kV busbar.

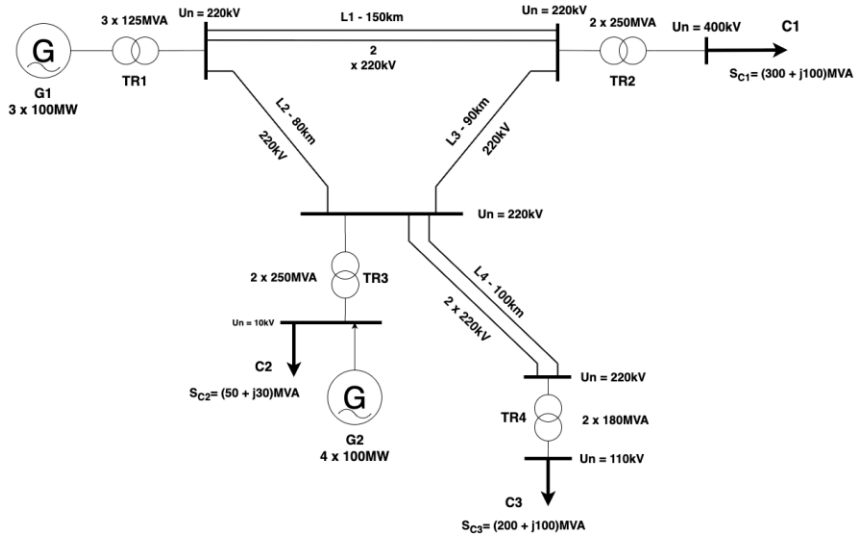


Figure 1 The Power system

MATLAB/Simulink provides specific blocks in its libraries, that enable the modelling of power system elements. The *Simscape Electrical toolbox* provides prebuilt blocks for generators, transmission lines, transformers, loads, faults, and relays. The data presented in Tables 1, 2 and 3 were used to calculate the parameters of the power system components [21], [22], [23]

Tabel 1 Characteristics of generators [21]

Power Plant	Type	Number of generators	Nominal power P_n (MW)	Nominal voltage U_n (kV)	Nominal power factor cosρ	Synchronous reactance X	Transient reactance X	Surge reactance X
G1	TB2-100-2	3	100	15.75	0.85	1.803	0.203	0.138
G2	TB1-100-2	4	100	10.5	0.85	1.795	0.283	0.183

Tabel 2 Characteristics of transformers [21]

Station	No. of transformers	Nominal power S_n (MVA)	Nominal primary voltage U_{np} (kV)	Nominal secondary voltage U_{ns} (kV)	Short-circuit voltage u_{sc} (%)	Short-circuit active power losses p_c (kW)	No-load power losses p_j (kW)	No-load current i_0 (%)	Primary/ secondary connection
TR1	3	125	15	242	11	380	135	3	Δ/Y_0
TR2	2	250	220	420	13	700	250	1	Y_0/Y_0
TR3	2	250	10	242	14	650	240	2.5	Δ/Y_0
TR4	2	180	121	220	12	760	320	3.2	Y_0/Y_0

Tabel 3 Characteristics of transmission lines [21]

Line	No. of circuits	Nominal voltage U_n (kV)	Length km	Resistance			Inductive reactance			Capacitive susceptance		
				Ω/km	of circuit Ω	of line Ω	Ω/km	of circuit Ω	of line Ω	μS/km	of circuit μS	of line μS
L1	2	220	150	0.07	10.5	5.25	0.432	64.8	32.4	2.63	394.5	789
L2	1	220	80	0.07	5.6	5.6	0.421	33.68	33.7	2.71	216.8	216.8
L3	1	220	90	0.07	6.3	6.3	0.421	37.89	37.9	2.71	243.9	243.9
L4	2	220	100	0.07	7	3.5	0.432	43.2	21.6	2.63	263	526

Using specialized blocks available in the MATLAB-Simulink libraries, the model shown in Figure 2 was built.

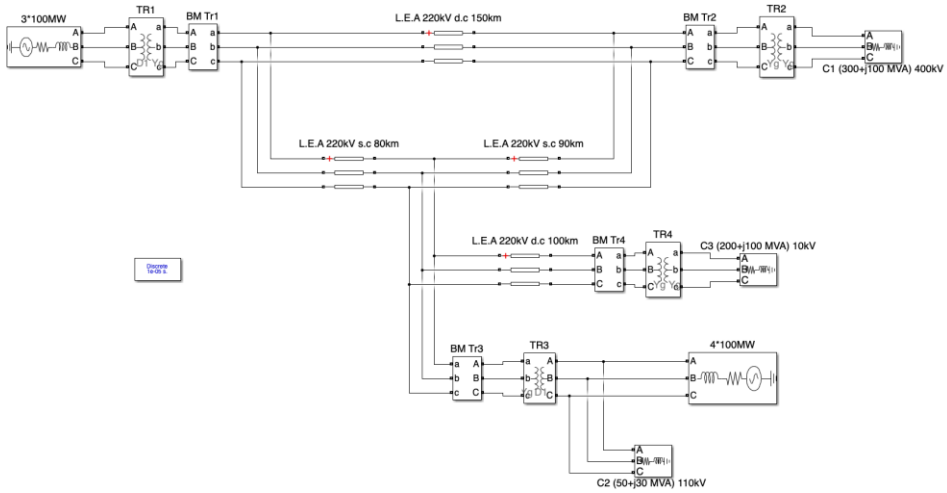


Figure 2 The MATLAB/Simulink model

The validation of the numerical model was done by analysing the normal operating regime. Thus, the values of the voltages and currents in the normal operating mode from the Simulink model were compared against those presented in the literature [21]. The results obtained from the simulations show minimal errors (relative current error: 0.052 %; relative voltage error: 0.183 %).

3. Program implementation in MATLAB/Simulink

The working method implemented in this article is presented in Figure 3.

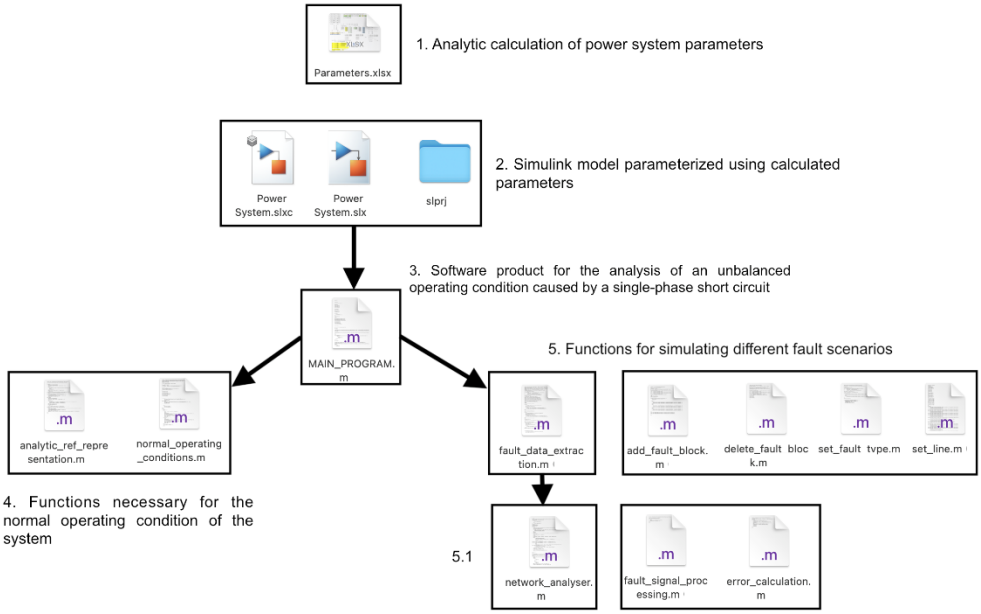


Figure 3 Workflow implemented in the project

The six stages on which the program was built are described in detail. In the file *Parameters.xlsx*, the necessary calculations were performed to determine the parameters of the equivalent schemes of the power system. The Simulink program, shown in Figure 2, is saved as *Power Systems.slx*. The *MAIN_PROGRAM.m* block defines the program code that contains all the functions subsequently called. The .m files labeled with number 4 define the stage for displaying voltage and current values as graphs under normal operating conditions. The function block labeled with number 5 contains the functions for fault simulation, the definition of fault conditions, and the extraction of data related to the power system during a single-phase fault. In stage 5.1, the .m files are defined to perform fault-signal processing, error calculation, and the graphical representation of voltage and current in the power system under fault conditions.

To analyse the influence of the fault location on the values of voltages and currents, a single-phase-to ground fault along the 220 kV double-circuit transmission line L1 (Figure 1) is considered. For different fault locations along the entire length of the L1 power line, instead of a single quadripolar equivalent scheme the transmission line L1 was modelled by a series of identical quadripolar equivalent schemes, each corresponding to a certain length. Specifically, for L1 having 150 km, 10 quadripolar equivalent schemes were considered, each corresponding to a 15 km stretch of line. In this way, the influence of the fault location can be studied by moving the block modelling the single-phase-to-ground fault (Figure 4).

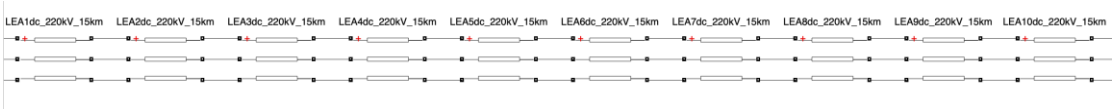


Figure 4 The model of the overhead transmission line L1

The Simulink environment allows the simulation of a single-phase short-circuit in the electrical network using a fault block ("*Three-Phase Fault*") [20]. The classic way of introducing the fault block into the Simulink model is usually done manually by the user, by positioning it in the desired location. To analyse the influence of the fault location on the values of voltages and currents it is necessary to move the fault block in Simulink after each quadripolar equivalent schematic modelling the overhead transmission line. To make this task less time-consuming, a MATLAB program, called *Function „add_fault_block()”*, that automatically moves the fault block after each quadripolar equivalent scheme (Figure 5) was created. The program *Function* "*add_fault_block()*" contains functions that use predefined commands: *"add_block()"* to add the block at the desired location in the modelled system and *"add_line()"* to make connections between the transmission line sections.

```
function [L1,L2,L3,blk]=add_fault_block(nr_block)
% add_fault_block - add fault block ...

x=((nr_block-1)*135)+50;
y=300;
% poz=[x,y,width,length];
poz=[x,y,x+50,y+50];

blk=add_block('powerlib/Elements/Three-Phase Fault', ...
    'SEE/Three-Phase Distributed Line/Fault', ...
    'Position',poz);

switch nr_block
    case 1
        connections_out=["LEA1dc_220kV_15km/RConn 1","LEA1dc_220kV_15km/RConn 2","LEA1dc_220kV_15km/RConn 3"];
    case 2
        connections_out=["LEA2dc_220kV_15km/RConn 1","LEA2dc_220kV_15km/RConn 2","LEA2dc_220kV_15km/RConn 3"];
    case 3
        connections_out=["LEA3dc_220kV_15km/RConn 1","LEA3dc_220kV_15km/RConn 2","LEA3dc_220kV_15km/RConn 3"];
    case 4
        connections_out=["LEA4dc_220kV_15km/RConn 1","LEA4dc_220kV_15km/RConn 2","LEA4dc_220kV_15km/RConn 3"];
    case 5
        connections_out=["LEA5dc_220kV_15km/RConn 1","LEA5dc_220kV_15km/RConn 2","LEA5dc_220kV_15km/RConn 3"];
    case 6
        connections_out=["LEA6dc_220kV_15km/RConn 1","LEA6dc_220kV_15km/RConn 2","LEA6dc_220kV_15km/RConn 3"];
    case 7
        connections_out=["LEA7dc_220kV_15km/RConn 1","LEA7dc_220kV_15km/RConn 2","LEA7dc_220kV_15km/RConn 3"];
    case 8
        connections_out=["LEA8dc_220kV_15km/RConn 1","LEA8dc_220kV_15km/RConn 2","LEA8dc_220kV_15km/RConn 3"];
    case 9
        connections_out=["LEA9dc_220kV_15km/RConn 1","LEA9dc_220kV_15km/RConn 2","LEA9dc_220kV_15km/RConn 3"];
end

L1=add_line('SEE/Three-Phase Distributed Line',connections_out(1),'Fault/LConn 1');
L2=add_line('SEE/Three-Phase Distributed Line',connections_out(2),'Fault/LConn 2');
L3=add_line('SEE/Three-Phase Distributed Line',connections_out(3),'Fault/LConn 3');

end
```

Figure 5 Function „add_fault_block()”

The input parameters of the function include the desired position of the fault block and the total number of faults to be simulated. For the automatic definition of the connections, the program uses a *"switch"* structure which sets the appropriate connection pairs for each section. In addition, the function returns variables containing information about the added block, allowing it to be subsequently modified or deleted depending on the simulated scenario. Figure 6 presents the function that changes the position of the fault block along the transmission line L1.

```
for i=1:block_step:nr_blocks
   [L1,L2,L3,blk]=add_fault_block(i);
   set_fault_type(fault_parameters,F_resistance(1),G_resistance(1));
   out=sim('SEE.slx');
   data_current4=out.Trafo1(:,1:3);
   data_voltage4=out.Trafo1(:,13:15);
   for j=1:3
      date_fault_current_location(:,(i-1)*3+j)=data_current4(:,j);
      date_fault_voltage_location(:,(i-1)*3+j)=data_voltage4(:,j);
   end
   delete_fault_block(L1,L2,L3,blk);
end
```

Figure 6 Extraction of three-phase system data of currents and voltages for analysing the influence of the distance between the fault location and the transformer substation TR1

Thus, using the functions *"add_fault_block()"* and *"set_fault_type()"* the defect block is added and set accordingly. The fault block parameters values are given in Table 1.

Tabel 1 Fault block parameters for analysing the influence of the distance from the fault location to the transformer substation TR1

Faulted phases	Fault duration (s)	Fault resistance– F_resistance (Ω)	Ground resistance – G_resistance (Ω)
Phase 1	[0.04 0.25]	0.1	0.1

In Figure 7 the transition of the fault block between those ten sections of the L1 equivalent quadripolar schemes is presented.

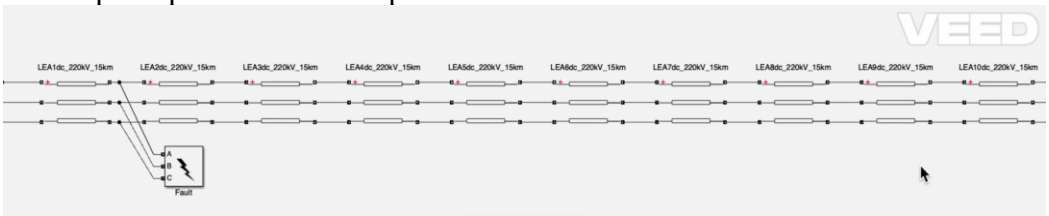


Figure 7 The transition of the fault block between line L1 sections

4. Results

A single-phase-to-ground fault can occur anywhere along a transmission power line. To determine how the fault location influences the amplitude of the single-phase-to-ground fault current, such a fault along the line L1 was simulated. Figure 8 shows the voltages measured on the busbars of the TR1 transformer station in the event of a fault occurring at different distances from it.

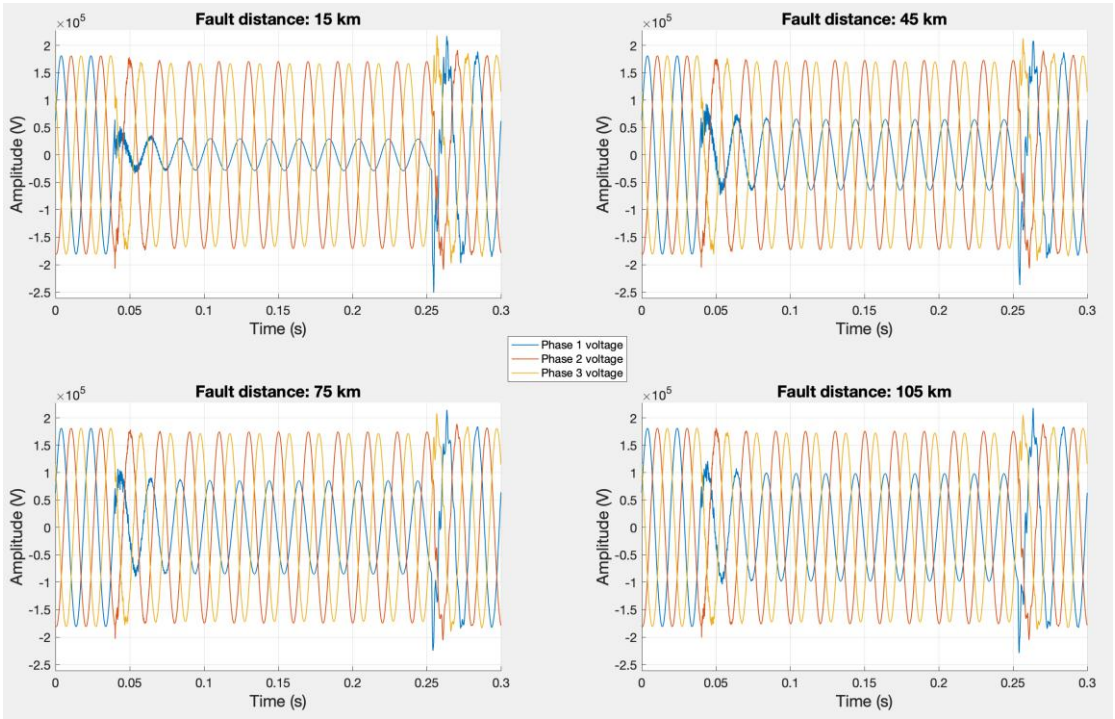


Figure 8 Voltage amplitude as a function of the distance from the fault

Figure 9 presents the currents for the four distances considered.

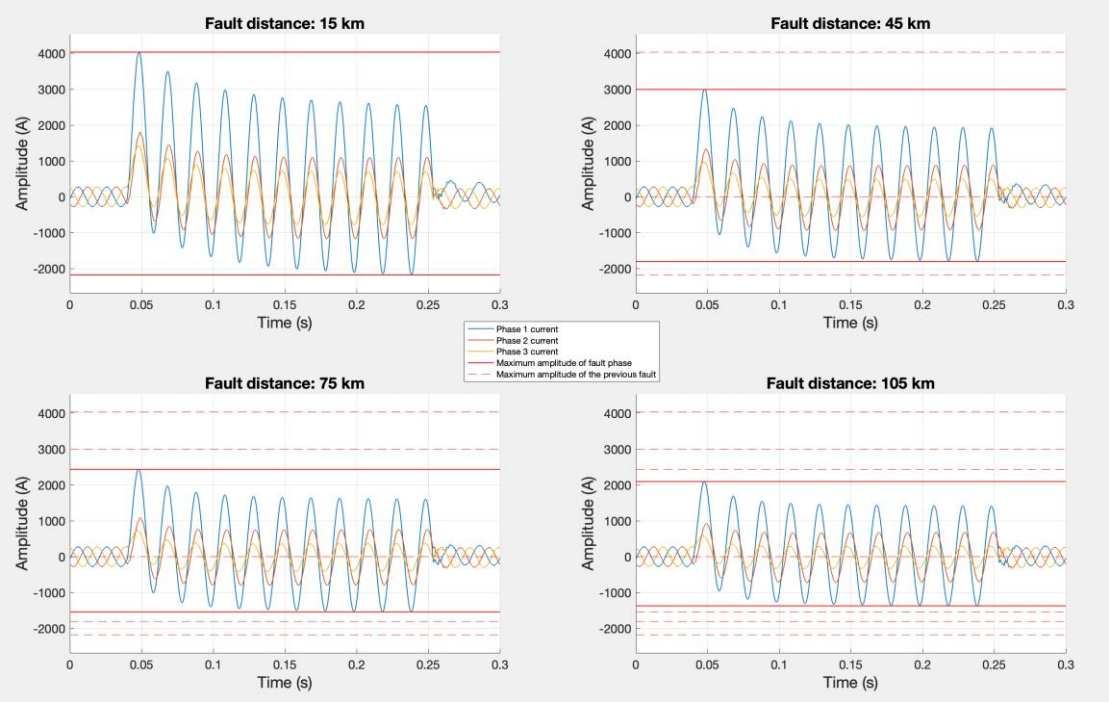


Figure 9 Current amplitude as a function of the distance from the fault

It is observed that the amplitude of the currents decreases with increasing distance from the busbars of the station.

5. Conclusions

MATLAB-Simulink is a widely used platform for power system modelling, simulation, and analysis.

This paper had two main objectives: firstly, the implementation in the MATLAB/Simulink environment of a program for a faster fault simulation, and, secondly, the analysis of the influence of the fault location on the single-phase-to-ground fault current in a power transmission network. The advantage of the proposed MATLAB program is that it makes it possible to estimate the distribution of fault currents considering the influence of various factors on the values of currents and voltages, i.e., the location of the fault at, practically, any transmission line tower.

Realistic fault scenarios were simulated, providing interesting outputs well correlated with other related work [12], [13], [14], [15], [21], [22], [24]. It has been shown that the amplitude of the currents decreases with increasing distance from the station busbars. As the distance between the fault location and the substation increases, the voltage has a more attenuated decrease. This behaviour is explained by the influence of the longitudinal impedance and transverse admittance of the power line. As the values of these transmission lines parameters increase, they influence the ground fault current and voltage amplitude significantly. The transient duration also decreases with increasing distance between the fault location and the transformer substation.

References

[1] "MATLAB." Accessed: May 18, 2025. [Online]. Available: https://www.mathworks.com/products/matlab.html
[2] "Single-Pole Reclosing of a Three-Phase Line - MATLAB & Simulink." Accessed: May 31, 2025. [Online]. Available: https://www.mathworks.com/help/sps/ug/three-phase-line-single-pole-reclosing.html
[3] "Three-Phase Series Compensated Network - MATLAB & Simulink." Accessed: May 31, 2025. [Online]. Available: https://www.mathworks.com/help/sps/ug/three-phase-series-compensated-network.html
[4] "IEEE 13 Node Test Feeder - MATLAB & Simulink." Accessed: May 31, 2025. [Online]. Available: https://www.mathworks.com/help/sps/ug/ieee-13-node-test-feeder.html
[5] "Unified Power Flow Controller (UPFC) Phasor Model - MATLAB & Simulink." Accessed: May 31, 2025. [Online]. Available: https://www.mathworks.com/help/releases/R2025a/sps/ug/upfc-phasor-model.html
[6] "Initializing a 5-Bus Network with the Load Flow Tool of Powergui - MATLAB & Simulink." Accessed: May 31, 2025. [Online]. Available: https://www.mathworks.com/help/sps/ug/initializing-a-5-bus-network-with-the-load-flow-tool-of-powergui.html
[7] "Initializing a 29-Bus, 7-Power Plant Network with the Load Flow Tool of Powergui - MATLAB & Simulink." Accessed: May 31, 2025. [Online]. Available: https://www.mathworks.com/help/sps/ug/initializing-a-29-bus-7-power-plant-network-with-the-load-flow-tool-of-powergui.html
[8] "Four Coupled 66-kV Cables - MATLAB & Simulink." Accessed: May 31, 2025. [Online]. Available: https://www.mathworks.com/help/sps/ug/computation-of-r-l-and-c-cable-parameters.html
[9] "STATCOM Phasor Model - MATLAB & Simulink." Accessed: May 31, 2025. [Online]. Available: https://www.mathworks.com/help/sps/ug/statcom-phasor-model.html
[10] "Flickermeter on a Distribution STATCOM - MATLAB & Simulink." Accessed: May 31, 2025. [Online]. Available: https://www.mathworks.com/help/sps/ug/flickermeter-on-a-distribution-statcom.html
[11] "Performance of Three PSS for Interarea Oscillations - MATLAB & Simulink." Accessed: May 31, 2025. [Online]. Available: https://www.mathworks.com/help/sps/ug/performance-of-three-pss-for-interarea-oscillations.html
[12] F. Dawalibi and G. B. Niles, "Measurements and Computation of Fault Current Distribution on Overhead Transmission Lines," *IEEE Power Eng. Rev.*, vol. PER-4, no. 3, pp. 33–34, Mar. 1984, doi: 10.1109/MPER.1984.5525534.
[13] M. Vintan, "About the Coupling Factor Influence on the Ground Fault Current Distribution on Overhead Transmission Lines," *AECE*, vol. 10, no. 2, pp. 43–47, 2010, doi: 10.4316/aece.2010.02007.
[14] M. Vinţan, "Ground fault location influence on AC power lines impedances," *Proceedings of The Romanian Academy*, vol. Series A: Mathematics, Physics, Technical Sciences, Information Science, no. Volume 15, Number 3, pp. 254–261, 2014.
[15] M. Vintan, "Computing the Ground Fault Current Distribution on Electrical Power Transmission Networks," *Proceedings of The Romanian Academy*, vol. Series A: Mathematics, Physics, Technical Sciences, Information Science, no. Volume 19, Number 3, pp. 463–471, 2018.
[16] L. M. Popović, *Practical Methods for Analysis and Design of HV Installation Grounding Systems*. Elsevier, 2018. Accessed: Sep. 08, 2025. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/C20170012211
[17] M. Kezunovic, S. Lotfifard, and J. Ren, *Design, Modeling and Evaluation of Protective Relays for Power Systems*, 1st ed. 2016. Cham: Springer International Publishing : Imprint: Springer, 2016.
[18] M. A. Bucarelli, D. Agostini, A. Geri, F. M. Gatta, M. Maccioni, and M. Paulucci, "Investigating the impact of short-circuit faults in different neutral configurations: a real case study," in *2023 AEIT International Annual Conference (AEIT)*, Rome, Italy: IEEE, Oct. 2023, pp. 1–5. doi: 10.23919/AEIT60520.2023.10330344.
[19] B. Azimian, R. S. Biswas, and A. Pal, "Application of AI and Machine Learning Algorithms in Power System State Estimation," in *Smart Cyber-Physical Power Systems*, 1st ed., A. Parizad, H.

R. Baghaee, and S. Rahman, Eds., Wiley, 2025, pp. 93–130. Accessed: Mar. 23, 2025. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/9781394334599.ch5

[20] "Three-Phase Fault - Implement programmable phase-to-phase and phase-to-ground fault breaker system - Simulink." Accessed: May 17, 2025. [Online]. Available: https://www.mathworks.com/help/sps/powersys/ref/threephasefault.html

[22] M. Eremia, Y. H. Song, and N. Hatziargyriou, *Electric power systems. Vol. 1*. Bucureşti: Editura Academiei Române, 2006.

[23] C. Rucăreanu, P. Militaru, V. Hotoboc, V. Chiricescu, and V. Prioteasa, *Linii electrice aeriene și subterane. Îndreptar (Overhead and cable transmission lines, in Romanian)*. București: Editura Tehnică, 1989.

[24] C. Solea, D. Toader, M. Vintan, M. Greconici, D. Vesa, and I. Tatai, "Framework for distribution network modelling and fault simulation using MATLAB," in *2022 International Conference and Exposition on Electrical And Power Engineering (EPE)*, Iasi, Romania: IEEE, Oct. 2022, pp. 118–123. doi: 10.1109/EPE56121.2022.9959758.

[25] O. Crișan, *Sisteme electroenergetice (Power systems, in Romanian)*. București: Editura Didactică și Pedagogică, 1978.

[26] F. M. Shakiba, S. M. Azizi, M. Zhou, and A. Abusorrah, "Application of machine learning methods in fault detection and classification of power transmission lines: a survey," *Artif Intell Rev*, vol. 56, no. 7, pp. 5799–5836, Jul. 2023, doi: 10.1007/s10462-022-10296-0.

[27] D. Toader, M. Vintan, C. Solea, D. Vesa, and M. Greconici, "Analysis of the Possibilities of Selective Detection of a Single Line-to-Ground Fault in a Medium Voltage Network with Isolated Neutral," *Energies*, vol. 14, no. 21, p. 7019, Oct. 2021, doi: 10.3390/en14217019.

[28] "Programmatic Modeling Basics - MATLAB & Simulink." Accessed: Sep. 08, 2025. [Online]. Available: https://www.mathworks.com/help/simulink/ug/approach-modeling-programmatically.html

DOI: 10.2478/ijasitels-2025-0002

Subsummation automata in 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

One of the most important constraints faced by the development of artificial intelligence systems is the limited computing power. The logic problems involved in the construction of artificial intelligence systems are most often NP-complete in terms of computational complexity. In the case of conceptual graphs, logical inference is based on the subsummation relation between conceptual graphs. In this paper, we use homomorphisms of conceptual graphs as a model for the subsummation relation. Therefore, the problems related to the existence of relations between two conceptual graphs such as homomorphisms, isomorphisms or generalization, specialization relations become fundamental problems. Although all these decision problems are, in general, NP-complex, they can be efficiently solved in many cases, by certain strategies, without making too much compromise in terms of generality. In this paper, we propose a method to find and manipulate homomorphisms of conceptual graphs based on the concept of finite automaton. Thus, using the finite automaton to recognize predicate homomorphisms and moving a large part of the complexity of finding a homomorphism to the knowledge base construction phase represents our main contribution in this work.

Keywords: conceptual graph, knowledge base, conceptual graph homomorphism, subsummation relation, subsummation automaton

1 Introduction

A knowledge-based system requires both a formalism for specifying knowledge and reasoning mechanisms on this knowledge in order to solve different types of problems. The language of classical first-order logic (FOL) remains the main formalism used for specifying knowledge. Classical logic, in its current form, has sophisticated reasoning mechanisms that can handle complex problems. However, two important problems arise: on the one hand, not all users master mathematical logic, and on the other hand, the computational complexity of some reasoning exceeds the computing capacity available today. Therefore, a large part of the research effort in this area focuses on finding a compromise between the expressiveness of the knowledge representation formalism and the efficiency of the reasoning mechanisms.

A modeling process begins with the conceptualization of the domain, that is, with the detection of all the concrete elements in the modeling domain and their replacement with concepts [6] [7]. Thus, the conceptualization provides a common vocabulary,

which integrates heterogeneous elements in the modeling domain. A complete specification of the conceptualization of the domain is called an ontology [3].

The conceptualization of the domain therefore provides the basic component of the ontology, the vocabulary formed by a set of terms, which name entities and relationships between these entities, and which represent the concepts necessary for modeling the domain.

Depending on the objectives of the model, the semantics of the concepts and relationships included in the vocabulary can be specified in natural language in the case of informal ontologies, or in a formal, axiomatic or prototype-based language in the case of formal ontologies.

Formal ontologies form the basic support for reasoning in the modeled domain. Therefore, the information contained in the ontology will have to support logical reasoning in the modeled domain. For this purpose, vocabulary elements are structured by equivalence relations, in order to capture terms that define the same concepts, or by partial order relations to specify hierarchies of generalization and specialization.

The effort made by researchers to find a compromise between the expressiveness of the knowledge representation formalism and the efficiency of reasoning mechanisms has led to several specifications, among which two types of representations stand out, based on semantic networks, namely: description logics and graph-based representations [4].

Although the language offered by description logics allows the specification of only decidable formulas from ordinal one logic based on unary and binary predicates, practice has demonstrated that they cover a large part of the modeling requirements [5].

When knowledge in a knowledge base is represented as a graph, it is confused with this graph which is called Knowledge Graph. The semantics of a Knowledge Graph is defined by the domain ontology. Over time, a lot of Knowledge Graphs have been built, including: Google Knowledge Graph [8], DBpedia [9], Yago [10], WordNet [11].

In 1976 Sowa [1] defines a version of Knowledge Graph, based on a bipartite multigraph which he calls a conceptual graph (CG) and which represents an important fragment of first-order logic. Inference in conceptual graphs is based on a subsummation relation specified by graph homomorphisms.

One of the most important constraints faced by the development of artificial intelligence systems is limited computing power. Even the simplest artificial intelligence problems, appeal to logic problems that from the point of view of computational complexity are most often NP-complete. In the case of conceptual graphs, logical inference is based on the subsummation relation between conceptual graphs. In this paper, we use homomorphisms of conceptual graphs as a model for the subsummation relation. Therefore, problems related to the existence of relations between two conceptual graphs, such as homomorphisms, isomorphisms or generalization-specialization relations between two conceptual graphs, become fundamental problems. Although all these decision problems are, in general, NP-complex, they can be efficiently solved in many cases, by certain strategies, without making too much compromise in terms of generality.

In this paper, we propose a method to find and manipulate homomorphisms of conceptual graphs based on the concept of finite automaton. Thus, the use of the finite automaton for recognizing homomorphisms and moving a large part of the complexity

of finding a homomorphism to the knowledge base construction phase represents our main contribution in this paper.

Section 2 of the paper contains Background Notions, section 3 introduces the finite automaton as a mechanism for recognizing isomorphisms of conceptual graphs, section 4 presents an algorithm for constructing a finite automaton that recognizes homomorphisms of conceptual graphs. In the last section, some conclusions and observations are presented.

2 Background Notions

Conceptual graphs were initially defined as graphical formalisms for modeling conceptual schemas specific to database systems, but soon found their place in many fields, especially related to artificial intelligence. Starting with the 80s, three directions in which they developed crystallized, namely: as a mechanism for representing knowledge and reasoning on this knowledge, as an interface between natural language and knowledge bases, and as a graphical language for representing an important fragment of first-order logic.

Most Knowledge Graph models are based on directed graphs with nodes of a single type, which represent conceptual entities, and edges between two nodes of the graph that express a semantic relationship between the two entities [13].

In the case of Conceptual Graphs, we have two types of nodes, namely: nodes that represent conceptual entities or concepts and nodes that represent conceptual relations, in which several conceptual entities participate, specified by edges of the graph. Conceptual relations are, most of the time, facts or actions expressed by verbs. The nodes that represent these actions or facts, together with the participants, which are the neighbouring entities, form elementary conceptual graphs in the shape of a star.

The nodes of the conceptual graph are marked with corresponding labels from the vocabulary of the modeled domain, depending on their meaning. Thus, the nodes that represent conceptual entities receive a type label and, possibly, a label that represents a representative of the respective type, and the nodes that represent actions receive only a type label. Also, because conceptual relationships will be represented by atomic formulas in first-order logic, and neighbouring entities represent the parameters of these formulas, the edges connecting neighbours are numbered in the order of their appearance in the parameter list.

Therefore, if the domain vocabulary is $V=(V_C, V_R, E)$ where: V_C contains the types of concepts, V_R, contains the types of relations, and E contains the individual labels, we can define a basic conceptual graph G over the vocabulary V as a tuple (V,G, η), where V is the vocabulary, $G=(C,R, \Gamma)$, is a bipartite multigraph, and $\eta=(\eta_C,\eta_R,\eta_\Gamma)$ is a marking application of the nodes and arcs of the graph (G) defined as follows: $\eta_C:C\rightarrow T_C\times (E\cup\{*\})$, $\eta_R:R\rightarrow T_R$; $\eta_\Gamma:\Gamma\rightarrow N$. We will sometimes denote the elements of Γ by triplets of the form (r,i,c) where r is a relational node, c is a conceptual node neighbouring the relational node r and i is the edge marking that represents the position of c in the parameter list of the relation c.

Both on the set of conceptual entities and on the set of conceptual relations an equivalence relation can be introduced called coreference relation. This equivalence relation determines classes of entities that specify the same entity or the same relation.

Basic conceptual graphs admit coreference only at the level of individual entities. Simple conceptual graphs are an extension of basic conceptual graphs that do not restrict coreference. Both basic conceptual graphs and simple conceptual graphs can be brought to a special form called normal form by cumulating all coreferent concepts into a single representative concept [2].

Practically, a conceptual graph in normal form no longer has different conceptual nodes that represent the same concept. It follows that, by normalizing conceptual graphs, the difference between normal basic conceptual graphs and normal simple conceptual graphs is eliminated. In the rest of this paper, we will use the name conceptual graph for both basic conceptual graphs and simple conceptual graphs.

To simplify the presentation, in the following, we will use the notation $\eta(x)$ for any $x \in C \cup R \cup \Gamma$, with the following meaning:

$$\eta(x)=\begin{cases}\eta_C(x)\ if\ x \in C\\ \eta_R(x)\ if\ x \in R\\ \eta_\Gamma(x)\ if\ x \in \Gamma\end{cases}$$

We will also use the simplified notation $G=(C,R,V,\eta)$, for a conceptual graph where, obviously, C is the set of conceptual nodes, R is the set of relational nodes, V is the vocabulary and η is the mapping of nodes and edges.

If $G= (C,R,V,\eta)$ is a conceptual graph then a conceptual subgraph G' of G is a conceptual graph $G'=(C',R',V',\eta')$ such that $C' \subseteq C$, $R' \subseteq R$, $V \subseteq V$ and η' is the corresponding restriction of η to the components of G'.

We observe that any conceptual subgraph of G can be obtained from G by repeatedly deleting isolated conceptual relations or nodes, no conceptual nodes neighbouring any relation can be deleted without also deleting the corresponding relation.

The subsummation relation in the case of conceptual graphs can be defined by homomorphisms.

If $G=(C_G,R_G,V,\eta_G)$ and $H=(C_H,R_H,V,\eta_H)$ are two conceptual graphs with a common vocabulary V then a homomorphism from G to H is a polymorphic mapping $\varphi: G \rightarrow H$ that maps conceptual nodes in G to conceptual nodes in H, relational nodes in G to relational nodes in H with the following conditions: if $(r,i,c) \in G$ then $(\varphi(r), i, \varphi(c)) \in H$ and for all nodes $x \in C_G \cup R_G$, we have $\eta_H(\varphi(x)) \leq \eta_G(x)$.

If there is a homomorphism from a conceptual graph G to a conceptual graph H then we say that G subsumes H and we denote this fact by $G \geqslant H$. The subsummation relation $G \geqslant H$ represents the fact that the semantics specified by H logically implies the semantics specified by G.

But determining the homomorphisms of conceptual graphs is not a simple problem [2]. In this paper we will show that using the notion of finite automaton for computing the homomorphisms of conceptual graphs comes with important advantages.

A finite, initial and deterministic automaton is a tuple $M = (I, S, O, f, g, s_0)$ where I, S, O are nonempty and finite sets called the input alphabet, the set of states and the output alphabet and $f: S \times I \rightarrow S$, $g: S \times I \rightarrow O$ are the transition and output functions of the automaton [12].

3 Finite automaton and conceptual graph isomorphisms

Conceptual graphs are generally defined as bipartite graphs, but they can be defined, perhaps even more intuitively, as hypergraphs [2]. A hypergraph H is defined as a pair H=(X, A), where: X is a set of objects called the vertices of H, and A is a family of multisets with objects in X, called the hyperedges of H. If the multisets in A are ordered then the hypergraph is called an ordered hypergraph.

The ordered hypergraph can conveniently replace the bipartite graph in the definition of the basic conceptual graph. Thus, a conceptual hypergraph H, on a vocabulary $V=(T_C,T_R,E)$ is a triple $H = (X, A, \eta)$, where: (X, A) is an ordered hypergraph, and η is the labelling application that attaches to a vertex x a pair of labels $(type(x),marker(x))\in(T_C\in(E\cup\{*\}))$ and to a hyperedge r, a label $type(r)\in T_R$.

Each vertex of a conceptual hypergraph represents a conceptual entity and each of its hyperedges represents a conceptual relation with the arity given by the cardinality of the hyperedge. Therefore, there is a natural correspondence between a conceptual hypergraph and a basic conceptual graph. If H=(X, A, η), is a conceptual hypergraph then the conceptual graph equivalent to H is G=(C,R,V,η), with the property that, C is isomorphic to X, R is isomorphic to A, and the labelling application η is common.

We observe that if H=(X, A, η), is a conceptual hypergraph, then each element of the set A represents a StarCG, that is, a CG formed by a single relation node together with its neighbors. To simplify our approach, we will specify each element $a\in A$, by a list $\alpha=[a_1,\ldots a_n,r]$, where $a_1,\ldots a_n$ are nodes that represent conceptual entities and r is a node that represents a conceptual relation. We will denote the set of conceptual nodes components of a, by $C(\alpha)=\{a_1,\ldots a_n\}$, and the relational node component of a by $R(\alpha)$.

In the following, we will consider only hypergraphs that do not have isolated conceptual nodes. Under these conditions, a conceptual hypergraph without isolated vertices is completely specified by the set A together with the labelling application η. Therefore, we will specify a conceptual hypergraph H without isolated nodes as a pair H=(A, η), where the meanings of A and η are as specified above.

Since the pair H=(A, η), represents a conceptual graph where A is the set of corresponding StarCGs and η is the labelling application, we will call the pair H=(A, η) with the meaning specified above a conceptual graph.

Example 1. The conceptual graph in Figure 1 can be specified as a pair H=(A, η), where: $A=\{[a_2,a_1,r^1], [a_1,a_3,a_2,r^2], [a_3,a_4,r^3]\}$ and η is defined as follows: η(a1)=" Profesor", η(a2)="UndergradStudent", η(a3)="UndergradCourse", η(a4)="BachelorDegree", η(r1)=" HasMentor", η(r2)="Teaches", η(r3)="Curriculum".

If $H_1=(A_1, \eta_1)$ and $H_2=(A_2, \eta_2)$, are two conceptual graphs, without isolated vertices, then a homomorphism from H_1 to H_2, is a map $\varphi:A_1\rightarrow A_2$, which maps each StarCG $\alpha=[a_1,\ldots a_n,r^\alpha]\in A_1$, to a StarCG $\beta=[b_1,\ldots b_n,r^\beta]\in A_2$, such that $b_i=\varphi(a_i)$, $\eta_1(a_i)\geqslant\eta_2(b_i)$ for all i=1,…,n and $r^\beta=\varphi(r^\alpha)$, $\eta_1(r^\alpha)\geqslant\eta_2(r^\beta)$. It is obvious that the problem regarding the existence of a homomorphism between two conceptual graphs is equivalent to the problem regarding the existence of a homomorphism between two conceptual hypergraphs.

An isomorphism between two conceptual graphs $H_1=(A_1, \eta_1)$ and $H_2=(A_2, \eta_2)$, is a bijective mapping φ, from A_1 to A_2, with the property that for any StarCG $\beta=\varphi(\alpha)$, we have:

i) if $a_i \in C(\alpha)$ and $r^{\alpha}=R(\alpha)$ then $b_i \in C(\beta)$ and $r^{\beta}=R(\beta)$

ii) if $b_i \in C(\beta)$ and $r^{\beta}=R(\beta)$ then $a_i \in C(\varphi^{-1}(b_i))$ and $r^{\alpha}=R(\varphi^{-1}(r^{\beta}))$

iii) for all components $x \in C(\alpha) \cup R(\alpha)$ we have $\eta_1(x)= \eta_2(\varphi(x))$

Under these conditions if H=(A, η), is a conceptual graph then a conceptual subgraph H' of H, is a pair H'=(A', η') where $A' \subseteq A$ and η' is the restriction of η to the components that appear in A'. If A' contains a single StarCG, we will also use the name StarCG for the conceptual graph H'.

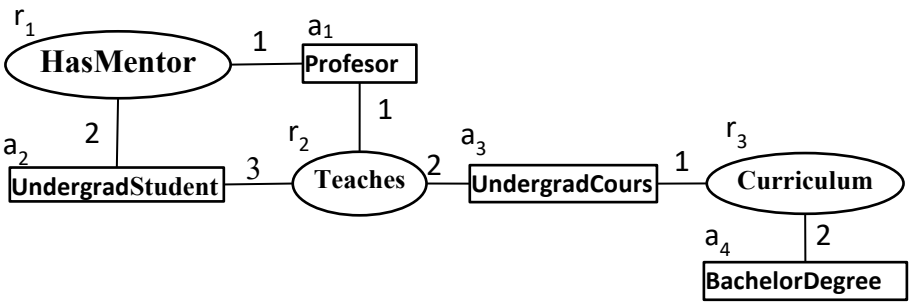


Figure 1. H – Conceptual Graph

Example 2. The conceptual graph H'=(A', η'), where: A'={[a_2,a_1,r^1], [a_1,a_3,a_2,r^2], [a_3,a_4,r^3]} and η' is defined as follows: η'(a1)="Profesor", η'(a2)="UndergradStudent", η'(a3)="UndergradCourse", η'(a4)="BachelorDegree", η'(r1)=" HasMentor", η'(r2)="Teaches", η(r3)="Curriculum" is a conceptual subgraph of H.

In this paper we will consider that all the conceptual graphs we use are brought to normal form. Normal form is not a restriction because any conceptual graph can be brought to normal form in polynomial time [2]. We observe that if $\eta=(\eta_C,\eta_R,\eta_\Gamma)$, is the marking application of a normalized conceptual graph then the component η_C is injective.

Next, we will specify a finite automaton associated with a conceptual graph G, which recognizes all the component StarCGs of G. If H=(A, η) is a conceptual graph, then each StarCG, $\alpha_k \in A$, is of the form $\alpha_k=[a_{k_1}, a_{k_2}, \ldots, a_{k_{|r_k|}}, r_k] \in A$, where we denote by $|r_k|$ the arity of r_k. The finite automaton attached to the StarCG component α_k (Figure 2) is defined as follows:

$M_k = (I_k, S_k, f_k, s_{0,k}, F_k)$, where:

$I_k=\{\eta(a_{k_1}), \eta(a_{k_2}), \ldots, \eta(a_{k_{|r_k|}}), \eta(r_k)\}$,

$S_k=\{a_{k_1}, a_{k_2}, \ldots, a_{k_{|r_k|}}, r_k\}$, $F_k=\{r_k\}$,

$f_k : S_k \times I$ is the transition function defined as follows: $f_k(s,\eta(a_k))=a_k$;

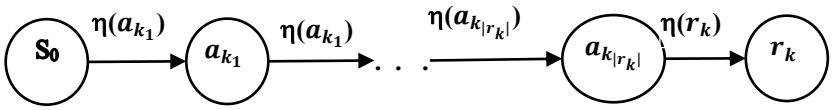


Figure 2. The finite automaton that recognizes a StarCG

For each StarCG component of a conceptual graph we can construct an elementary finite automaton.

Example 3. For the StarCG components of the conceptual graph H in Figure 1 we can construct the finite automata in Figure 3, defined as follows:

$M_1=((I_1, S_1, f_1, s_{0,1}, F_1)$, where:

I_1={Profesor, UndergradStudent, HasMentor},

$S_1=\{S_0, a_1, a_2, r_1\}$, $F_1=\{r_1\}$,

$f_1:S_1\times I_1$ is the transition function defined as can be seen in Figure 3.

$M_2=((I_2, S_2, f_2, s_{0,2}\ F_2)$, where:

I_2={Profesor, UndergradCours, UndergradStudent, Teaches},

$S_2=\{S_0, a_1, a_3, a_2, r_2\}$, F2=$\{r_2\}$,

$f_2:S_2\times I_2$ is the transition function defined as can be seen in Figure 3.

$M_3=((I_3, S_3, f_3, s_{0,3}\ F_3)$, where:

I_3={UndergradCours, BachelardDegree, Curriculum},

$S_3=\{S_0, a_3, a_4, r_3\}$, F3=$\{r_3\}$,

$f_3:S_3\times I_3$ is the transition function defined as can be seen in Figure 3.

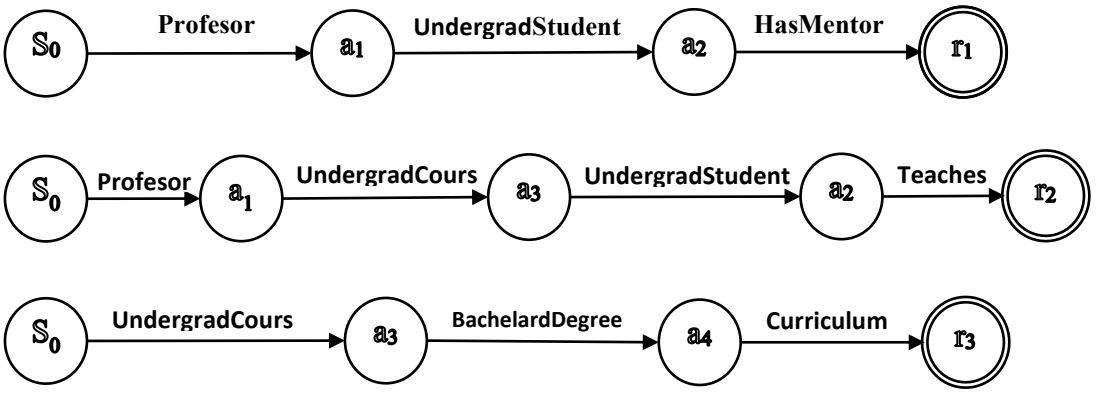


Figure 3. Finite automata associated with the conceptual graph H

Let there be two finite automata $M_1 = (I_1, S_1, f_1, F_1)$ and $M_2 = (I_2, S_2, f_2, F_2)$. The direct sum of the two automata is the automaton $M=M_1\oplus M_2=(I_1\cup I_2, S_1\sqcup S_2, f_1\oplus f_2, F_1\sqcup F_2)$, where we denoted by $\sqcup$ the disjoint sum of two sets of states, and $f_1\oplus f_2$ and $g_1\oplus g_2$ are defined as follows:

$$f_1\oplus f_2(s,i)=\begin{cases} f_1(s,i) \text{ if } s\in S_1 \\ f_2(s,i) \text{ if } s\in S_2 \end{cases}$$

$$g_1\oplus g_2(s,i)=\begin{cases} g_1(s,i) \text{ if } s\in S_1 \\ g_2(s,i) \text{ if } s\in S_2 \end{cases}$$

Example 4. The direct sum of the three automata specified in Example 3, $M=M_1\oplus M_2\oplus M_3$ can be seen in Figure 4.

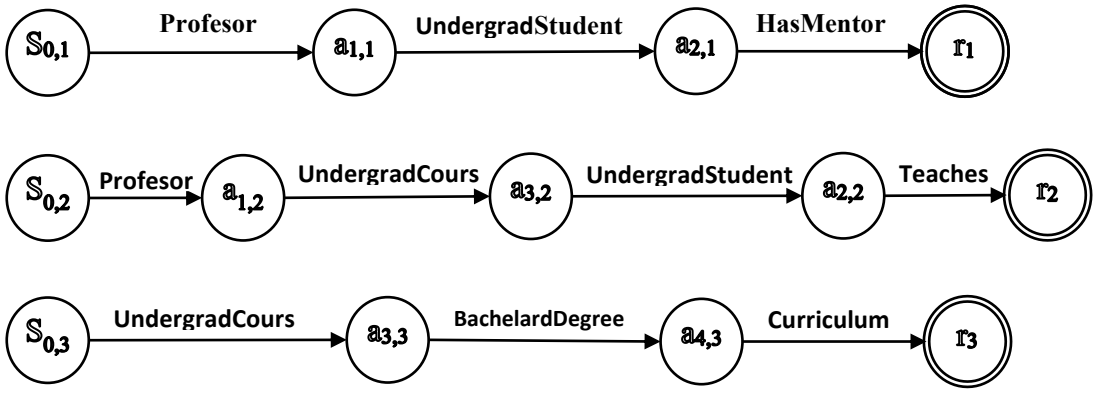


Figure 4. Direct sum of the automata in Figure 3

Direct sum only renames the states of the automata so that they are distinct. The direct sum of two conceptual graphs is a specialization operation that does not connect the two graphs. The direct sum of two finite automata behaves in the same way. We present below an algorithm that connects several automata into a single finite automaton. Moreover, the resulting automaton is deterministic.

The CONECT algorithm. This algorithm connects several finite automata into a single deterministic finite automaton.

Input: The input to the algorithm is represented by a set of finite automata M_k, $1 \le k \le n$.

Output: The initial deterministic finite automaton $M=(I, S, f, S_0, F)$ equivalent to the direct sum of the finite automata Mk, $1 \le k \le n$, from the input.

Method: In the first step, the algorithm calculates the nondeterministic finite automaton $N=(I_N, S_N, f_N, F_N)$, as the direct sum of all the automata M_k, $1 \le k \le n$. Then, the algorithm creates new states that are subsets of the set S_N, of states of the automaton N. The algorithm also defines a new transition function between the new states.

P1. The direct sum $N=M_1 \oplus M_2 \oplus \ldots \oplus M_n=(I_N, S_N, f_N, F_N)$ is made. $I \leftarrow I_N$ is made.

P2. The initial state is defined $S_0=\{s_{0,1}, s_{0,2}, \ldots, s_{0,n}\}$, where each $s_{0,k}$ is the initial state of the automaton M_k. Make $S \leftarrow \{S_0\}$ and $m \leftarrow 0$.

P3. As long as there are unmarked states in S, execute steps P4 and P5.

P4. Choose an unmarked state s from S and mark it.

P5. For each input $e \in I_N$, calculate the set $P_e=\{x \in S_N | \ x= f_N(s,e)\}$. If $P_e \neq \emptyset$, make $m \leftarrow m+1$, $S_m= P_e$, $S \leftarrow S \cup \{S_m\}$. Define $f(s,e) \leftarrow S_m$.

Example 5. The result of applying the CONECT algorithm for the automata specified in Example 3 can be seen in Figure 5.

If H is a conceptual graph then we will denote by M(H), the automaton attached to H. The automaton M(H) is obtained by applying the CONECT algorithm to the elementary automata corresponding to the component StarCGs of H.

We will say that a conceptual graph $G=(A_G, \eta_G)$ is recognized by the automaton M if the automaton M recognizes all the components of the StarCGs in AG.

It is obvious that if G=(A_G, η_G) and H=(A_H, η_H) are conceptual graphs and the conceptual graph G is recognized by the automaton M(H), then G is isomorphic to a conceptual subgraph of H.

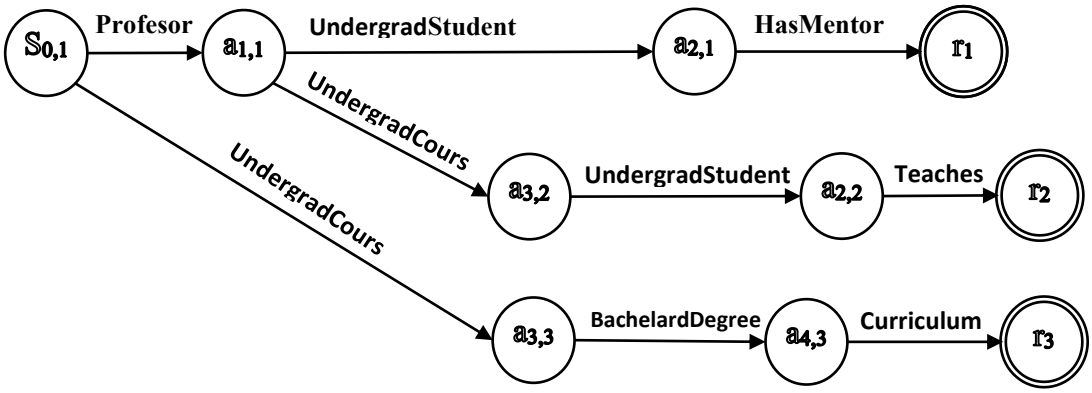


Figure 5. The finite automaton associated with the conceptual graph H

4 Finite automata and conceptual graph homomorphisms

As we saw in the previous section, a bijective homomorphism between two conceptual graphs is an isomorphism between the two bipartite graphs in the definition of a conceptual graph that preserves labels. Also in the previous section, we saw that to a conceptual graph H we can associate a finite automaton that recognizes any conceptual graph isomorphic to a conceptual subgraph of H.

Coreference and generalization relations are of utmost importance in specifying and determining inference between logical formulas expressed by conceptual graphs [14]. The generalization relation is defined at the level of conceptual and relational labels and extends to the level of conceptual graphs as a subsummation relation.

We have several important results regarding the connection between the generalization relation and conceptual graph homomorphisms [2]. One of these results is that a conceptual graph G is a generalization of a conceptual graph H if and only if there is a homomorphism from G to H. Also, a conceptual graph G is a generalization of a conceptual graph H if and only if H is a specialization of G. In this section we will define a finite automaton associated with a conceptual graph H that recognizes any conceptual graph G, with the property that there is a homomorphism of conceptual graphs from G to H. Our approach is based on the observation that, if we have two conceptual graphs G and H, then there is a homomorphism from G to H, if and only if there is a generalization of H isomorphic to G.

First, we will define three generalization operations applicable to StarCGs. The generalization operations for StarCGs are unary operations that apply to a StarCG G and return another StarCG H that is a generalization of G, that is, there is a homomorphism from H to G. These operations are, Split, Duplicate and Increase and will be applied to a StarCG G=(A={α},η) as described below.

The operation Split(G,c,c_1,c_2,l_1,l_2) receives as input a StarCG G, a conceptual node c existing in G, two names c_1 and c_2 for the new conceptual nodes and two new labels $l_1 \neq l_2$; $l_1, l_2 \notin C(\alpha)$, $l_1 \geq \eta(c)$ si $l_2 \geq \eta(c)$. The operation will replace the conceptual node c

with two conceptual nodes c_1 and c_2, of the same type and will modify the labeling application as follows: $\eta(c_1)=l_1$ and $\eta(c_2)=l_2$.

We note that a concept node cannot be divided into two with the same label because the result is no longer normalized. A conceptual node can be divided into two concepts with different labels but both greater than or equal to the StarBG label.

The Duplicate(G,r,r_1) operation receives as input a StarCG G, a relational node r existing in G and a name r_1 for a new relational node. The operation simply adds a new relational node r_1, of the same type as r, with the same label and with the same neighbours.

The Increase(G,x,l) operation receives as argument a StarCG G, a conceptual or relational node x and a label $l \notin C(\alpha)$, $l>\eta(x)$. The operation modifies the labelling application such that $\eta(x)=l$.

Example 6. Consider the StarCG G=(A, η), from Figure 6, where: A={[a_1,a_3,a_2,r^2] } and η is defined as follows: $\eta(a_1)$="Profesor", $\eta(a_2)$="UndergradStudent", $\eta(a_3)$="UndergradCourse", $\eta(r_2)$="Teaches".

Suppose that in the vocabulary of the conceptual graph G we have the following hierarchy of concepts: TeachingStaff ⩾ Profesor, Student ⩾ UndergradStudent, Course ⩾ UndergradCourse, Degree ⩾ BachelorDegree , Attends ⩾ Teaches.

With this hierarchy the elementary automata in Figure 7 are the automata associated with some examples of generalizations of G.

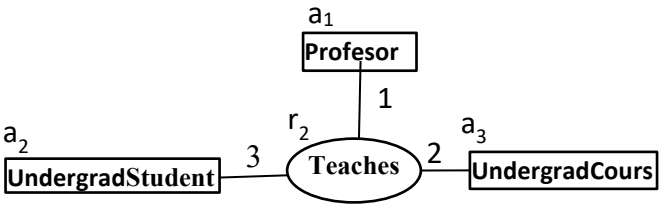


Figure 6. Elementary Conceptual Graph (StarCG)

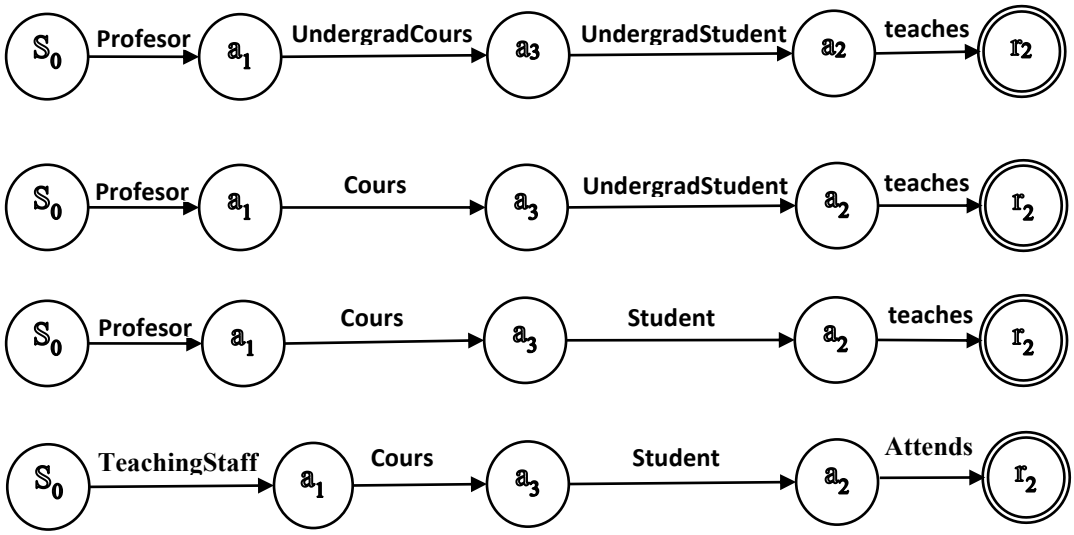


Figure 7. Examples of elementary automata attached to generalizations

Let us now consider a conceptual graph H=(A, η) , where A={$\alpha_1,\alpha_2,\ldots,\alpha_n$}. Then the conceptual graph H is the union of its constituent StarCGs of the form $H_k=(A_K, \eta_k)$, where each component A_K= {$\alpha_k=[a_{k_1}, a_{k_2}, \ldots, a_{k_{|r_k|}}, r_k]$} and η_k is the restriction of the map η to A_K.

Let H=(A, η), A={$\alpha_1,\alpha_2,\ldots,\alpha_n$} be a conceptual graph and $H_k=(A_K, \eta_k)$, $1\leq k\leq n$. We will denote by $\boldsymbol{\mathcal{G}}_H$ the set of all normalized conceptual graphs that generalize StarCG components of H:$\boldsymbol{\mathcal{G}}_H$={G| ∃ k≥1 such that G⩾ H_k}.

If we have a conceptual graph G=(A, η) then for any two StarCGs α,β∈A we have $\eta_\alpha(x)= \eta_\beta(x)= \eta(x)$ ∀x∈C(α)∩C(β). Therefore, two StarCGs, $G=(A_G, \eta_G)$ and $H=(A_H, \eta_H)$ can be assembled into a single conceptual graph (A, η) if and only if $\eta_G(x)= \eta_H(x)$ ∀x∈C(A_G)∩C(A_H) si $\eta_G(r)= \eta_H(r)$ ∀x∈R(A_G)=R(A_H).

We will now introduce a partial union operation of two normalized conceptual graphs.

Let $G=(A_G,\eta_G)$ and $H=(A_H,\eta_H)$ be two conceptual graphs such that for any two StarCGs α∈A_G and β∈A_H we have $\eta_G(x)= \eta_H(x)$ ∀x∈C(α)∩C(β). The union of G with H is the conceptual graph P=($A_G \cup A_H$, η) where η is defined as follows:

$$\grave{}\eta(x)=\begin{cases}\eta_G(x)\ if\ x \in C(G)\cup R(G)\\ \eta_H(x)\ if\ x \in C(H)\cup R(H)\\ \eta_G(x)\ if\ x \in \big(C(G)\cap C(H)\big)\cup\big(R(G)\cap R(H)\big)\end{cases}$$

We must note that the union of two conceptual graphs can only be done when the two labelling applications coincide on the common conceptual or relational nodes. In our case we will also add the condition that the result of the union of two normalized conceptual graphs is also a normalized conceptual graph.

For a conceptual graph H, we now consider the set of all conceptual graphs that can be obtained by the operation of partial union of graphs in $\boldsymbol{\mathcal{G}}_H$, that is, the closure of this set which we denote by $\boldsymbol{\mathcal{G}}_H^*$. It is obvious that $\boldsymbol{\mathcal{G}}_H^*$ contains all normalized generalizations of H.

We have a very important result for our approach, namely that, if we have two conceptual graphs G and H, then there is a homomorphism from G to H, if and only if there is a generalization of H isomorphic to G [2].

Using the CONECT algorithm applied to the set $\boldsymbol{\mathcal{G}}_H$, we obtain an automaton that recognizes all conceptual graphs in $\boldsymbol{\mathcal{G}}_H^*$. We denote by $\boldsymbol{\mathcal{M}}$(H) the automaton: $\boldsymbol{\mathcal{M}}$(H)=CONECT($\boldsymbol{\mathcal{G}}_H$) and call it a subsummation automaton.

We have seen that a normalized conceptual graph $H=(A_H, \eta_H)$ is recognized by the automaton M if the automaton M recognizes all StarCG components in A_H. From this it follows that, if $G=(A_G, \eta_G)$ and $H=(A_H, \eta_H)$ are conceptual graphs and the conceptual graph G is recognized by the automaton $\boldsymbol{\mathcal{M}}$(H), then G is isomorphic to a conceptual graph $G_1\in\boldsymbol{\mathcal{G}}_H^*$, that is, there is an isomorphism η:G→G_1. But since G_1 is a generalization of H, it follows that there is a natural homomorphism π:G_1→H. It follows that the map φ=πoμ, is a homomorphism from the conceptual graph G to the conceptual graph H.

5 Observations and conclusions

The ontological component of a knowledge base contains general knowledge about the domain, which is called ontological knowledge. But knowledge bases contain, in addition to ontological knowledge, various facts and statements about specific entities, which are logical formulas with truth values, and which are called factual knowledge.

Thus, a knowledge base models an open world, which includes information about a potentially infinite set of entities, by using universal quantifiers applied to unknown individuals. The main purpose of a knowledge base is achieved by the associated inference engine, which deduces new knowledge by logical reasoning on some implicit information in the knowledge base.

In the case of conceptual graphs, logical inference is based on homomorphisms of conceptual graphs. This paper proposes the use of finite automata to determine the existence of homomorphisms between two conceptual graphs. Our approach involves moving the complexity of the problem from the phase of determining the homomorphisms to the phase of building the knowledge base.

We note that in our approach the knowledge base must contain, in addition to the conceptual graph, the attached finite subsummation automaton. We can also observe that the subsummation automaton can reach a very large number of states. To eliminate this shortcoming, this automaton will have to undergo an optimization process that will be followed in subsequent works.

References

[1] Sowa, John F., “Conceptual graphs for a database interface, IBM Journal of Research and Development”, 20:4, 336-357, (1976).
[2] Michel Chein, Marie-Laure Mugnier, “Graph-based Knowledge Representation, Computational Foundations of Conceptual Graphs”, Springer-Verlag London Limited, (2009).
[3] Thomas R. Gruber, “A translation approach to portable ontology specifications”, Knowledge Acquisition Volume 5, Issue 2, June 1993,Pages 199-220, (1993)
[4] Lehmann D, Magidor M, “What does a conditional knowledge base entail?” Artificial Intelligence, Volume 55, Issue 1, May 1992, Pages 1-60, (1992)
[5] Bienvenu M, Ortiz M, “Ontology-mediated query answering with data-tractable description logics”. Lecture notes of the 11th international reasoning web summer school. LNCS, vol 9203, Springer, Berlin, pp 218–307, (2015)
[6] 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).
[7] Daniel-Cristian Craciunean, “Categorical Mechanisms in Multi-level Modeling Methods”, Publishing House of the "Lucian Blaga" University of Sibiu, (2023).
[8] Ehrlinger L, Wöß W, “Towards a definition of knowledge graphs”, Joint Proceedings of the Posters and Demos Track of 12th International Conference on Semantic Systems - SEMANTiCS2016 and 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS16) 48(1–4):2, (2016)

[9] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z., "DBpedia: A Nucleus for a Web of Open Data." In: Aberer, K., et al. The Semantic Web. ISWC ASWC 2007 2007. Lecture Notes in Computer Science, vol 4825. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-76298-0_52, (2007)
[10] Rebele T, Suchanek F, Hoffart J et al, "Yago: a multilingual knowledge base from wikipedia, wordnet, and geonames." In: International semantic web conference, Springer, p 177–185, (2016)
[11] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi, "WordNet::Similarity - Measuring the Relatedness of Concepts." In Demonstration Papers at HLT-NAACL 2004, pages 38–41, Boston, Massachusetts, USA. Association for Computational Linguistics, (2004).
[12] K. V. N Sunitha, N. Kalyani, "Formal Languages and Automata Theory", Pearson India Education Services Pvt. Ltd, (2015)
[13] Bordes A, Weston J, Collobert R et al, "Learning structured embeddings of knowledge bases", Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI, San Francisco, California, USA, August 7-11, (2011)
[14] 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, (2017)

DOI: 10.2478/ijasitels-2025-0007

Implementation of Heuristic Algorithms for Simulating Crisis Situations in the Medical System

Pătrăuşanu Andrei, Buta Andra-Paraschiva, Florea Adrian

Abstract

Healthcare systems face significant challenges during crises such as pandemics or mass-casualty events, where resource shortages and patient overflow require rapid, optimized decisions. This paper proposes a simulation-based approach to model and improve hospital resource allocation using a heuristic method—Genetic Algorithms (GA). The aim is to explore how intelligent algorithms can support decision-making under pressure by assigning patients to limited ICU beds and available doctors, considering constraints such as treatment duration, medical priority, and resource availability.

The core method involves evolving allocation strategies over multiple generations, using tournament selection, crossover, mutation, and fitness-based evaluation to optimize both resource usage and patient coverage. The simulation is implemented as a desktop application in C#, with a SQL Server database and an interactive GUI that allows users to run scenarios, configure parameters, and visualize outcomes.

Compared to a First-Come-First-Served (FCFS) baseline, the GA consistently achieves higher efficiency, treating more high-priority patients and reducing resource bottlenecks. The original contribution lies in the dual-resource optimization model and its integration into a flexible, user-friendly tool. Results demonstrate that heuristic-driven simulations can support better planning and training in emergency healthcare environments.

Keywords: genetic algorithm; heuristic optimization; hospital resource allocation; crisis simulation; healthcare systems; emergency planning

1. Introduction

In recent years, healthcare systems around the world have been repeatedly tested by large-scale crises such as pandemics, natural disasters, and other emergency events. These scenarios typically lead to significant resource imbalances, overwhelmed medical staff, and critical delays in patient care. As hospitals operate under intense pressure, the need for intelligent tools that can assist with real-time decision-making becomes increasingly urgent.

Traditional resource allocation methods, such as First-Come-First-Served (FCFS), often fail to deliver efficient or equitable outcomes in such high-demand situations. In contrast, heuristic and metaheuristic algorithms—particularly Genetic Algorithms (GAs) [1] —have shown great promise in solving complex scheduling and optimization problems where multiple constraints must be balanced.

This paper presents a simulation-based approach for improving hospital resource allocation during crisis conditions using a Genetic Algorithm. The proposed system

models real-world constraints by assigning patients to limited ICU beds and medical staff, taking into account treatment duration, resource availability, and patient priority.

The goal of this research is to demonstrate how intelligent algorithms can support healthcare professionals in planning and training for crisis events by simulating alternative allocation strategies. A desktop application was developed for this purpose, offering an interactive interface, parameter configuration, and visual comparison between heuristic and traditional methods.

2. Related Work

Numerous studies have addressed the problem of resource allocation in healthcare systems, especially during crisis scenarios such as pandemics or mass-casualty events. Traditional scheduling methods are often insufficient in such contexts due to the high complexity and uncertainty involved. As a result, heuristic and metaheuristic algorithms have emerged as effective tools for optimizing resource distribution under pressure.

In [2] the authors proposed a Genetic Algorithm (GA)-based scheduling model for patient assignment, treating the problem similarly to a job-shop scheduling task. Their results showed improved efficiency and fairness compared to manual allocation methods. The authors of [3] extended this approach by integrating realistic medical workflows, emphasizing better hospital stay management through evolutionary algorithms. Other studies have explored alternative metaheuristics. The work from [4] employed Particle Swarm Optimization (PSO) for inter-hospital medical staff coordination, while [5] proposed a hybrid system combining deep learning with genetic search to improve dynamic healthcare planning. Additionally, [6] adapted GA for emergency nurse scheduling, incorporating both hard and soft constraints to maintain fairness and coverage.

In [7], the authors used the NSGA-II [8] algorithm to optimize the bed allocation in ICU, considering different levels of uncertainty: the amount of emergency patients and their corresponding length of stay, but also the length of stay of the current ICU patients. In [9] the authors present a hybrid approach of a Discrete Event Simulation and a GA (genetic algorithm) in order to minimize the waiting time for rehabilitation by varying the care unit capacity. A similar approach is present in [10], where the main optimization flow is driven by NSGA-II algorithms and the solutions evaluated by the Discrete Event Simulation. The results showed a promising alternative to the bed allocation in Brazil hospitals. In [11], a genetic algorithm approach is followed to obtain a score-based priority mechanism to efficiently use the healthcare resources in hospital units.

Other similar works, such as [12] use concepts such as operation research [13] to evaluate the responsiveness in healthcare systems under the pressure of a medical crisis. In [14] a hybridization of Genetic Algorithm and extreme learning machine to identify optimal medical events that contribute to mortality of patients is described as a part of

a larger prediction mechanism. Works such as [15] exploits the parallelism of individuals evaluation in genetic algorithms in the field of patients distribution. In [16] two strategies such as first come first served and shortest processing time rules are used in conjunction within an heuristic algorithm. Their approach is used to schedule patient treatment depending on the bed allocation in a hospital in Australia. The results concluded a reduction of 8% of waiting time.

A comprehensive study of resource allocation managements in crisis circumstances has been accomplished in [17]. One of the main conclusions of the paper was that the analyzed SRA (Scarce Resource Allocation) protocols resulted in ambiguous lottery system allocation, hence the need for standardization and refinement. The impact of COVID-19 struck many areas of treatment and healthcare. Thesis [18] emphasizes the use of a genetic algorithm with a local search operator for finding optimal configuration of a surrogate of an initial agent -based system used for hospital scheduling areas.

It is also worth mentioning frameworks and tools [19], [20], [21] which prove the effectiveness of implementation and integration of heuristic search in fields such as length of stay, bed allocation scheduling etc.

Despite these contributions, most existing systems either target a single resource type or are limited to theoretical frameworks. In contrast, the present work introduces a dual-resource allocation model (beds and doctors), integrated into a simulation tool with a user-friendly interface that enables practical experimentation and comparative analysis.

3. Methodology

3.1 Problem Formulation

This study addresses the allocation of limited hospital resources—specifically ICU beds and doctors—during crisis scenarios. The goal is to create an optimized allocation plan that reduces waiting time and maximizes the treatment of high-priority patients. The problem is modeled as a constrained multi-objective optimization task, considering both treatment priorities and total resource usage.

3.2 System Architecture

The proposed solution is implemented as a Windows Forms desktop application using C#. It integrates with a SQL Server database to store and retrieve entities such as patients, beds, doctors, and allocations.

The application follows a three-layer architecture:

- Presentation Layer: user interface for simulation control.
- Business Logic Layer: implementation of Genetic Algorithm and evaluation logic.
- Data Access Layer: interaction with the underlying SQL database.

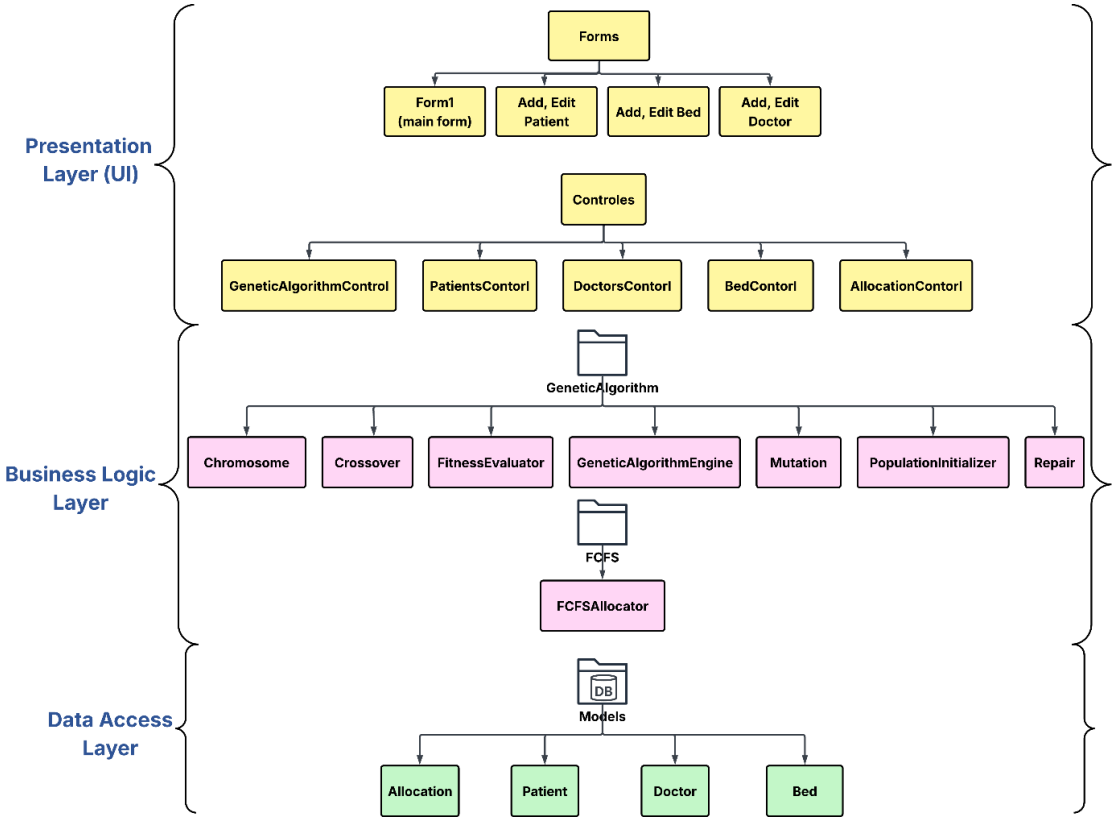


Figure 1 – System Architecture Overview

3.3 Data Representation

Each patient record includes multiple fields: a unique ID, full name, age, time of arrival in the hospital system, treatment priority (a numerical value reflecting urgency), bed duration (estimated time spent occupying a bed), and doctor duration (estimated time required by a doctor for treatment).

Beds and doctors are treated as independent and limited resources. Each bed or doctor can attend to only one patient at a time, and patients are treated in the order in which they are assigned. This reflects real-world constraints in hospital settings where resource conflicts must be avoided.

To ensure simulation accuracy and solution feasibility, a valid allocation must assign every patient exactly once to both a bed and a doctor. Partial assignments or duplicate assignments are considered infeasible and are corrected during the repair phase of the algorithm. The data model enforces these constraints to simulate realistic medical workflows, supporting the calculation of metrics like treated patients, resource utilization, and makespan.

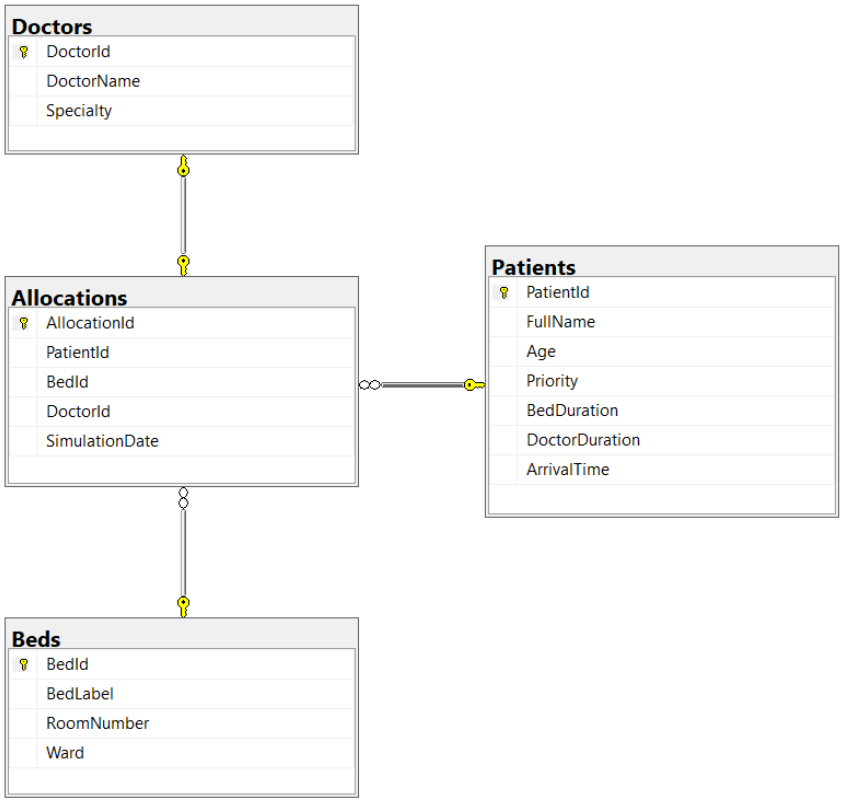


Figure 2 – Database Entity Relationship Diagram

3.4 Chromosome Representation

Each potential solution, or chromosome, encodes a full allocation plan of patients to two types of hospital resources: ICU beds and doctors. The chromosome has a dual-segment structure:

- BedSegments: a list of patient IDs assigned to each bed.
- DoctorSegments: a list of patient IDs assigned to each doctor.

This representation ensures that each patient receives both required resources, and that resource usage and treatment times can be calculated accurately.

At runtime, a chromosome might look like this:

- BedSegments[0] = [2, 5, 7] → Bed 1 is assigned to Patients 2, 5, and 7
- DoctorSegments[0] = [5, 2] → Doctor 1 is treating Patients 5 and 2

This mapping allows the algorithm to simulate how patients are distributed across limited resources and how much time each resource will be occupied.

3.5 Fitness Function

The fitness function evaluates each chromosome based on:

- The number of patients successfully treated (i.e., assigned both a bed and a doctor),
- The total priority of treated patients (higher-priority patients improve the score),
- The overall makespan (total time resources are occupied), which the algorithm aims to minimize.

The goal is to maximize treated priority while minimizing resource overload. The final fitness is normalized between 0 and 1.

$$\text{fitness} = \frac{1}{\text{makespan}} * \text{total\_priority}$$

$$normalized\_fitness \frac{fitness}{max\_possible\_fitness}$$

Where:

- makespan = the maximum time any resource (bed or doctor) is occupied
- total_priority = sum of priorities for all patients successfully treated
- max_possible_fitness = a normalization constant calculated from the ideal scenario

3.6 Genetic Operators

The algorithm follows the classical GA steps:

1. **Initialization**: A population of chromosomes is randomly generated by distributing patient IDs across resources without duplication.
2. **Selection**: Tournament selection is used to choose parent chromosomes based on fitness.
3. **Crossover**: A modified Partially-Mapped Crossover (PMX) combines two parents to produce offspring, applied separately to both segments.
4. **Mutation**: A uniform swap mutation randomly exchanges two patients in a segment, with a configurable mutation rate .
5. **Repair**: A repair function ensures all patients are assigned exactly once to each resource.
6. **Elitism**: The best solution of each generation is carried over to preserve high-quality solutions.

3.7 Parameters

The performance of the genetic algorithm depends significantly on the configuration of its parameters. These parameters influence convergence speed, solution quality, and exploration capability. The following parameters are used in this study:

- Population Size: Determines the number of chromosomes per generation. A higher value increases diversity but also computation time. In the simulations, values between 20 and 200 were tested.
- Number of Generations: Sets how many iterations the algorithm will perform. Higher values allow more time for convergence.
- Mutation Rate: Defines the probability of mutation in each chromosome.
- Tournament Size: Refers to the number of chromosomes selected at random for each tournament.

- Elitism: Ensures the best chromosome from each generation is preserved in the next. This prevents loss of high-quality solutions.
- Crossover Method: PMX was used due to its ability to preserve relative order and segment structure.

Parameter tuning was done manually by testing various combinations and observing the impact on solution quality.

Parameter	Min	Max	Default
Population Size	20	200	50
Number of Generations	50	500	100
Mutation Rate (%)	5%	40%	20%
Tournament Size	2	20	5

Table 1 – Genetic Algorithm Parameters

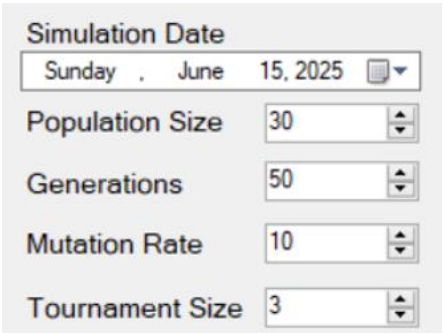


Figure 3 – Genetic Algorithm Parameters Panel

3.8 Simulation Workflow

Before running a simulation, the application loads data from the database (patients, doctors, beds) and user configures parameters such as population size, number of generations, mutation rate, and tournament size.

The algorithm runs for a predefined number of generations, during which fitness scores are tracked and visualized. Upon completion, the best solution is stored in the database and can be reviewed using the interface. The FCFS strategy is also applied as a baseline for performance comparison.

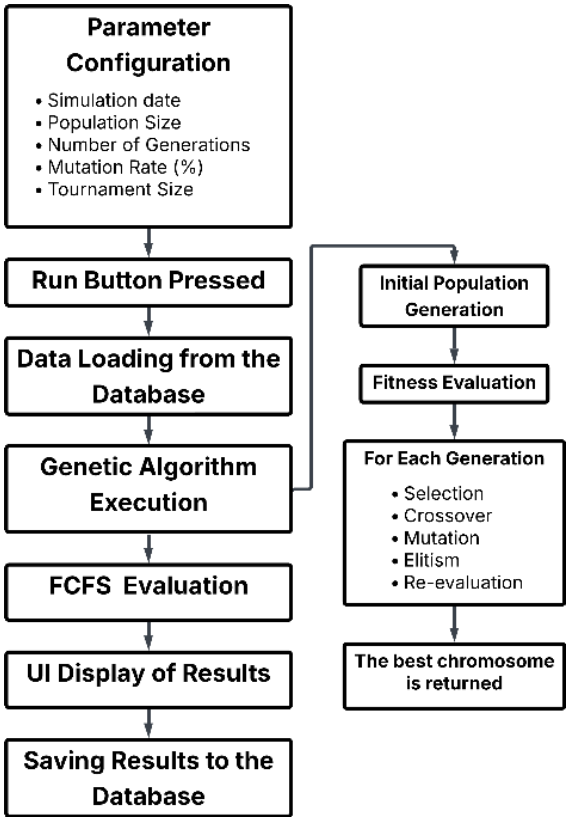


Figure 4 – Execution Flow

3.9 Baseline for Comparison – FCFS

To measure the effectiveness of the genetic algorithm, the application also includes a First-Come-First-Served (FCFS) baseline. This simple method assigns patients to beds and doctors in arrival order and does not perform optimization. Its fitness score is used as a reference point in simulation results.

The table below summarizes the main differences between the Genetic Algorithm and the First-Come-First-Served approach implemented in the application.

Criterion	**Genetic Algorithm (GA)**	**First-Come-First-Served (FCFS)**
Optimization goal	Minimize makespan, maximize treated priority	Assign based on arrival time only
Resource awareness	Balances usage across doctors and beds	Does not optimize resource distribution
Adaptability	Adapts to patient severity and resource duration	Fixed and rigid

Solution quality	Evolves better results over generations	Typically lower efficiency
Conflicts or duplicates	Avoided using repair function	No conflict resolution
Performance indicator	Fitness score (0 to 1), plotted per generation	One-time result, used as a benchmark

Tabel 2 – Comparative summary between GA and FCFS

4. Results and discussions

The effectiveness of the proposed genetic algorithm (GA) was evaluated by comparing it with a baseline First-Come-First-Served (FCFS) approach under various simulation scenarios. The results highlight improvements in patient prioritization, resource utilization, and overall system efficiency.

4.1 Experimental Setup

The dataset used for simulation includes a full week of data, with 40 patients per day, and a predefined number of 7 beds and 5 doctors. Each patient has attributes like arrival time, priority, bed usage duration, and doctor usage duration—these durations are calculated automatically based on a formula that considers patient priority and some randomness. All values are synthetic and were generated to reflect realistic variation in hospital demands.

$$\text{BedDuration} = \text{Priority} \times 30 + \text{random}(0, 30)$$

$$\text{DoctorDuration} = \text{Priority} \times 10 + \text{random}(0, 30)$$

The number of available resources (beds and doctors) can be modified at any time by the user, by adding or removing records directly from the application's interface. This allows running simulations under different stress levels, such as low capacity or patient surges.

Before each simulation, the user selects the parameters that will control the behavior of the genetic algorithm.

This flexibility allows users to experiment with different optimization strategies and evaluate how the algorithm responds.

4.2 Parameter Variation Impact

To analyze how GA parameters affect performance, various test scenarios were executed by adjusting population size, mutation rate, tournament size, and number of generations. All configurations were compared to the static baseline of FCFS (67.73% fitness).

Population Size	Generations	Mutation Rate (%)	Tournament Size	Fitness Max (%)	Fitness Avg (%)	FCFS
30	50	10	3	79.89	78.16	67.73
30	100	10	3	76.81	76.2	67.73
50	50	20	4	79.89	78.53	67.73
50	100	20	4	77.65	76.88	67.73
100	50	30	5	81.32	79	67.73
100	100	30	5	81.32	80.72	67.73

Table 4 – Comparison of Fitness Scores under Different GA Settings

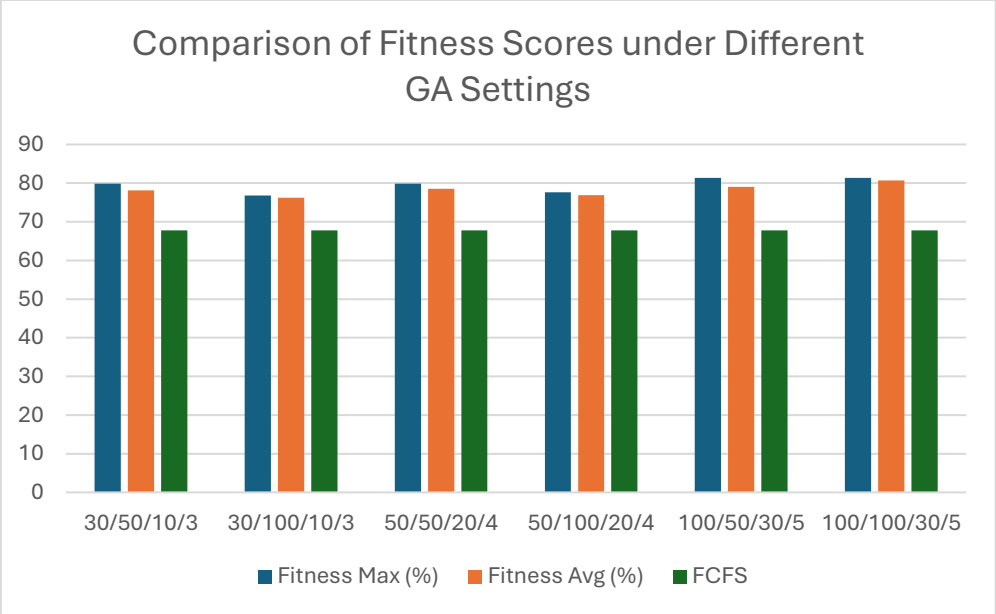


Figure 5 - Fitness Trends for Various Configurations

Table 4 and Figure 5 show that:

- A population of 100 and mutation rate of 30% achieved the highest max fitness (81.32%).
- More generations do not always improve results; best average fitness occurred around 90 generations.
- Moderate tournament size (3–5) provided a balance between selection pressure and diversity.
- The genetic algorithm consistently outperformed FCFS across all tested configurations.

4.3 Tournament Size Impact

Population Size	Generations	Mutation Rate (%)	Tournament Size	Fitness Max (%)	Fitness Avg (%)	FCFS
50	50	20	3	81.45	78.51	67.73
50	50	20	5	78.01	77.07	67.73
50	50	20	7	77.53	76.12	67.73
50	50	20	9	77.17	76.45	67.73

Table 5. Fitness Scores for Different Tournament Sizes

Table 5 analyze the effect of varying tournament size while keeping other parameters constant. A tournament size of 3 resulted in the highest max fitness (81.45%). Larger sizes slightly reduced performance, possibly due to early convergence.

4.4 Mutation Rate Impact

Population Size	Generations	Mutation Rate (%)	Tournament Size	Fitness Max (%)	Fitness Avg (%)	FCFS
50	50	5	3	80.14	78.24	67.73
50	50	10	3	79.63	78.18	67.73
50	50	15	3	79.51	77.7	67.73
50	50	20	3	79.38	76.94	67.73

Table 6 – Fitness Scores for Different Mutation Rates

Table 6 highlight the impact of mutation rates from 5% to 20%. The best fitness was achieved with a 5% rate. Excessive mutation disrupted convergence. A small rate improved exploration without destabilizing solution quality.

4.5 Generations Impact

Population Size	Generations	Mutation Rate (%)	Tournament Size	Fitness Max (%)	Fitness Avg (%)	FCFS
50	50	15	3	79.51	77.7	67.73
50	90	15	3	81.32	80.08	67.73
50	120	15	3	79	78.37	67.73
50	170	15	3	79.25	78.56	67.73

Table 7 – Fitness Evolution with Increasing Generations

As shown in Table 7 increasing generations from 50 to 90 yielded noticeable gains. Beyond that, improvement plateaued or declined slightly. This indicates diminishing returns in prolonged evolution.

4.6 Population Size Impact

Population Size	Generations	Mutation Rate (%)	Tournament Size	Fitness Max (%)	Fitness Avg (%)	FCFS
30	50	15	3	76.22	75.21	67.73
50	50	15	3	79.51	77.7	67.73
70	50	15	3	80.4	77.45	67.73
100	50	15	3	81.32	74.57	67.73

Table 8 – Fitness Evolution under Different Population Sizes

Table 8 summarize the impact of varying population size. A population of 70 offered the best balance, while 100 led to higher maximum fitness but lower average, indicating over-diversity.

In all experiments, GA results were significantly superior to those of FCFS, demonstrating the robustness and adaptability of evolutionary optimization for crisis-driven hospital resource allocation.

4.7 Observations and Insights

The simulation results highlight several important observations:

• The Genetic Algorithm consistently outperforms FCFS in terms of fitness, which reflects better allocation efficiency and treatment success.

• The algorithm adapts well to different resource constraints, even when the number of beds or doctors is limited.

• High-priority patients are more likely to be assigned when using GA compared to FCFS.

• The graphical interface and live fitness feedback allow quick interpretation of results, even for non-technical users.

These findings support the conclusion that evolutionary algorithms like GA can significantly improve hospital planning and decision-making, especially in scenarios with high demand and limited capacity.

5. Conclusions

This study proposed and implemented a simulation framework designed to support decision-making in hospital crisis situations through heuristic resource allocation. By combining a custom-built software system with a genetic algorithm-based optimization

strategy, the project successfully demonstrated how intelligent heuristics can enhance operational efficiency in constrained healthcare environments.

The developed Windows Forms application, written in C#, provides a modular interface for loading patient data, configuring simulations, visualizing outcomes, and comparing heuristic optimization against a baseline First-Come-First-Served (FCFS) strategy. The integration with a SQL Server database ensures realistic and reusable scenarios, making the tool suitable for both academic research and operational testing.

Experimental results confirmed that the Genetic Algorithm (GA) significantly outperforms the FCFS method in terms of treated patient priority, resource utilization, and overall system efficiency. The flexible design of the GA engine, including adjustable parameters such as population size, number of generations, mutation rate, and tournament size, allowed extensive exploration of the solution space. The analysis also highlighted how parameter tuning can influence convergence and fitness scores.

Overall, the system proved to be both functional and extensible, with clear potential for real-world impact in crisis management planning.

References

[1] M. Mitchell, "An Introduction to Genetic Algorithms," *MIT Press,* 1998.

[2] N. Jayanthi, S. Baskar and S. Kannan, "Patient Scheduling System for Medical Treatment Using Genetic Algorithm," *International Journal of Computer Applications,* vol. 37, no. 11, pp. 21-28, 2012.

[3] M. Kühn, M. Spies and D. Schwarzbach, "A Multi-objective Optimization Approach Based on Deterministic and Metaheuristic Techniques to Resource Management in Health Crisis Scenarios Under Uncertainty," *Expert Systems with Applications,* vol. 197, no. 116702, 2022.

[4] Y. Chen and Y. Lu, "Medical Staff Scheduling Based on Particle Swarm Optimization under Inter-Hospital Cooperation Mode," *International Journal of Environmental Research and Public Health,* vol. 19, no. 10, 2022.

[5] G. Girimurugan, G. Vinoth and A. Muthukumaran, "A novel framework of dynamic healthcare planning using genetic algorithm and deep learning techniques," *Journal of Ambient Intelligence and Humanized Computing,* vol. 12, p. 4235–4246, 2021.

[6] A. Heiniger, M. Reiss and S. Rachuba, "Workforce scheduling in hospital emergency departments using a genetic algorithm," *Health Care Management Science,* vol. 25, p. 132–147, 2022.

[7] F. Wan, J. Fondrevelle and T. e. a. Wang, "Two-stage multi-objective optimization for ICU bed allocation under multiple sources of uncertainty," *Sci Rep,* vol. 18925, no. 13, 2023.

[8] K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation,* vol. 6, no. 2, pp. 182-197, April 2002.

[9] C. Yan, N. McClure, S. P. Dukelow, B. Mann and J. Round, "Optimal Planning of Health Services through Genetic Algorithm and Discrete Event Simulation: A Proposed Model and Its Application to Stroke Rehabilitation Care," *MDM policy & practice,* vol. 7(2), no. 23814683221134098, 2022.

[10] B. Oliveira, J. Vasconcelos, J. Almeida and L. Pinto, "A Simulation-Optimisation approach for hospital beds allocation," *International Journal of Medical Informatics,* vol. 141, no. 104174, 2020.

[11] W. Feng, Z. Lou, N. Kong and H. Wan, "A multiobjective stochastic genetic algorithm for the pareto-optimal prioritization scheme design of real-time," *Operations Research for Health Care,* vol. 15, pp. 32-42, 2017.

[12] S. Farahi and K. Salimifard, "A simulation–optimization approach for measuring emergency department resilience in times of crisis,," *Operations Research for Health Care,* vol. 31, no. 100326, 2021.

[13] M. Nacht, "Operations Research," *International Encyclopedia of the Social & Behavioral Sciences,* pp. 10873-10876, 2001.

[14] G. Krishnan and S. Sowmya Kamath, "A novel GA-ELM model for patient-specific mortality prediction over large-scale lab event data," *Applied Soft Computing,* vol. 80, pp. 525-533, 2019.

[15] X. Pang, Y. Ge, K. Wang, A. Traina and H. Wang, "Patient assignement optimization in cloud healthcare systems: a distributed genetic algorithm," *Health Inf Sci Syst,* vol. 11, no. 30, 2023.

[16] W. Allihaibi, M. Cholette, M. Masoud, J. Burke and A. Karim, "A heuristic approach for scheduling patient treatment in an emergency department based on bed blocking," *International Journal of Industrial Engineering Computations,* vol. 11, pp. 1-20, 2020.

[17] J. S. Sohmer, S. Fridman, D. Peters and e. al, "Winning the Lottery: A Simulation Study Comparing Scarce Resource Allocation Protocols in Crisis Scenarios," *Cureus,* vol. 17, no. 1, 05 January 2025.

[18] T. Ogundare, "Surrogate assisted by genetic algorithm for simulation of a hospital-resource allocation: lockdown policy analysis for COVID-19 pandemic case," *MSc dissertation, University of Witwatersrand,* 2022.

[19] L. Luo, J. Li, X. Xu, W. Shen and L. Xiao, "A Data-Driven Hybrid Three-Stage Framework for Hospital Bed Allocation: A Case Study in a Large Tertiary Hospital in China," *Computational and mathematical methods in medicine,* no. 7370231, 2019.

[20] J. Jiang, D. Sheng, X. Chen, Q. Tian, F. Li and P. Yang, "Data-driven collaborative healthcare resource allocation in pandemics," *Transportation Research Part E: Logistics and Transportation Review,* vol. 192, no. 103828, 2024.

[21] S. Ridler, A. Mason and A. Raith, "A simulation and optimisation package for emergency medical services," *European Journal of Operational Research,* vol. 298, no. 3, pp. 1101-1113, 2022.

DOI: 10.2478/ijasitels-2025-0008

Achieving Clinical Reliability in Suicide Risk Detection: A Low-Resource Benchmark of RoBERTa vs. DistilBERT

Stefania-Eliza Berghia[1], *Adrian Barglazan*[1]
[1]*Computer Science and Electrical and Electronics Engineering Department, Faculty of Engineering, "Lucian Blaga" University of Sibiu, Romania*
{stefania.berghia, adrian.barglazan} @ulbsibiu.ro

Abstract

Detecting suicidal ideation in social media text is a critical public health objective, demanding high-accuracy, deployable models. This study addresses the challenge of achieving clinical reliability within severe hardware constraints. We conduct a comparative fine-tuning benchmark of two Transformer models, DistilBERT and RoBERTa, for binary classification of suicidal risk. The models were optimized on a balanced, 10,000-sample subset of the Reddit Suicide Detection Dataset under CPU-only, low-resource constraints. Prediction robustness is achieved through a simple selection process that chooses the output with the highest confidence score from the two models. RoBERTa achieved a peak F1-Score of 97.4% and 97.40% accuracy, substantially outperforming DistilBERT (94.1% F1-Score). Crucially for safety-critical applications, error analysis confirmed RoBERTa's ethical superiority by achieving a 40% reduction in False Negatives compared to DistilBERT. The validated framework establishes that high-performance, ethically robust risk detection is feasible under resource limitations, enabling the safe integration of the final classification system with a constrained LLaMA 3 conversational module for proactive support.

Keywords: suicidal ideation detection, transformer models, clinical NLP

1 Introduction

Suicide represents a major global public health concern, with early identification of suicidal ideation remaining a persistent challenge. Many individuals who experience severe psychological distress do not actively seek clinical support, yet they often express early warning signs in digital environments. Social media platforms such as Reddit provide an open, large-scale stream of unsolicited textual data, which has contributed to the development of digital phenotyping approaches for understanding mental states from online behaviour. Language expressed in communities like r/SuicideWatch offers valuable indicators of emotional crises, while general subreddits provide contrasting baseline content.

These characteristics create an opportunity to develop non-intrusive, scalable, language-based systems capable of detecting risk signals in real time. However, effective deployment in practice requires models that are not only accurate but also computationally accessible and clinically reliable, particularly in low-resource settings where specialized hardware is not available.

This study addresses these challenges by providing two key contributions. First, it presents a rigorous low-resource benchmark comparing two prominent Transformer-based language models, Distilled Bidirectional Encoder Representations from Transformers (DistilBERT) and Robustly Optimized BERT Pretraining Approach (RoBERTa), for suicidal ideation detection using a balanced 10,000-sample subset of the Suicide Detection Dataset. Second, it examines model errors with particular attention to false negatives, the most consequential error type in suicide-risk detection, as these represent missed high-risk cases. Finally, the work demonstrates a practical application through a web-based detect-and-support pipeline that integrates the hybrid model with an empathetic conversational agent based on the Large Language Model Meta AI, version 3, 8-billion parameters (LLaMA3-8B).

2 Related work

Transformer-based architectures form the foundation of modern text classification. [1] introduced the Bidirectional Encoder Representations from Transformers (BERT) model, which leverages masked language modeling and bidirectional self-attention to capture contextual dependencies. Building on this foundation, [2] proposed DistilBERT, a compressed version that maintains most of BERT's performance while reducing inference cost. RoBERTa was introduced in [3] and improves BERT by applying dynamic masking, removing the Next Sentence Prediction objective, and scaling training with larger batch sizes and corpora.

Transformers have also enabled progress in multimodal mental health detection. [4] presented the Multimodal Hierarchical Attention (MHA) model, which integrates Convolutional Neural Network (CNN)-based textual features, ResNet-18 visual representations, temporal metadata, and emotion lexicon signals using intra- and inter-modality attention mechanisms. Evaluated on a clinically validated Weibo dataset of 10,000 users, the model achieved 89.7% accuracy, outperforming Term Frequency–Inverse Document Frequency (TF-IDF) and Support Vector Machine (SVM), CNN, Bidirectional Long Short-Term Memory (BiLSTM), Bidirectional Encoder Representations from Transformers (BERT), and RoBERTa.

A related multimodal approach was introduced in [5] through the Emotion-based Reinforcement Attention Network (ERAN). ERAN uses a TextCNN emotion extractor, a BiLSTM encoder, and a reinforcement-learning agent that selects emotionally salient posts. Evaluated on the MDD dataset, it obtained 90.6% accuracy, outperforming Naive Bayes, SVM, Long Short-Term Memory (LSTM), BiLSTM-Attention, BERT-base, and RoBERTa-base. Ablation experiments confirmed the importance of both the emotion extractor and the reinforcement-learning–based post-selection mechanism.

Complementing these models, [6] introduced the Multimodal Hierarchical Attention Graph Convolutional Network with Vision Transformer (MHA-GCN_ViT), which integrates Graph Convolutional Networks (GCN) for modeling EEG inter-channel dependencies with a Vision Transformer (ViT) for spectral audio features. On the MODMA dataset, MHA-GCN_ViT achieved 89.03% accuracy, outperforming SVM, CNN, and earlier multimodal frameworks.

Research specifically addressing suicidal ideation detection is more limited. [7] compared multiple Transformer-based models on annotated Reddit posts, reporting that RoBERTa achieved 90.5% accuracy and an 89.3% F1-score, outperforming LSTM-based architectures through improved long-range contextual modeling. A

complementary approach is presented in [8], which combines Linguistic Inquiry and Word Count (LIWC) psycholinguistic features with word embeddings and Extreme Gradient Boosting (XGBoost). While effective, feature-engineered pipelines require extensive manual preprocessing relative to Transformer-based models, which learn contextual features directly from text.

3 Proposed methodology

The proposed system for suicidal ideation detection is built around a supervised text-classification framework that combines modern Transformer architectures with a lightweight hybrid decision mechanism. The methodological pipeline comprises four stages: dataset preparation, text preprocessing, model fine-tuning, and the integration of a confidence-based hybrid inference strategy designed to enhance prediction robustness in low-resource environments.

3.1 Dataset description

This study utilizes the Suicide Detection Dataset [9], a publicly available corpus on the Kaggle platform. The dataset was curated by crawling posts from the Reddit social media platform. Its composition is binary, consisting of posts from two distinct subreddits:

- r/SuicideWatch: A community providing support for individuals in suicidal crises. Posts from this subreddit are labeled as 'suicide'.
- r/teenagers: A general-interest subreddit for teenagers to discuss daily life. Posts from this subreddit are labeled as 'non-suicide'.

The original dataset contains 232,074 posts, equally balanced between the two classes.

3.2 Corpus curation for low-resource environments

Training large Transformer models on a dataset of over 230,000 samples requires significant computational resources, specifically high-VRAM GPUs. A primary objective of this study was to develop a high-performance model under significant hardware constraints (a laptop with an Intel Core i7-1065G7 CPU, 8 GB RAM, and no dedicated GPU).

To this end, a representative subset of 10,000 samples was randomly selected from the full dataset. This approach serves a dual purpose: first, it makes the experiment feasible on the available hardware, and second, it provides a valuable, reproducible benchmark for other researchers working in similarly low-resource environments. The use of random_state=42 during sampling ensures that this exact subset can be replicated for future studies.

After selection and cleaning, the final subset comprised 9,996 posts. The random sampling process preserved the near-perfect class balance of the original corpus, as detailed in Table 1.

Table 1. Class distribution within the selected dataset

Class	Samples	Percentage	Source
Suicidal	5021	50.25%	Subreddit SuicideWatch

Non-suicidal	4975	49.75%	Subreddit teenagers
Total	9996	100%	Reddit via Pushshift API

3.3 Text preprocessing

Prior to fine-tuning the models, each Reddit post underwent a standardized preprocessing sequence designed to reduce noise while preserving semantically relevant content. Text normalization involved lowercasing all characters, removing hyperlinks, email addresses, and platform-specific tokens such as user mentions or hashtags. Superfluous whitespace and repeated characters were trimmed, and non-essential symbols were filtered to produce cleaner textual input.
After normalization, posts were converted into model-ready representations using the tokenizers associated with each model. DistilBERT employs the WordPiece tokenizer inherited from BERT, whereas RoBERTa uses a Byte-Pair Encoding (BPE) tokenizer. Both tokenizers generated input_ids and attention_mask tensors, with sequences truncated or padded to a maximum length of 32 tokens, an empirically determined value based on the dataset's typical post lengths.

3.4 Transformer architectures

The models used in this study are based on the Transformer encoder architecture introduced by Vaswani et al. [4], which replaces recurrence and convolution with a multi-head self-attention mechanism capable of modeling long-range dependencies in parallel. This design is particularly suitable for analyzing user-generated content on social media, where emotional cues may appear at any point in the text and where conventional grammatical patterns are frequently absent. Because the suicidal ideation detection task is a sequence-classification problem, only the encoder component of the Transformer is required.
The Transformer encoder is composed of stacked layers, each containing a multi-head self-attention sublayer followed by a position-wise feed-forward network. These layers are surrounded by residual connections and layer normalization. During fine-tuning for classification, a special prepended token—[CLS] for BERT-based models and <s> for RoBERTa—is used to summarize the meaning of the entire input sequence. The final hidden state corresponding to this token serves as the input to the classification head.

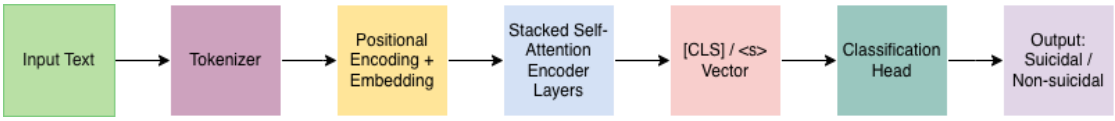


Figure 1. Transformer fine-tuning workflow

3.4.1 Model A: Fine-Tuned DistilBERT

The first model evaluated in this study is DistilBERT [2], a compressed version of BERT that reduces the depth of the encoder to six layers and totals approximately 66 million parameters. This compact architecture is well suited for CPU-only environments, offering substantially faster inference while retaining most of BERT's semantic modeling capabilities.
To adapt DistilBERT for binary classification, a dropout layer and a linear transformation nn.Linear(768, 2) were added on top of the [CLS] token representation.

The model was fine-tuned using the AdamW optimizer with a learning rate (LR) of 5e-5 for two epochs, following a linear warm-up schedule. A batch size of eight samples was selected as the largest configuration compatible with CPU memory constraints.
This training configuration ensures that the evaluation reflects DistilBERT's representational capacity under realistic low-resource conditions.

3.4.2 Model B: Fine-Tuned RoBERTa

The second model is the RoBERTa [3], a Transformer encoder with twelve layers and approximately 125 million parameters. RoBERTa differs from BERT-style models through several pretraining refinements, such as dynamic masking, removal of the Next Sentence Prediction objective, and the use of larger corpora, which together yield more robust contextual embeddings.
RoBERTa was fine-tuned using the same classification head as DistilBERT—an nn.Dropout layer followed by nn.Linear(768, 2)—to ensure that any performance differences arise strictly from the encoder representations rather than classifier complexity. The model was trained using the AdamW optimizer with a LR of 2e-5 for four epochs, also employing a linear warm-up scheduler.
This lower LR reflects RoBERTa's greater parameter count and results in more stable convergence during CPU-based fine-tuning.

3.4.3 Inference-time prediction selection

Although both DistilBERT and RoBERTa achieve strong results, their strengths differ: DistilBERT offers faster, more efficient inference, while RoBERTa provides richer contextual representations. To take advantage of both in the deployed system, a lightweight inference-time strategy is employed, designed to choose the prediction from the model that is most certain about its own result.
This selection strategy does not rely on traditional ensemble methods such as voting, averaging, or training an additional meta-classifier. Instead, both models process the input text in parallel, and each produces a probability distribution over the two classes. The mechanism identifies the highest probability value (i.e., the model's internal confidence) from each distribution. The final label is assigned by simply selecting the prediction from the model with the larger confidence score.
This approach adds only a single comparison after the standard forward passes of the two models, introducing virtually no computational overhead. Despite its simplicity, this selection method provides a practical robustness advantage by dynamically choosing between a faster model and a deeper, more expressive one on a case-by-case basis. This allows the system to flexibly exploit DistilBERT's efficiency and RoBERTa's representational power without additional training or architectural modifications.

4 Experiments and results

The models were trained on the 72% training set, with hyperparameters tuned against the 8% validation set. The final, optimized models were then evaluated a single time on the 20% held-out test set. This allocation (20% of the total dataset) was purposefully set high to ensure a more robust and statistically reliable evaluation of critical metrics—particularly False Negatives (FN)—for this high-stakes task.

4.1 Hyperparameter Optimization

The fine-tuning process involved a systematic sensitivity analysis of the LR and the number of training epochs to identify the optimal configuration for maximum generalization and stability for both DistilBERT and RoBERTa models.

4.1.1 DistilBERT Optimization

The optimization study for DistilBERT, an efficient 6-layer architecture, focused on finding the optimal configuration to maximize performance while preventing premature overfitting. Sensitivity analysis, visualized in Fig. 2, confirmed that performance peaked at a higher LR of 5e-5, with metrics declining for lower rates (e.g., 2e-5). Analysis of epoch count, shown in Fig. 3, revealed that the model achieved maximum generalization after only 2 epochs of training, with subsequent epochs leading to performance degradation attributable to overfitting. The final configuration adopted for the model used a LR of 5e-5, trained for two epochs with a batch size of 8 and a maximum sequence length of 32 tokens.

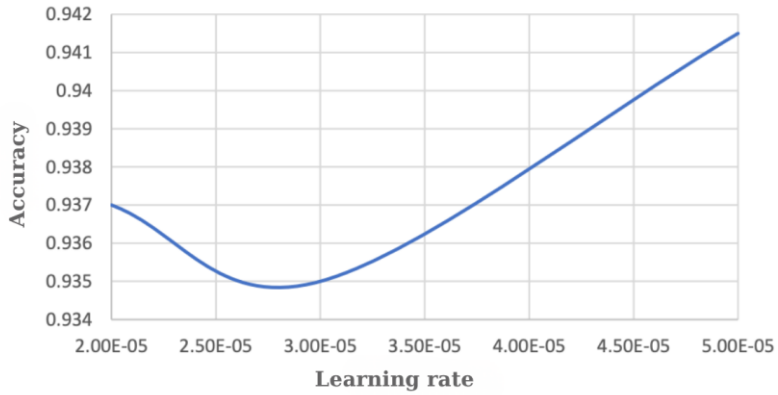


Figure 2. Influence of LR on model accuracy

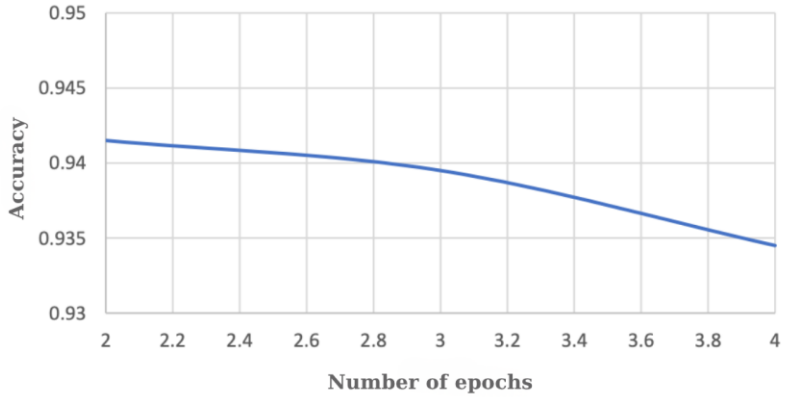


Figure 3. Influence of number of epochs on model accuracy

4.1.2 RoBERTa Optimization

Given RoBERTa's deeper, 12-layer architecture and superior representational power, a slower fine-tuning strategy was required. The hyperparameter analysis demonstrated an inverse relationship between LR and performance: metrics maximized at the lowest tested rate, 2e-5, confirming the necessity of slower, meticulous parameter updates to ensure stable convergence as shown in Fig. 4. Unlike DistilBERT, RoBERTa benefited

from extended training, showing continuous metric improvement and reaching peak performance after 4 epochs (Fig. 5). This confirmed that the deeper model required more steps to capture nuanced semantic patterns without exhibiting premature overfitting. The final configuration adopted for the model used a LR of 2e-5, trained for four epochs with a batch size of 8 and a maximum sequence length of 32 tokens.

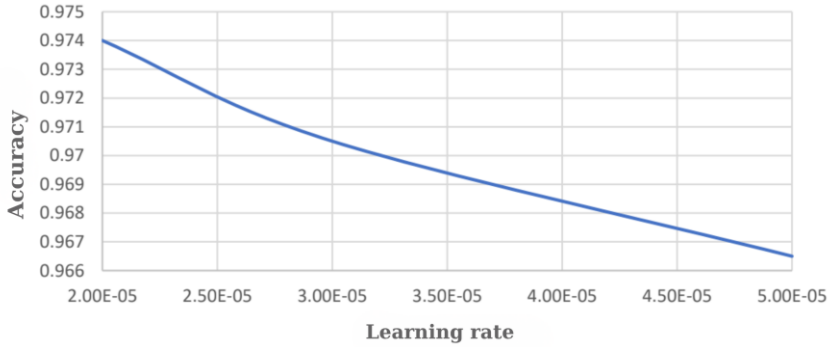


Figure 4. Influence of LR on model accuracy

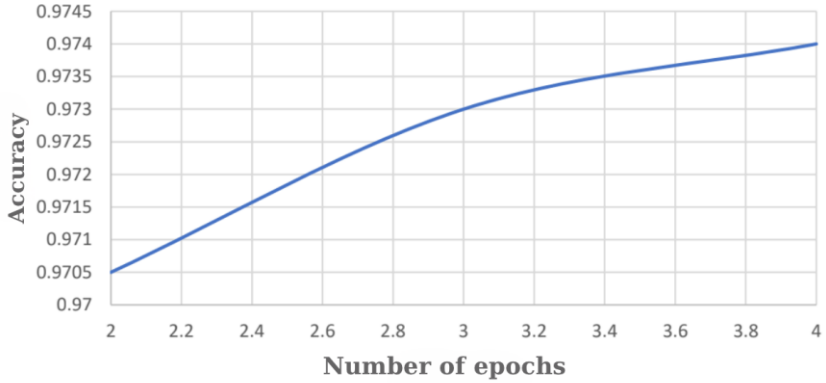


Figure 5. Influence of number of epochs on model accuracy

4.2 Performance evaluation

The two optimally configured models were evaluated on the N=2,000 test set, confirming their complementary roles and establishing the basis for the hybrid fusion mechanism.

A head-to-head comparison confirmed RoBERTa's substantial superiority in both generalization and critical error reduction, essential for this high-stakes task. The performance metrics are summarized in Table 2.

Table 2. Comparative performance DistilBERT vs RoBERTa

Model	Accuracy	Precision	Recall	F1 Score
DistilBERT	94.15%	94.30%	95.20%	94.10%
RoBERTa	97.40%	97.70%	97.10%	97.40%

RoBERTa consistently outperformed DistilBERT across all generalized metrics, achieving a peak accuracy of 97.40% and an F1-Score of 97.4% (3.3 percentage points higher than DistilBERT's 94.1% F1-score). The RoBERTa model demonstrated superior recall (97.1%) and precision (97.7%), indicating exceptional capability in both detecting actual risk messages and minimizing false warnings. This performance

validates RoBERTa's efficacy and ethical superiority, establishing it as the high-accuracy anchor of the hybrid system. This exceptional performance is largely attributed to the meticulous data curation and selection strategy on the perfectly balanced subset, which allowed the model to learn highly specific, decisive linguistic features efficiently.

4.3 Error analysis

The design philosophy of the implemented hybrid system prioritizes the minimization of FN above all other metrics, aligning with the highest ethical mandate for suicide risk detection systems. An FN represents a failure to flag a high-risk message, potentially leading to catastrophic consequences. Conversely, a False Positive (FP) is tolerable, resulting only in the unnecessary offer of supportive resources (such as the LLaMA 3 chatbot access).
The DistilBERT model achieved balanced True Positives (TP) and True Negatives (TN), but generated 48 critical FNs—missed opportunities for intervention—which represented the primary vulnerability of deploying DistilBERT as a standalone system. The 69 FPs were considered acceptable, reflecting a conservative ethical approach that prioritizes safety over nuisance warnings. The full breakdown of its classification errors on the test set is presented in Table 3.

Table 3. DistilBERT confusion matrix

	Predicted Suicide	Predicted Non-Suicide
Actual Suicide	TP = 957	FN = 48
Actual Non-Suicide	FP = 69	TN = 926

The RoBERTa model demonstrated high reliability by substantially minimizing the critical error rate. As detailed in Table 4, it reduced the number of FN to only 29 and the FP to 23. This means RoBERTa achieved a 40% reduction in the critical FN rate compared to DistilBERT (48 FN vs. 29 FN), validating its ethical superiority in this high-stakes task. The low FP rate further highlights RoBERTa's precision in distinguishing genuine risk from non-suicidal emotional expression.

Table 4. RoBERTa confusion matrix

	Predicted Suicide	Predicted Non-Suicide
Actual Suicide	TP = 976	FN = 29
Actual Non-Suicide	FP = 23	TN = 972

The prediction selection mechanism serves as a strategic deployment of the system's conservative ethical bias. By dynamically adopting the prediction with the highest certainty, the system ensures that RoBERTa's superior sensitivity—demonstrated by its significantly lower FN rate—drives the final decision in ambiguous or high-stake cases. This reliable and ethically conservative classification enables the safe and responsible activation of the constrained LLaMA 3 support module. This ensures advanced conversational aid is only triggered when a robust, high-certainty risk signal is present.

5 Conclusions and further work

The experimental results definitively validate the efficacy of the comparative methodology for suicide risk detection from social media text. The RoBERTa architecture demonstrated substantial superiority, achieving an F1-score of 97.4% and, most significantly, a 40% reduction in the critical FN rate compared to the computationally lighter DistilBERT model. This FN reduction confirms the ethical requirement for prioritizing deep, robust architectures over sheer computational speed in high-stakes mental health applications.

A core contribution of this work is the validation of a resource-efficient fine-tuning methodology that achieved state-of-the-art-level performance on a constrained dataset subset, utilizing only CPU resources, demonstrating feasibility for low-resource environments.

Finally, the resultant hybrid system, integrated with an ethically constrained, prompt-guided LLaMA 3 conversational module via the Groq API, establishes a viable, safe, and responsible prototype for an early intervention, detect-and-support pipeline in digital mental health.

To transition the prototype into a fully robust and clinically relevant tool, several key development directions are necessary.

The most immediate required improvement is the scaling of the training process. The next phase of research must utilize advanced GPU computational environments to fine-tune the optimal RoBERTa model configuration on the full original dataset of 232,074 examples to validate robustness across a wider and noisier linguistic distribution.

Further research should focus on multimodal and linguistic expansion. This includes integrating non-textual features such as temporal patterns, posting frequency, and interaction metadata to improve predictive accuracy, as shown in related research, as well as expanding the data ingestion pipeline to handle inputs from multiple social media platforms and developing a variant capable of processing texts in Romanian.

For practical implementation, clinical validation and real-time intervention capabilities are essential. Future work must involve formal collaborations with mental health professionals to perform clinical evaluation, establish calibrated risk thresholds based on standard diagnostic criteria, and integrate professional feedback for continuous improvement. Developing a secure native mobile application component capable of real-time monitoring and implementing automated, location-aware emergency alerts (such as integration with national emergency services like 112) would transform the system into a complete tool for proactive suicide prevention.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, *"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,"* arXiv preprint arXiv:1810.04805, 2018.

[2] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *"DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,"* arXiv preprint arXiv:1910.01108, 2019.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, *et al.*, *"RoBERTa: A Robustly Optimized BERT Pretraining Approach,"* arXiv preprint arXiv:1907.11692, 2019.

[4] J. Li, Z. An, W. Cheng, J. Zhou, F. Zheng, and B. Hu, *"MHA: a multimodal hierarchical attention model for depression detection in social media," Health Information Science and Systems*, vol. 11, no. 1, pp. 1–13, 2023.

[5] B. Cui, J. Wang, H. Lin, Y. Zhang, L. Yang, and B. Xu, *"Emotion-Based Reinforcement Attention Network for Depression Detection on Social Media: Algorithm Development and Validation," JMIR Medical Informatics*, vol. 10, no. 8, e37818, 2022.

[6] X. Jia, J. Chen, K. Liu, Q. Wang, and J. He, *"Multimodal depression detection based on an attention graph convolution and transformer," Mathematical Biosciences and Engineering*, vol. 22, no. 3, pp. 652–676, 2025.
[7] K. Hasan and J. Saquer, "A comparative analysis of transformer and LSTM models for detecting suicidal ideation on Reddit," arXiv preprint arXiv:2411.15404, 2024.
[8] A. Izmaylov, S. Malmasi, and A. Yates, *"Combining psychological theory with language models for suicide risk detection,"* in *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2405–2414, 2023.
[9] K. Nikhileswar, D. Vishal, L. Sphoorthi, and S. Fathimabi, "*Suicide Ideation Detection in Social Media Forums*," in *Proceedings - 2nd International Conference on Smart Electronics and Communication*, ICOSEC 2021, doi: 10.1109/ICOSEC51865.2021.9591887, 2021.

DOI: 10.2478/ijasitels-2025-0005

Sensorless Vector Control of a Surface Permanent Magnet Synchronous Motor

Gabriela Crăciunaş[1], Alina Cristina Viorel[1], Iulian Bouleanu[1]

[1]Computer Science and Electrical Engineering Department, "Lucian Blaga" University of Sibiu, Romania, E-Mail: gabriela.craciunas@ulbsibiu.ro; alina.viorel@ulbsibiu.ro; iulian.bouleanu@ulbsibiu.ro

Abstract

This paper analyzes the operation of the surface permanent magnet synchronous motor (SPMSM) in a sensorless configuration, with the main objective of speed estimation. It is presented a vector control strategy independent of position sensors, implemented through a voltage inverter with prescribed currents. Advanced algorithms, including the Gopinath observer, are used to accurately determine the rotor speed and position. This plays an essential role in increasing the stability and accuracy of the control system, by reducing estimation errors and ensuring a rapid response to load variations. Performance validation was performed through simulations in the MATLAB/Simulink environment, with the results demonstrating both the accuracy and reliability of sensorless control under various operating conditions. The conclusions highlight the efficiency of the proposed method and its relevance for the development of modern electric drive technologies.

Keywords: surface permanent magnet synchronous motor, vector control, Gopinath observer.

1 Introduction

Surface permanent magnet synchronous motor (SPMSM) has become benchmark solutions in modern applications due to their low energy consumption, advantageous power-to-volume ratio, and ability to respond quickly to load variations. They are frequently found in traction systems for electric vehicles, in automated production lines with industrial robots, but also in other drives that require high performance. For the proper operation of these machines, very precise control of the rotor position and speed is required [1]–[4]. In the conventional approach, this involves integrating encoder or resolver sensors, but these increase the price of the equipment, complicate assembly and can become vulnerable points in demanding industrial conditions.

An alternative to mechanical sensors is estimation techniques based on measured electrical signals. Multiple methods are described in the specialized literature that uses both mathematical models of the engine and numerical processing algorithms. For example, by applying the equations of state and using stator currents and voltages, information about the rotor position and speed can be deduced. Another method relies on magnetic flux analysis to determine the stator space vector, also based on stator voltages and currents [2], [5]–[8].

In this paper, an SPMSM is studied in a sensorless control regime, with the emphasis on determining the rotor speed using the reduced order adaptive observer developed by Gopinath. The theoretical foundations of the motor and the vector control strategies applicable to sensorless control are described. A control scheme without position sensors is also presented, the motor being powered by a voltage inverter with prescribed currents. [5], [8]–[11]. The performance of the method was evaluated through simulations performed in MATLAB/Simulink, the results confirming the accuracy and stability of the solution.

2 Mathematical model

2.1 Permanent magnet synchronous motor model

To deduce the structure of the Gopinath reduced order adaptive observer, we start from the SPMSM engine model represented in a d-q axis rotor mobile system. Thus, the stator voltage equations are,

$$u_d = R_s i_d - p\omega_r L_s i_q + L_s \frac{di_d}{dt} \tag{1}$$

$$u_q = R_s i_q + p\omega_r L_s i_d + p\omega_r \phi_0 + L_s \frac{di_q}{dt} \tag{2}$$

Where,

$$i_s = i_d + j i_d, u_s = u_d + j u_d \tag{3}$$

The calculation relationship for the electromagnetic torque of the surface permanent magnet synchronous machine is

$$T = \frac{3}{2} p \phi_0 i_q \tag{4}$$

To determine the speed of the car, the dynamic equation of motion is integrated,

$$J \frac{d\omega_r}{dt} = T - T_l - F\omega_r \tag{5}$$

Thus, the canonical form of the system of equations of the surface permanent magnet synchronous machine model results,

$$\begin{gathered} \frac{di_d}{dt} = -\frac{R_s}{L_s} i_d + p\omega_r i_q + \frac{1}{L_s} u_d \\ \frac{di_q}{dt} = -\frac{R_s}{L_s} i_q - -p\omega_r \frac{\phi_0}{L_s} - p\omega_r i_d + \frac{1}{L_s} u_q \\ \frac{d}{dt}\omega_r = \frac{3}{2J} p \phi_0 i_q - \frac{F}{J}\omega_r - \frac{T_l}{J} \\ \frac{d}{dt}\theta_r = \omega_r \end{gathered} \tag{6}$$

Where,
R_s - *the stator resistance*
L_s - *the stator inductance*
ϕ_0 - *the permanent magnet flux linkage*
p - *the pole pair number*

ω_r - *the rotor speed*
θ_r - *the angular position*
F - *the friction coefficient*
T_l - *the load torque*
J - *the moment inertia*

Studying equations (6), it is observed that the multiplication of two state variables $(\omega_r i_d)$ and $(\omega_r i_q)$ is performed, which results in a non-linear system. This shortcoming must be corrected in order to implement the Gopinath reduced order adaptive observer, thus transforming the system into a linear one. Thus, the system of equations (6) becomes linear if the order of the state variables is changed, resulting in,

$$\frac{d}{dt}\begin{bmatrix}\omega_r\\ \theta_r\\ i_d\\ i_q\end{bmatrix} = \begin{bmatrix} -\frac{F}{J} & 0 & 0 & \frac{3}{2J}p\phi_0 \\ 1 & 0 & 0 & 0 \\ pi_q & 0 & -\frac{R_s}{L_s} & 0 \\ -\frac{p}{L_s}(\phi_0 + L_s i_d) & 0 & 1 & -\frac{R_s}{L_s} \end{bmatrix} t \cdot \begin{bmatrix}\omega_r\\ \theta_r\\ i_d\\ i_q\end{bmatrix} + \begin{bmatrix} 0 & 0 & \frac{1}{J} \\ 0 & 0 & 0 \\ \frac{1}{L_s} & 0 & 0 \\ 0 & \frac{1}{L_s} & 0 \end{bmatrix} \cdot \begin{bmatrix}u_d\\ u_q\\ T_l\end{bmatrix} \quad (7)$$

Where,

$$A = \begin{bmatrix} -\frac{F}{J} & 0 & 0 & \frac{3}{2J}p\phi_0 \\ 1 & 0 & 0 & 0 \\ pi_q & 0 & -\frac{R_s}{L_s} & 0 \\ -\frac{p}{L_s}(\phi_0 + L_s i_d) & 0 & 1 & -\frac{R_s}{L_s} \end{bmatrix} t \quad (8)$$

$$B = \begin{bmatrix} 0 & 0 & \frac{1}{J} \\ 0 & 0 & 0 \\ \frac{1}{L_s} & 0 & 0 \\ 0 & \frac{1}{L_s} & 0 \end{bmatrix} \quad (9)$$

2.2 Gopinath observer model

The Gopinath observer is designed to estimate electrical and mechanical quantities in systems where sensors have been eliminated. In this paper, the motor speed and the rotor position relative to the stator will be estimated. These estimates are possible if we intervene on the state vectors,

$$x = \begin{bmatrix}\omega_r\\ \theta_r\\ i_d\\ i_q\end{bmatrix} = \begin{bmatrix}x_e\\ y\end{bmatrix} \quad (10)$$

Where,

$$x_e = \begin{bmatrix}\omega_r\\ \theta_r\end{bmatrix}, y = \begin{bmatrix}i_d\\ i_q\end{bmatrix} \quad (11)$$

With these adjustments, the matrix equation (7) becomes,

$$\frac{d}{dt}\begin{bmatrix} x_e \\ y \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \cdot \begin{bmatrix} x_e \\ y \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \cdot \begin{bmatrix} u_d \\ u_q \\ T_l \end{bmatrix} \tag{12}$$

Where,

$$a_{11} = \begin{bmatrix} -\frac{F}{J} & 0 \\ 1 & 0 \end{bmatrix}, a_{21} = \begin{bmatrix} pi_q & 0 \\ -\frac{p}{L_s}(\phi_0 + L_s i_d) & 0 \end{bmatrix} \tag{13}$$

$$a_{12} = \begin{bmatrix} 0 & \frac{3}{2J}p\phi_0 \\ 0 & 0 \end{bmatrix}, a_{22} = \begin{bmatrix} -\frac{R_s}{L_s} & 0 \\ 0 & -\frac{R_s}{L_s} \end{bmatrix} \tag{14}$$

$$b_1 = \begin{bmatrix} 0 & 0 & \frac{1}{J} \\ 0 & 0 & 0 \end{bmatrix}, b_2 = \begin{bmatrix} \frac{1}{L_s} & 0 & 0 \\ 0 & \frac{1}{L_s} & 0 \end{bmatrix} \tag{15}$$

In the specialized literature [12],[13] the expression of the Gopinath observer is,

$$\frac{d}{dt}\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = F \cdot \begin{bmatrix} \omega_r \\ \theta_r \end{bmatrix} + G \cdot \begin{bmatrix} u_d \\ u_q \\ T_l \end{bmatrix} + H \cdot \begin{bmatrix} i_d \\ i_q \end{bmatrix} \tag{16}$$

Where, $z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}$ (17), the vector of state variables of the Gopinath observer.

$$\begin{bmatrix} \omega_r \\ \theta_r \end{bmatrix} = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + L \cdot \begin{bmatrix} i_d \\ i_q \end{bmatrix} \tag{18}$$

$L = \begin{bmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{bmatrix}$ (19), matrix of unknown parameters. This matrix determines the stability of the observer and it is obtained by the pole placement method.

$$F = a_{11} - L \cdot a_{21} \tag{20}$$

$$G = b_1 - L \cdot b_2 \tag{21}$$

$$H = a_{12} - L \cdot a_{22} + F \cdot L \tag{22}$$

The substitutions of expressions (13) and (19) will be made in relation (20) and the matrix expression F will be obtained,

$$F = \begin{bmatrix} -\frac{F}{J} - l_{11}pi_q + l_{12}\frac{p}{L_s}(\phi_0 + L_s i_d) & 0 \\ 1 - l_{21}pi_q + l_{22}\frac{p}{L_s}(\phi_0 + L_s i_d) & 0 \end{bmatrix} \tag{23}$$

The polynomial that determines the poles assigned to the observer is,

$$P_0(s) = \det(s \cdot I_2 - F) = s\left[s + \frac{F}{J} + l_{11}pi_q - l_{12}\frac{p}{L_s}(\phi_0 + L_s i_d)\right] \tag{24}$$

By solving the above equation, a null solution is determined. It is further assumed that there is a solution $p_1 \neq 0$ and it is considered that the constants $l_{11} = l_{21} = l_{22} = 0$ do not influence the dynamics of the process. Thus, with these working hypotheses, the matrices *G* and *H* are determined,

$$G = \begin{bmatrix} 0 & -\frac{l_{12}}{J} & -\frac{1}{J} \\ 0 & 0 & 0 \end{bmatrix} \tag{25}$$

$$H = \begin{bmatrix} 0 & \frac{3}{2J}p\phi_0 + l_{12}\left(\frac{R_s}{L_s} - \frac{F}{J} + l_{12}\frac{p}{L_s}(\phi_0 + L_s i_d)\right) \\ 0 & l_{12} \end{bmatrix} \tag{26}$$

After substitutions in equation (16), the detailed expression of the Gopinath observer is obtained,

$$\frac{dz_1}{dt} = \left[-\frac{F}{J} + l_{12}\frac{p}{L_s}(\phi_0 + L_s i_d)\right] z_1 - \frac{l_{12}}{L_s}u_q - \frac{T_l}{J} + \\ + \left(\frac{3}{2J}p\phi_0 + l_{12}\frac{R_s}{L_s}\right) i_q + l_{12}\left[-\frac{F}{J} + l_{12}\frac{p}{L_s}(\phi_0 + L_s i_d)\right] i_q \tag{27}$$

$$\frac{dz_2}{dt} = z_1 + l_{12} \cdot i_q \tag{28}$$

If equation (27) is solved and relation (18) is taken into account, the two estimated electrical quantities are determined as follows: speed and angular position of the rotor, $(\omega_{r_{estim}}, \theta_{r_{estim}})$,

$$\omega_{r_{estim}} = z_1 + l_{12} \cdot i_q \tag{29}$$

$$\theta_{r_{estim}} = z_2 \tag{30}$$

Where,

$$l_{12} = \left(p_1 + \frac{F}{J}\right) \cdot \frac{L_s}{p(\phi_0 + L_s i_d)} \tag{31}$$

3 Motor control with current controlled voltage inverter

The modelling and simulation of the vector control of the SPMSM powered by a current controlled VSI voltage inverter in a sensorless system is presented in the block diagram in the figure Fig.1.

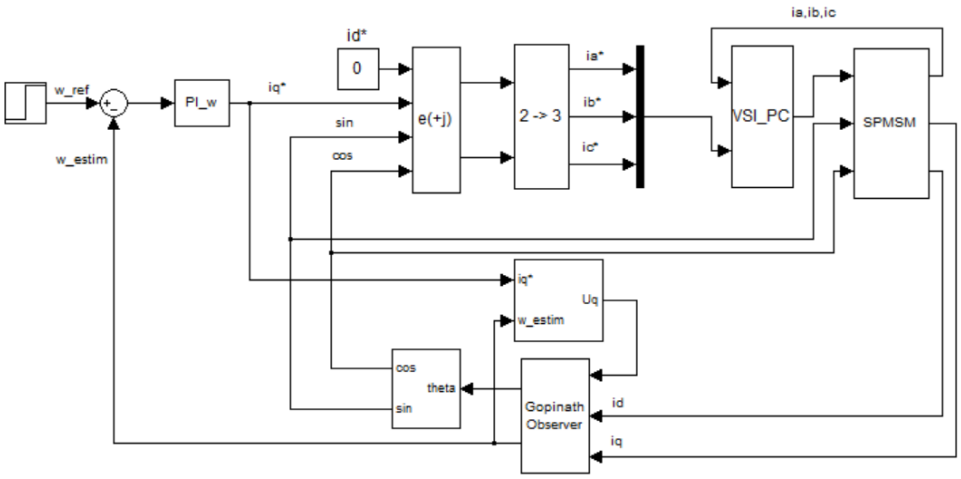


Figure 1. Control block diagram of SPMSM in sensorless system

It is considered as a regulation strategy for the synchronous motor with magnets, imposing the maintenance of zero reactive current ($i_d^* = 0$) and the active component (i_q^*) is the one that controls the electromagnetic torque developed by the motor. This current component is the result of the speed controller „PI-w". Both current components are considered to be in the rotating system joint with the rotor. Transforming the two components (i_d^*, i_q^*) into the fixed three-phase system, ($e^{+j}, 2 \rightarrow 3$) results in the prescribed instantaneous values of the phase currents (i_a^*, i_b^*, i_c^*).

$$\frac{di_d}{dt} = \frac{u_d}{L_s} - \frac{R_s}{L_s} i_d + p\omega_r i_q \tag{32}$$

$$\frac{di_q}{dt} = \frac{u_q}{L_s} - \frac{R_s}{L_s} i_q - p\omega_r i_d - \frac{p\omega_r \phi_0}{L_s} \tag{33}$$

This model describes the behavior of the synchronous motor in the system joint with the rotor (and the rotating field). To obtain the Matlab/Simulink model of the SPMSM that is similar to a real motor (powered in the fixed reference system), the model described by equations (32) –(33) must be preceded by two transformation blocks: ($3 \rightarrow 2$)si(e^{-j}), which transform the three-phase supply currents into a two-phase system (first transformation), rotating integrally with the rotor (second transformation). Following the integration of the equations of state in the rotating system, the phase currents are obtained by applying the inverse transformations (e^{+j}) and ($2 \rightarrow 3$). The model thus obtained is masked in the "SPMSM" block, in the general control scheme Fig.1. The estimation of the stator flux and rotor speed is performed in the "Gopinath Observer" block, a block that receives the two-phase stator currents of the motor as input. This rotor speed value is used both in the coordinate transformation blocks (for calculating the electrical angle $\theta_{r_{estim}}$) and in the speed loop, thus closing the control chain.

4 Computation results

To validate the proposed control model, MATLAB/Simulink was used as a simulation environment, where several simulations were performed. The model also uses a SPMSM with the following constructive parameters, Table 1.

For the accuracy of the simulation results and their correct visualization on the graphs, 2s simulation times and a sampling step of 10ms were used. Thus, from the graph of rotor speed variation Fig.2 and electromagnetic torque, Fig.3, we can see the stability of the motor's operation with variations in stator resistance. Three values ($R_s = 0.5\Omega$; $R_s = 1\Omega$; $R_s = 1.5\Omega$) were imposed for this resistance and the operating dynamics of the motor were monitored over a 2s period.

Table 1. The parameters of SPMSM

$m = 3$	$\phi_0 = 0.13Wb$
R_s =1 Ω	J =0.01 kgm^2
L_s =0.0032 H	p =3

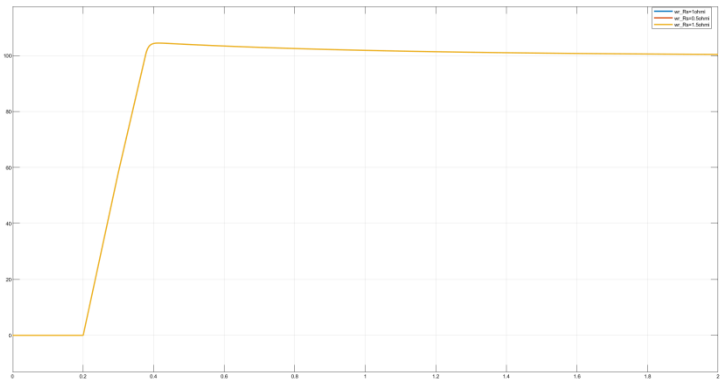


Figure 2. Rotor speed response for ($R_s = 0.5\Omega;\ R_s = 1\Omega;\ R_s = 1.5\Omega$)

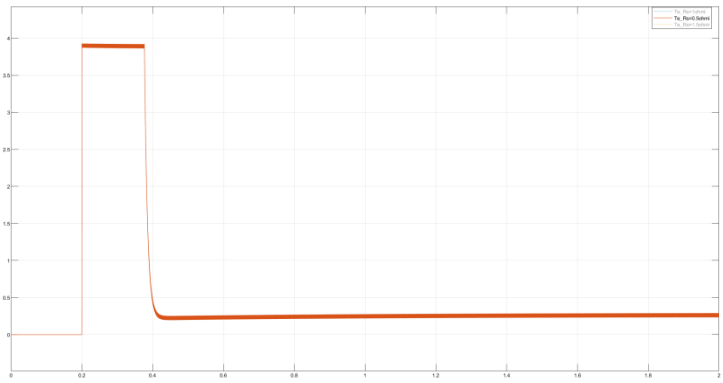


Figure 3. Electromagnetic torque response for ($R_s = 0.5\Omega;\ R_s = 1\Omega;\ R_s = 1.5\Omega$)

In the typical "current-controlled inverter" configuration we have an internal current regulation loop (PI/PI d–q decoupling), with high bandwidth compared to mechanical dynamics. The current regulator forces the d-q axis components of the current to follow the references, compensating for the terms in the electrical equations (dependent on R_s). Thus, for an ideal PI controller that is fast enough, the required control voltage is calculated by the controller so that the current error is zero: that is, the voltage required to compensate for the resistance losses and produce the desired current. So, the variation of R_s produces a change in the voltage value, but the regulator adjusts it to keep the electric current constant.

As observed from the specialized literature [14], in voltage control, the current is not fixed by a regulator — the stator resistance R_s enters directly into the equations of state and influences the current, and therefore the torque and speed.

Another problem analyzed in this paper is the confirmation of the robustness of the Gopinath-type flux observer implemented for the surface permanent magnet synchronous motor (SPMSM). The purpose of this analysis is to demonstrate the observer's ability to correctly estimate the stator flux components even under conditions of variations in machine parameters or rotation speed. Fig.4 illustrates the behavior of the system for a reference rotation speed $\omega_{r_{ref}} = 100\, rad/sec$ and $\omega_{r_{ref}} = 20\, rad/sec$, a regime in which the effects of nonlinearities and resistive terms become more pronounced, and flux estimation is normally more difficult. The obtained results show that the Gopinath observer provides a stable and precise estimation of the stator flux, without oscillations or drifts over time, which confirms the robustness of the method to speed variations and the influence of stator resistance.

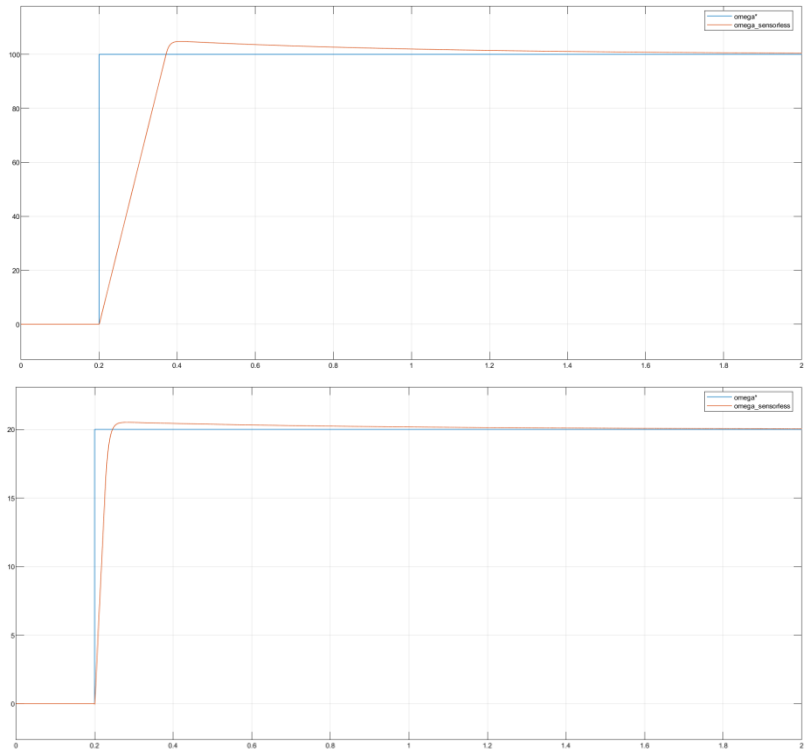


Figure 4. Speed dynamic response ($\omega_{ref} = 100rad/sec, \omega_{ref} = 20rad/sec$)

5 Conclusions

The paper presents an analysis of the performance of the vector control system applied to the permanent magnet synchronous motor (PMSM), with a focus on current control and the evaluation of the robustness of the Gopinath-type flux observer. The main objective is to investigate the stability of the system and the accuracy of rotor speed regulation in the presence of variations in electrical parameters, especially stator resistance.

The current control system uses separate loops to regulate the stator currents on the two orthogonal axes and to regulate the rotor speed. The performed simulations demonstrate that this structure ensures stability and high performance, maintaining the rotor speed independent of stator resistance variations. This confirms the efficiency and robustness of current control, providing a fast and precise response even to load changes or variable operating conditions.

The Gopinath flux observer, used for estimating stator flux and rotor speed, was evaluated in low-speed regimes ($\omega_{ref} = 20rad/\text{sec}$). The results indicate an accurate and stable estimate, demonstrating that the observer is robust to parametric variations and maintains consistent performance under practical operating conditions.

In conclusion, the study confirms that the combination of current vector control and the Gopinath observer provide a robust, stable and precise SPMSM drive system, capable of operating efficiently over the entire speed and load range. The obtained results can serve as a basis for practical implementations and for future research on adaptive observers and methods for online compensation of parametric variations.

References

[1] M.A. Hoque, M.A. Rahaman, *Speed and Position Sensorless Permanent Magnet Synchronous Motor Drives*, IEEE Anadien Conference on Electrical and Computer engineering, CCECE, pp. 689-692, 1994.
[2] P. Vas, *Sensorless Vector and Direct Torque Control*, Oxford University Press, Oxford,pp. 87-137, 1998.
[3] J. Liu, T. A. Nondahl, P. B. Schmidt, S. Royak and M. Harbaugh, *Rotor Position Estimation for Synchronous Machines Based on Equivalent EMF,* IEEE Trans. Ind. Appl., vol. 47, no. 3, pp. 1310-1318, May-June 2011.
[4] H. B. Azza, N. Zaidi, M. Jemli, and M. Boussak, *Development and Experimental Evaluation of a Sensorless Speed Control of SPIM Using Adaptive Sliding Mode-MRAS Strategy,* IEEE J. Emerg. Sel. Topics Power Electronics., vol. 2, no. 2, pp. 319-328, June 2014.
[5] X. Song, J. Fang, and B. Han, *Adaptive compensation method for high-speed surface PMSM sensorless drives of EMF-based position estimation error,* IEEE Trans. Power Electron., vol. 31, no. 2, pp. 1438–1449, Feb. 2016.
[6] G. Wang, Z. Li, G. Zhang, Y. Yu and D. Xu, *Quadrature PLL-Based High-Order Sliding-Mode Observer for IPMSM Sensorless Control with Online MTPA Control Strategy,* IEEE Trans. Energy Convers, vol. 28, no. 1, pp. 214-224, March 2013.
[7] S. Bolognani, R. Oboe and M. Zigliotto, *Sensorless full-digital PMSM drive with EKF estimation of speed and rotor position,* IEEE Trans. Ind. Electron., vol. 46, no. 1, pp. 184-191, Feb. 1999.
[8] N.P. Quang, J., A. Dittrich. *Vector Control of Three-Phase AC Machines*, Power Systems, Springer, vol 4, Jan 2008.
[9] F. Parasiliti, R. Petrella, M. Tursini, *Sensorless Speed Control of a PM Synchronous Motor by Sliding Mode Observer*, Industrial Electronics, 1997, ISIE'97, Proceedings of the IEEE International Symposium on, pp. 1106–1111, 1997.
[10] S. Seung-Ki, *Control of Electric Machine Drive System*, IEEE Press, Series on Power Engineering, Wiley-IEEE Press, pp 283-324, 2011.
[11] R.C. Panaitescu, *A Synchronous Motor Speed Control System Based on the Internal Angle Deviation Control*, Proceedings of the IV-th Conference on Optimiyation of Electric and Electronic Equipements, Braşov, May 12-14, Vol. I, pp 347-350, 1994.
[12] S. Nakashima, Inagki, I. and Miki, I., *Sensorless Initial Rotor Position Estimation of Surface Permanent Magnet Synchronous Motor*, IEEE Trans. Ind. Appl., vol.36, Nov./Dec., 2000, pp. 1598-1603.
[13] M. Tursini, R. Petrella and F. Parasiliti, *Initial Rotor Position Estimation Method for PM Motors*, IEEE Trans. Ind. Appl., vol.39, pp.1630-1640, Nov./Dec. 2003.
[14] G. Craciunas, *Operation of the Synchronous Motor with Surface Permanent Magnet in a Sensorless System with Speed Estimation*, Journal of Electrical and Electronics Engineering, Vol. 17, Nr.2, 2024.

DOI: 10.2478/ijasitels-2025-0003

Deepfake Voice Detection for Underrepresented Languages: A Romanian Case Study

Bălășoiu Robert-Alin[1], Ivașcu Ioana-Daniela[2] , Mihu Cantemir[2], Muntean Robert-Andrei[2], Olescu Marco Leon[3], Resiga Sorana-Ioana[4]

[1]*Politehnica University of Timisoara, Romania*

[2]*Lucian Blaga University of Sibiu, Romania*

[3]*NTT DATA Romania*

[4]*Babes Bolyai University of Cluj Napoca, Romania*

Abstract

The rapid advancement of artificial intelligence and its subsequent application in deepfake media present a significant global security concern. Despite this widespread issue, effective detection solutions are notably absent for less commonly used languages. This paper proposes a potential solution for identifying generated audio in Romanian. The solution centres on an SVM-based algorithm, previously demonstrated to perform effectively in English language tests, adapted with a dataset specifically tailored to the Romanian language. The resulting language-specific model exhibits better performance in differentiating between authentic and synthetic Romanian audio, thereby offering an improvement over general-purpose, Anglocentric systems. This constitutes an effective strategy for developing localised solutions for a language with limited resources.

Keywords: Deepfake detection, Synthetic speech, Romanian language, Support Vector Machine (SVM)

1. Introduction

Artificial intelligence has been one of the most researched topics in recent history. This field has experienced an exponential growth, faster than any of its predecessors; thus, it has created a new cluster of problems that we are still trying to grasp. One of the most pressing issues is the rise of deepfake media. While it existed long before through digital manipulation, AI makes it much more accessible and realistic than ever before. The dangers of it have been thoroughly studied and analysed, but more importantly, they have been felt as a shockwave in our society. Because of its implications in both social and personal spaces, it became imperative that we find tools to mitigate this threat.

Current research in the domain has already identified various methods for handling this problem, from watermarks embedded in the content as it is created to tools that assess the media and assign it a likelihood of being generated by AI. One problem encountered

is the lack of accuracy of these tools for underrepresented languages [1]. English, Spanish, German, and other widely used languages have the advantage of extensive training databases, allowing these tools to be honed and perfected. Unfortunately, Romanian is not among these languages.
The purpose of this paper is to explore findings on this language: how pre-existing tools perform in detecting audio generated by AI, as well as a proposed model tailored to handle this language. The different models that have been tested underperformed, producing false positives on audio that a native speaker would easily categorise correctly. Because of that, a Support Vector Machine (SVM) model has been trained on two databases, MLAAD[2] for generated audios, and UPSDATRO [3] for real audio. Such a focused approach has managed to classify correctly between real and fake audio, especially in a controlled test environment. Despite that, the shortcomings caused by a lack of training data leave room for improvement when it comes to scaling the model and increasing accuracy.
This study confirms that developing targeted, language-specific models is an important advancement in audio detection for languages with limited data resources. Yet, the main problem is caused by the lack of data that can be used for training and testing such models.

2. The AI Detection Model

The proposed detection algorithm was implemented in Python, using librosa, scikit-learn, and joblib for audio processing, classification, and model persistence. The system architecture is designed around a Support Vector Machine (SVM) classifier, trained to distinguish between AI-generated and authentic Romanian speech using Mel-Frequency Cepstral Coefficients (MFCCs) [4] as acoustic features. The implementation is adapted from an open-source project [5].

2.1. Data Processing and Feature Extraction

Before classification, the model converts each audio file into a compact feature representation using MFCCs. MFCCs are among the most widely used feature sets in speech analysis, as they approximate the human ear's sensitivity to frequency and capture a voice's timbral characteristics.
The function `extract_mfcc_features()` loads each .wav file using librosa, computes 13 MFCC coefficients per frame (`n_mfcc=13`), and returns the mean of all frames to form a 13-dimensional vector per sample. This ensures a consistent input size regardless of the audio file's duration.

Listing 1: MFCC Feature Extraction

```
mfccs = librosa.feature.mfcc(y=audio_data, sr=sr,
                             n_mfcc=n_mfcc,
                             n_fft=n_fft,
                             hop_length=hop_length)

return np.mean(mfccs.T, axis=0)
```

This process is identical during both training and inference, ensuring that the model interprets new data consistently.

2.2. Dataset Integration

Two datasets were used to train and evaluate the model:

- **UPSDATRO** – a curated collection of *authentic Romanian speech* samples, serving as the positive (real) class.
- **MLAAD (Machine Learning Audio Anti-Deepfake)** – a collection of *AI-generated* Romanian voices created by text-to-speech (TTS) systems such as Bark or multilingual models, representing the synthetic (fake) class.

Each dataset directory is recursively scanned for .wav files, which are processed to extract MFCCs and labelled accordingly (0 for real, 1 for fake). The combined dataset is then split into training and testing subsets (80/20 ratio) using `train_test.split()` with stratification to preserve class balance.

2.3. Model Architecture

The classifier is a **Support Vector Machine (SVM)** [6] implemented through sklearn.svm.SVC. A linear kernel was selected for its balance between performance and interpretability, and probability estimates were enabled to provide confidence scores during inference.

Listing 2: SVM Classifier Initialisation

```
svm_classifier = SVC(kernel='linear', random_state=42, probability=True)
svm_classifier.fit(X_train_scaled, y_train)
```

Before training, all features are standardised using StandardScaler() to ensure uniform scaling across dimensions, which is crucial for SVMs to compute meaningful margins between classes

2.4. Model Training and Evaluation

Training involves fitting the SVM to the normalised training data and evaluating it on the test set. The model's performance is measured using:

- **Accuracy**: proportion of correctly classified samples.
- **Confusion matrix**: a detailed view of true vs. false predictions for both real and fake classes.

Listing 3: Model Evaluation Metrics

```
accuracy = accuracy_score (y_test,
y_pred )
confusion_mtx = confusion_matrix
(y_test, y_pred )
```

The trained model and scaler are serialised using joblib as svm_audio_classifier.pkl and scaler.pkl, respectively, allowing efficient reuse without retraining.

2.5. Inference Pipeline

For inference, the trained model loads once at runtime:

Listing 4: Model Loading for Inference

```
svmmodel = joblib.load(" svm_audio
classifier .pkl") scaler =
joblib.load("scaler.pkl")
```

Given a new Romanian audio file, the system:

1. Extracts MFCC features.
2. Normalises them using the saved scaler.
3. Uses the SVM's predict proba() method to estimate class probabilities.

The output includes a predicted label ("Real" or "Fake") and a confidence score expressed as a percentage:

Listing 5: Inference Output Generation

```
prediction_proba = svm_model. predict_proba (
features_scaled )[0] confidence = prediction
proba [ prediction_index ] *100
```

2.6. Implementation Summary

Table 1. Summary of Model Implementation

Stage	Description
Feature Extraction Classifier	MFCCs (13 coefficients) computed with librosa
Preprocessing	Support Vector Machine (SVC, linear kernel, probability=True)

Datasets	Standardisation using StandardScaler UPSDATRO (real), MLAAD (fake)
Evaluation metrics Output	Accuracy, Confusion Matrix “Real” or “Fake” + confidence percentage
Model persistence	Saved using joblib

This lightweight, interpretable approach enables high performance even with limited linguistic data, providing a scalable framework for language-specific audio authenticity detection.

3. Experimental Works and Results

3.1. Databases and Training Configuration

To evaluate whether a classical machine-learning method can effectively detect Romanian deepfake speech, we used two publicly available audio datasets representing both authentic and synthetic Romanian voices. For genuine speech, we relied on the **UPSDATRO dataset**, a curated corpus containing **1,001 recordings** of native Romanian speakers. These recordings include a variety of voices, dialects, genders, and background conditions, which provide natural acoustic diversity. For synthetic data, we used the **MLAAD dataset**, which contains **2,637 AI-generated Romanian audio samples** created using multiple modern text-to-speech (TTS) systems. This dataset is part of a multilingual anti-spoofing collection specifically designed for deepfake research.

The choice of these two datasets was guided by the fact that **Romanian is a low-resource language**, with few open datasets for voice anti-spoofing research. Many of the most common deepfake detection models are trained primarily on English or Mandarin, where massive speech corpora are available. However, transferring such models directly to Romanian often results in poor performance due to phonetic and prosodic differences. Thus, our goal was to test whether even a relatively small, language-specific dataset could produce a reliable detection model.

Although the two datasets differ in size, combining them created a **diverse yet balanced training foundation**. Altogether, the combined dataset contained **3,637 samples (1,001 real and 2,637 fake)**. Before training, all audio files were standardised in format, converted to mono, and resampled to a consistent rate to minimise potential artefacts caused by inconsistent sampling frequencies. The data were randomly shuffled and split into **80% for training (2,909 samples)** and **20% for testing (728**

samples), following a common machine-learning convention that reduces overfitting and ensures fair generalisation testing.

To further investigate how dataset structure affects learning, a **second experiment** was conducted with a perfectly balanced corpus by randomly selecting **1,000 real and 1,000 fake recordings**. This setup allowed us to examine whether class balance impacts model generalisation. The same 80/20 ratio was applied, resulting in **1,600 training samples (800 real, 800 fake)** and **400 testing samples (200 real, 200 fake)**. This comparative design helped identify whether the imbalance of the initial configuration caused the model to bias toward either class.

Table 2. Training Configurations

Case	Real Samples	Fake Samples	Total
Initial configuration	1,001	2,637	3,637
Balanced configuration	1,000	1,000	2,000

In both scenarios, the model was trained using **short-form Romanian speech clips**, most of which were clean and of moderate duration. While this structure offered clear advantages for training consistency and computational simplicity, it also limited the exposure of the model to background noise, long speech sequences, and spontaneous conversational dynamics. These factors, as will be discussed later, played an important role in the real-world evaluation phase.

3.2 Audio Duration

The average duration of the audio clips used in training ranged between **2 and 10 seconds**, with most samples centred around **approximately 8 seconds**. This time range was not arbitrary; short-duration speech segments are common in audio classification studies because they capture enough acoustic variability for meaningful feature extraction while maintaining manageable computational complexity. Each clip generally contained a single spoken sentence or phrase, allowing the model to focus on the **core spectral and prosodic characteristics** of the voice without distractions from prolonged pauses or background interference.

Short recordings have both benefits and drawbacks. On the positive side, shorter clips make it easier for traditional machine-learning models, such as **Support Vector Machines (SVM)**, to process input data since these models rely on fixed-length feature vectors and do not model temporal dependencies directly. This makes MFCC-based feature extraction particularly effective, as it compresses the essential characteristics of speech into a compact representation. However, the drawback is that short clips may not include sufficient speech diversity. Variations in tone, rhythm, and articulation that

occur across longer recordings might not be captured, which can reduce model robustness when faced with unfamiliar speaking patterns.

In our dataset, **variability in duration influenced model confidence**. The SVM tended to perform best on clips with durations close to the average training length, as the MFCC features extracted from such samples were consistent with what the model had learned. Extremely short or very long recordings sometimes resulted in unstable predictions because the averaged MFCC features were less representative of typical speech behaviour. This observation suggests that future datasets should aim for greater uniformity in clip duration or employ segment-based approaches that split long recordings into smaller, consistent frames.

Furthermore, short-duration training data mirrors how Romanian speech often appears in real-world scenarios, in brief utterances from phone recordings, short social media posts, or voice assistants. By focusing on these smaller clips, our model simulated practical conditions where users might submit or encounter only a few seconds of speech for verification. However, as demonstrated in Section 3.4, when exposed to long, noisy, or mixed-content recordings, performance decreased significantly. This reveals that **duration variability is not just a technical detail but a major factor shaping generalisation and robustness**.

Table 3. Audio Duration Statistics

Metric	Value
Average Duration	8.32 s
Minimum Duration	0.44 s
Maximum Duration	43.58 s
Standard Deviation	6.78 s
Estimated Total Duration	13.0 h

3.3 Model Training and Evaluation

Our deepfake voice detector was implemented as a **lightweight pipeline** combining Python, **Librosa** for feature extraction, and **scikit-learn** for classical machine learning. Each audio sample was first processed to extract **MFCCs**, a compact representation of how sound energy is distributed across perceptually meaningful frequency bands. The MFCCs were then summarised by computing the **mean of 13 coefficients**, producing a single fixed-length feature vector per recording. This step simplified the input structure while preserving enough acoustic information for distinguishing real and synthetic speech.

Before training, we applied a **StandardScaler** to normalise all features, ensuring that each dimension contributed equally to the learning process. The model was then trained using an **SVM** with a linear kernel. The linear SVM was chosen for three main reasons:

1. It performs well with small and medium-sized datasets, making it ideal for low-resource languages.

2. It offers interpretability by providing clear decision boundaries between real and fake samples.
3. It is computationally efficient, requiring no GPUs and minimal parameter tuning.

For both dataset configurations, an **80/20 split** was used for training and testing. The initial configuration produced **2,909 training samples** and **728 testing samples**, while the balanced configuration included **1,600 training samples** and **400 testing samples**.

Table 4. Initial Split

Set	Train	Test
Real (initial)	800	201
Fake (initial)	2,109	528
Total	**2,909**	**728**

Table 5. Balanced Split

Set	Train	Test
Real (balanced)	800	200
Fake (balanced)	800	200
Total	**1,600**	**400**

The results obtained in controlled testing were **remarkably high**. The model trained on the larger, unbalanced dataset achieved **100% test accuracy**, while the balanced version achieved **99.5%**, with only a few misclassifications. The confusion matrices revealed near-perfect separation between real and fake speech, indicating that the SVM successfully captured distinguishing MFCC-based patterns.

However, such high accuracy must be interpreted cautiously. These values were achieved in a **lab-like environment**, using clean data closely resembling the training samples. While this confirms that the SVM learned to differentiate synthetic from genuine speech under controlled circumstances, it does not necessarily imply the same level of performance in uncontrolled settings. **While these results demonstrated strong classification capability in a controlled setup, the next step was to test whether such performance holds in an uncontrolled, real-world environment.**

3.4 Real-World Evaluation

To assess how the trained model behaves with natural, uncontrolled audio, we performed a **real-world evaluation** using **60 Romanian TikTok clips** and manually collected samples from various online sources. The dataset contained **30 short clips (2–**

5 seconds) and **30 long clips (approximately 8 seconds to over 3 minutes)**, covering a mix of real and AI-generated voices. This test simulated how a detection system might operate in practice, analysing user-uploaded or spontaneous speech data with variable background noise and compression.

When evaluated on this real-world dataset, the model's behaviour diverged significantly from its laboratory results. While our testing yielded an average accuracy of 45%, reflecting the challenge of analyzing raw, unfiltered data that contain real-life interferences, this still represents a significant improvement over our initial tests on the multipurpose models (Gemini, Claude, ChatGPT), which averaged only 20% in the same scenarios. On short clips, it correctly recognised all real recordings but **misclassified nearly all synthetic ones as real**, suggesting that speech generated by modern TTS systems closely mimicked the acoustic cues the SVM had learned to associate with authenticity. For longer clips, the issue became more pronounced — **the vast majority of samples, whether real or fake, were classified as real.**

This bias toward the "real" class suggests that the model's decision boundary, learned from clean MFCC averages, may **underfit the broader variability** of real-world audio. Background sounds, room echo, speech overlap, and microphone differences altered the spectral structure of the audio in ways that the model had never encountered during training. These effects weakened its ability to recognise subtle synthesis artefacts.

Despite the drop in accuracy, these tests were highly instructive. They highlighted that real-world recordings introduce **domain shift**, a condition where test data differ systematically from the training data. In the context of deepfake detection, domain shift can stem from environmental noise, variable speaking styles, or compression artefacts typical of social media platforms. Understanding this mismatch is essential for improving model robustness. A practical next step would involve collecting more diverse samples or augmenting existing data with simulated noise and distortions to make the system better prepared for real-world variability.

4. Experimental Observations and Conclusions

This research highlights both the strengths and limitations of classical machine-learning methods in Romanian deepfake voice detection. The experiments demonstrated that even a simple SVM classifier, when trained on well-prepared and balanced datasets, can achieve excellent performance under controlled conditions. This confirms that **traditional, interpretable models** can remain valuable tools for languages where deep-learning resources and large datasets are not yet available.

At the same time, the **performance gap between controlled and real-world evaluations** reveals an important limitation. The model struggled when exposed to unpredictable or noisy inputs, particularly long, spontaneous recordings. It tended to classify unfamiliar acoustic conditions as "real," suggesting that **MFCC-only features** are insufficient to capture the full range of synthetic voice artefacts in diverse

environments. The strong bias observed toward real samples underscores the need for richer, context-aware features or hybrid architectures.

Two key lessons emerged from these findings:

1. **Language-specific machine-learning models** can outperform generic, multilingual detectors when properly trained, even if they rely on classical techniques.
2. **Diverse and realistic datasets** are essential for robust real-world performance, particularly in underrepresented languages where data scarcity amplifies overfitting risks.

Future work will focus on expanding the dataset with additional speaker profiles, longer and noisier recordings, and a broader range of TTS-generated voices. Implementing **audio segmentation**, **data augmentation**, and **hybrid SVM-neural approaches** could further enhance adaptability. By combining the interpretability of SVMs with the feature-learning capabilities of neural networks, future systems could balance accuracy and efficiency while maintaining accessibility for Romanian researchers and institutions.

Ultimately, this project demonstrates that even in the absence of large-scale computational infrastructure, **meaningful progress can be achieved through careful dataset design, balanced experiments, and systematic evaluation**. The lessons learned from this Romanian case study may serve as a foundation for developing deepfake detection frameworks for other low-resource languages facing similar challenges.

References

[1] Joshi, Pratik, Sebastin Santy, Amar Budhiraja, Kalika Bali, Monojit Choudhury. *The State and Fate of Linguistic Diversity and Inclusion in the NLP World.*, Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 6282–6293, ISBN N/A, Online, 2020.
[2] Müller, Nicolas M., Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, Konstantin Böttinger. *MLAAD: The Multi-Language Audio Anti-Spoofing Dataset.*, arXiv preprint arXiv:2401.09512, ISBN N/A, Online, 2024.
[3] Păiș, Vasile, Verginica Barbu-Mititelu, Elena Irimia, Radu Ion, Dan Ioan Tufis. *USPDATRO: Under–Represented Speech Dataset from Open Data – Case Study on the Romanian Language.*, Zenodo Repository, DOI:10.5281/zenodo.7898232, ISBN N/A, Online, 2024.
[4] Ahmad, Wali Muhammad. *AI Voice-Detection (TensorFlow).*, GitHub Repository, https://github.com/WaliMuhammadAhmad/AIVoice-Detection/tree/master/tensorflow, ISBN N/A, Online, 2025.
[5] Sahidullah, Md., Tomi Kinnunen, Cemal Hanilçi. *A Comparison of Features for Synthetic Speech Detection.*, Proceedings of Interspeech 2015, 2087–2091, ISBN N/A, Dresden, Germany, 2015.
[6] Cortes, Corinna, Vladimir Vapnik. *Support-Vector Networks.*, Machine Learning, Volume 20, Issue 3, Pages 273–297, ISSN 0885-6125, Springer, New York, 1995.

DOI: 10.2478/ijasitels-2025-0009

SDR receiver for power measurement with internal noise compensation

Iulian Bouleanu[1], Mădălina Antoneac[1], Macarie Breazu[1], Alina Cristina Viorel[1], Gabriela Crăciunaș[1]

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

{iulian.bouleanu, madalinamaria.antoneac, macarie.breazu, alina.viorel, gabriela.craciunas} @ulbsibiu.ro

Abstract

This paper presents a solution for implementing a power measurement receiver made with PlutoSDR. The paper proposes the dynamic elimination of the contribution of internal noise to the measured signal level by compensating with the level of internal noise produced by the receiver in the adjacent RF band. Our approach is based on the fact that the internal noise level in the working band is approximately equal to that in the frequency band in the immediate vicinity. As far as we know, our work is the first to use the idea of measuring the noise in an adjacent band to estimate the internal noise of the measuring instrument. Another aspect necessary to ensure a complete measurement solution that does not require pre-measurement calibration concerns the compensation of the nonlinearities of the reception path. The compensation was automatically performed by modifying the amplification of the reception path depending on the working frequency. For high levels of the measured CW signal (-50 dBm ... -10 dBm) the solution allows obtaining measurement errors of less than 0.55 dB. From the perspective of measurement uncertainty, with a value of $U = \pm 1$ dB (k=2), the instrument with the method used fits very well into the requirements for pre-compliance EMC measurements. For levels lower than -60 dBm, the error starts to increase rapidly, so that at values of -70 dBm the signal can no longer be measured.

Keywords: Software Defined Radio (SDR), Spectrum Analysis, Internal Noise Correction, Signal Processing

1 Introduction

The electromagnetic spectrum represents one of the most valuable resources of the modern interconnected world. In a time when wireless applications have experienced explosive growth, accurate monitoring and measurement of energy levels of electromagnetic waves is absolutely necessary to verify compliance with international regulations on human exposure to EM fields or to ensure the electromagnetic compatibility of electrical and electronic systems [7].

Conventional spectrum analyzers, although more accurate, are much more expensive, heavier and bulkier and are not suitable for continuous electromagnetic field mobile monitoring situations.

The purpose of this article is to present a solution for the design and calibration of an electromagnetic field measurement receiver that is sufficiently accurate, lightweight and cost-effective in order to facilitate measurements under mobile conditions by placing, for example, the entire measurement system on a drone. The key to achieving low costs and low weight is the use of SDR technology.

Software Defined Radio (SDR) is a flexible radio communication architecture where traditional hardware operations realized by analogical components - mixers, filters, amplifiers, modulators - are replaced by software algorithms running on a general-purpose processor or FPGA [14].

In a typical SDR system the analog front-end converts incoming electromagnetic waves into an intermediate frequency (IF) or baseband signal, which is then digitized by an Analog-to-Digital Converter (ADC). The digital signal is processed through a sequence of algorithms which performs operations such as filtering, demodulation, and spectral analysis [13]. This approach is possible as a result of using the equivalent bandgap theorem [16]. The front-end analog converter translates the spectral components contained in the RF band of interest to much smaller frequency values, to which the sampling theorem can be applied, maintaining the shape of the spectral density distribution.

In [3] Costouri et al. use the idea of measuring the noise in the adjacent band to estimate the external noise. As far as we know, our work is the first to use the idea of measuring the noise in an adjacent band to estimate the internal noise of the measuring instrument.

2 Theoretical Background and Method

2.1 Electromagnetic radiation

Electromagnetic (EM) radiation encompasses oscillations of electric and magnetic fields that propagate through space. These oscillations transport energy, and their behavior depends on parameters such as frequency, wavelength, phase and amplitude [9]. The electromagnetic spectrum covers all possible frequencies from near zero Hertz to infinite.

Electromagnetic waves can be produced naturally by a series of phenomena but also artificially with technologies developed by humanity in the last two centuries. As new technologies have begun to use more and more frequency bands, the problem of monitoring the electromagnetic spectrum in order to achieve effective management has become increasingly acute.

Sources generating electromagnetic waves produce spectral components in different widths of frequency bands and can propagate over significant distances. From a human perspective, they can be useful (for example, when they carry information) or parasitic (when they disrupt the functioning of some systems or create radiation dangerous to human health).

Signal propagation is influenced by path loss, atmospheric absorption, diffraction, and multipath reflection [5]. The power density (S) of an electromagnetic wave decreases proportionally to the square of the distance (r) [9], [16]:

$$S = \frac{P_t}{4\pi r^2} \quad (1)$$

where P_t is the transmitted power.

The power density of a source that generates electromagnetic waves is distributed in a continuous spectrum of frequencies and its level can vary both due to the propagation conditions and due to the operating principle of the source. As a result of the superposition of contributions from two or more sources, at each point in space there are different power distributions for the entire frequency spectrum.

Any electronic equipment, including those designed to perform measurements, generates parasitic electromagnetic waves by their very operation. These waves contribute to the formation of an electromagnetic noise floor that establishes the minimum threshold of an external electromagnetic wave that can be measured.

A signal in the frequency domain is typically represented by its Power Spectral Density (PSD), which quantifies power per unit bandwidth [11]:

$$PSD(f) = \frac{|X(f)|^2}{B} \quad (2)$$

where $X(f)$ is the Fourier transform of the time-domain signal $x(t)$ and B is the observation bandwidth.

The accurate estimation of PSD is central to EM field measurements. Modern communication networks operate across shared frequency bands [7]. To prevent interference and ensure compliance with emission regulations, continuous spectrum monitoring is necessary [7].

2.2 Mathematical model of real-time spectrum measurement

In order to correctly evaluate the EM field level at the receiver input it is necessary to mathematically model the contribution of internal noise and amplification variations produced by the translation of the spectrum from RF to the equivalent baseband at the signal processing stages.

Signal power can be evaluated in both the time domain and the frequency domain. In the frequency domain the evaluation involves summing the power contribution of each spectral component in the RF band. In the time domain the power is obtained by calculating an average value of the power of the measured signal samples over a large number of periods.

Since evaluating power in the time domain requires less processing power (it does not also require performing FFT), this solution was chosen in this work.

Let x(t) represent th e received signal at the SDR input. The measured digital signal *y(t)* after amplification and filtering can be expressed as [15]:

$$y(t) = G(t) \cdot x(t) + n(t) \quad (3)$$

where *G(t)* is the time-varying system gain, and *n(t)* represents additive noise (thermal + quantization).

The instantaneous power $P_{inst}(t)$ is computed as:

$$P_{inst}(t) = |y(t)|^2 \quad (4)$$

The mean received power over a time window *T* is [17]:

$$P_{avg} = \frac{1}{T}\int_0^T |y(t)|^2 dt \quad (5)$$

Expressed in logarithmic form, the result is [12]:

$$P_{dBm} = 10 \cdot \log_{10}\left(\frac{P_{avg}}{1mW}\right) \quad (6)$$

To compensate for internal offsets, noise and nonlinearity, a correction factor (C_{corr}) is introduced [18]:

$$P_{corr} = P_{dBm} - C_{corr} \quad (7)$$

Into the correction factor, two components can be identified:

$$C_{corr} = C_{nonlin} + C_{noise} \quad (8)$$

where:

- C_{nonlin} will correct the amplification nonlinearities of the radio reception down-convertor path branch in PlutoSDR.
  This can be determined as the difference between the actual, correct level of a CW signal and the level displayed by the receiver:

$$C_{nonlin} = P_{real_{dBm}} - P_{meas_{dBm}} \quad (9)$$

- C_{noise} will correct for internal noise produced by the receiver operation. Internal noise is a major limitation in SDR-based measurements [10]. Mixer leakage, quantization artifacts and the intermodulation products due to the nonlinear operation of some reception path stages introduce unwanted energy that affects power readings. The proposed system compensates for these effects by measuring the baseline noise floor when external signal is present C_{int_noise} correction constant is derived as:

$$C_{noise} = 10 \cdot \log_{10}\left(\frac{P_{noise}}{1mW}\right) \quad (10)$$

  where P_{noise} is the measured internal noise power in the immediate vicinity of the measured signal. In this category, in addition to thermal noise and quantization noise, leakage due to the local oscillator (LO) as well as parasitic products resulting from the nonlinear operation of some stages on the reception path will also have to be taken into account. Except for LO leakage, the other internal noise components are additive and dependent on the real-time bandwidth.

Thermal noise depends on the bandwidth of the signal taken into real-time analysis and can be determined mathematically with:

$$P_{th\_noise_{dB}} = 10 \cdot \log_{10}(kTB) \quad (11)$$

where: k – Boltzmann constant (1,38·10−23), T - receiver operating temperature in degrees Kelvin and B - real-time bandwidth.

To evaluate the power according to (5) a moving average filter of *N* samples is used to smooth measurement fluctuations:

$$P_{filtered}[k] = \frac{1}{N}\sum_{i=0}^{N-1} P_{corr}[k-i] \quad (12)$$

This process stabilizes output readings and eliminates short-term variations.

2.3 Calibration Method

The calibration method involves dynamically changing the value of the correction coefficients C_{nonlin} and C_{noise} in the signal processing flow. These correction terms need to be subtracted from measured power levels.

The nonlinearity coefficient of the reception path, C_{nonlin}, is specific to each receiver and can be obtained with the measurement setup shown in Fig. 1.

A calibrated signal generator provides a controlled RF signal, which is transmitted to the SDR platform acting as the receiving and processing device. In order for the level value at the signal generator input to be within a known error margin, a characterized RF cable must be used. The SDR platform captures the input signal and forwards it to a computing system running the GNU Radio software environment, where the digital signal processing flow is implemented. Within GNU Radio, several processing blocks perform filtering, power estimation, and visualization of the measured spectrum in real time.

The value of nonlinearity coefficient is frequency dependent and its determination requires measurements across the entire SDR platform frequency band on as many frequencies as possible. For each frequency into a chosen frequency set the actual level at the SDR input is compared with the level displayed after processing by the GNU Radio power measurement application.

The noise coefficient, C_{noise}, is also specific to each receiver. It is a well-known problem with SDR receivers that they produce parasitic components near zero frequency that form a peak. This is due to the coupling of the mixer with the local oscillator, the imperfections of the mixer, the offset of the amplifiers and the differences between the I and Q branches. For a more accurate assessment of the received signal level, it is desirable to avoid translating the measured signal spectrum over this spike.

In this situation, if the spectrum of the measured signal can be placed completely on one side or the other of the zero frequency, the other half can be used as a reference for evaluating the C_{noise}. The fact from which this work starts is that in this band, if it does not contain other external signals, the level of internal noise is approximately the same value as that in the measured signal band.

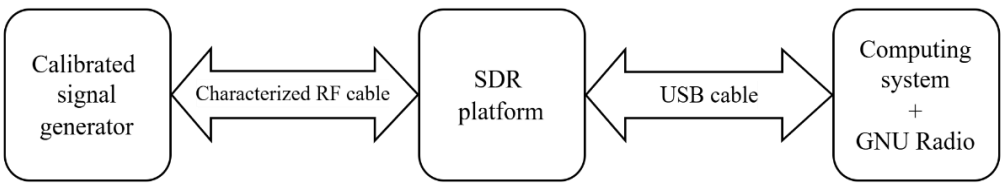


Figure 1. Block diagram of the measurement system used to determine the nonlinearity coefficient

3 Implementation, measurements and results

3.1 ADALM PlutoSDR

Pluto SDR is a SDR platform made by Analog Devices. This platform can digitize, process, and analyze radio signals in software, allowing flexible and reconfigurable implementations without specialized analog components [14].

PlutoSDR operates natively within the frequency range of 325 MHz to 3.8 GHz, a limitation determined by its internal components, particularly the AD9363 Radio Frequency Integrated Circuit (RFIC), which is optimized for this range. However, the device can be customized to extend its operational bandwidth, covering frequencies from 70 MHz up to 6 GHz [1]. This customization is achieved through a software modification involving a firmware update, which enables additional capabilities of the AD9363 RFIC that are normally disabled by default [1].

The device provides 12-bit ADC resolution and communicates via USB2.0 with a laptop or a desktop where data will be processed with GNU Radio, MATLAB or other SDR frameworks.

GNU Radio provides a graphical flow-based interface (GNU Radio Companion, GRC) that allows the design of signal processing chains without need of complex programming [4].

3.2 Flowchart for power measurement

The signal received by PlutoSDR via cable from the signal generator is first down-converted in frequency around the value of zero, digitized by the ADC located in PlutoSDR and passed into the computing system through the USB2.0 socket where GNU Radio perform data processing. The sequence of processing blocks used is presented in Figure 2.

The data is fed into the processing flow by the PlutoSDR Source block. It controls the reception side of the PlutoSDR platform by setting, among other things, the reception frequency (LO Frequency), the RF bandwidth and the receiver gain. The gain must be set manually so that it does not automatically adapt to the measurement conditions by introducing an uncontrollable variable into the received signal amplification chain. The sampling frequency is set with the *samp_rate* variable to the value of 196kHz.

The data from the PlutoSDR Source output are complex numbers that represent the I and Q values of the equivalent baseband signal samples. Because the RF signal is brought directly into the equivalent baseband and not on an intermediate frequency, if we view the signal spectrum at this point, we will notice parasitic components due to the frequency shift process performed on the PlutoSDR receive path.

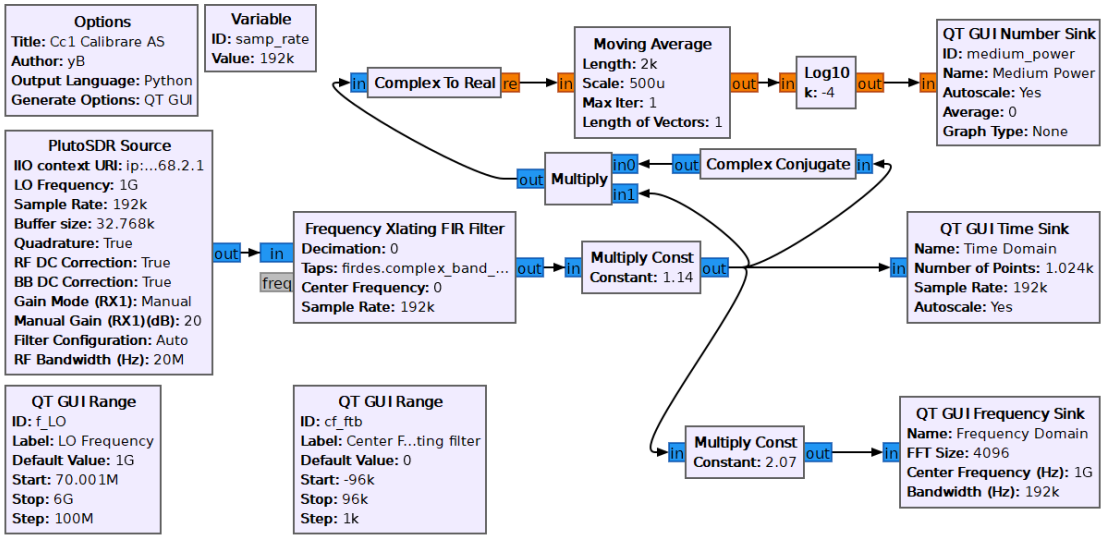


Figure 2. GNU Radio flowgraph of the SDR-based measurement chain

In order for the parasitic near zero components not to affect the measured signal, it is desirable that the spectrum of interest does not overlap the zero frequency. This is achieved with the Frequency Xlating FIR Filter, which can move the spectrum of interest around the zero frequency with a maximum value of (+/-)*samp_rate*/2. The center frequency of the XLating filter will be changed with the slider made with QT GUI Range with id: *ct_ftb*. Being a filter, it will also allow interactive filtering of spectral components [6]. Changing the bandwidth of this filter will introduce changes in the correction coefficients, but the analysis of these changes is not the subject of this study.

The Multiply Const blocks used in the processing flow will increase / decrease the value of the samples. They are used to allow the final level to be adjusted regardless of the working frequency and bandwidth. In order to achieve visual feedback regarding the level and shape of the measured signal, we considered it is necessary to present the result in three display modes:

1. In the time domain, with QT GUI Time Sink.
2. In the frequency domain, with QT GUI Time Frequency Sink.
3. Numerically, by displaying the signal power level, in dBm units, with QT GUI Number Sink.

The power level of the received signal is computed in the following steps:

1. The instantaneous power of each sample is obtained, according to (4), by multiplying the value by the complex conjugate of the value, and subsequently transforming the result into a real number (Complex to Conjugate, Multiply, Complex to Real).

2. Calculate the average power over a large number of periods N, according to (5), respectively (11) with the Moving Average block. The power is in linear power units.

3. Obtain the value in dBm according to (6) with the Log10 block.

3.3 Calibration

To obtain reliable results, the characterization of the device was performed using a calibrated signal generator Anna Pico APSIN20G [2].

A frequency step of 100 MHz is established and the PlutoSDR working band is divided by this value, determining the frequencies at which the frequency coefficients will be determined. Therefore, all measurements in this study will be performed on the same set of frequencies uniformly distributed in the range of 70 MHz to 6 GHz.

In a first step, the values of the multiplication constants in the processing flow will have to be set up. With the signal generator, a continuous wave (CW) signal with a level of -50 dBm is generated and the entire frequency band is scanned to identify the frequency at which the level is maximum. The constants are modified so that the displayed level is -50 dBm. It is taken into account that the cable between the generator and PlutoSDR has an attenuation that increases with frequency.

In the case of the PlutoSDR used, f_{LO}=4 GHz, C_1=1.14, C_2 = 2.07 and k= -4 were identified. The measurement result can be seen in the graphical user interface shown in Fig. 3. The measured signal is of the CW type: a single spectral component of -50dBm level in the frequency domain and two sinusoidal components, affected by noise, in the time domain (representing the real and imaginary part).

This is the signal that we want to measure with the highest possible accuracy. The noise floor is at -110 dBm. The spectral component corresponding to the 4 GHz frequency is shifted by approximately 50 kHz and is due to the PlutoSDR LO generator (adjusting this deviation is not the subject of this study). On the left side of this component (where PlutoSDR shows 4 GHz) a group of spectral components in a bandwidth of a few kHz is observed. These frequencies represent the near-zero parasitic spike components. In the spectral domain, a group of other parasitic components is also observed, with values

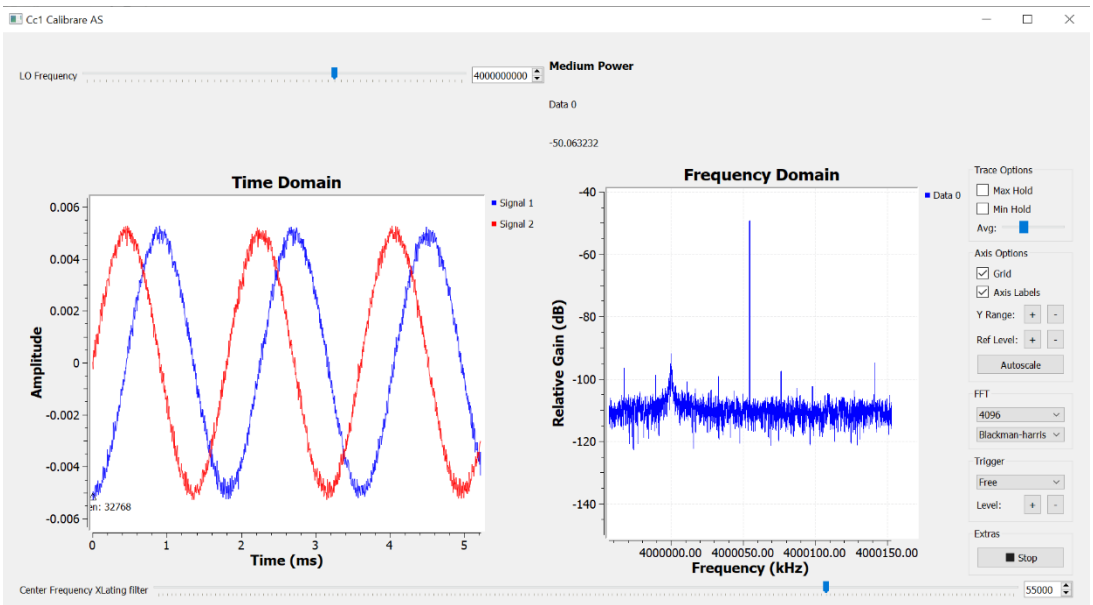


Figure 3. GUI with results for f_{LO}=4 GHz and N=-50dBm

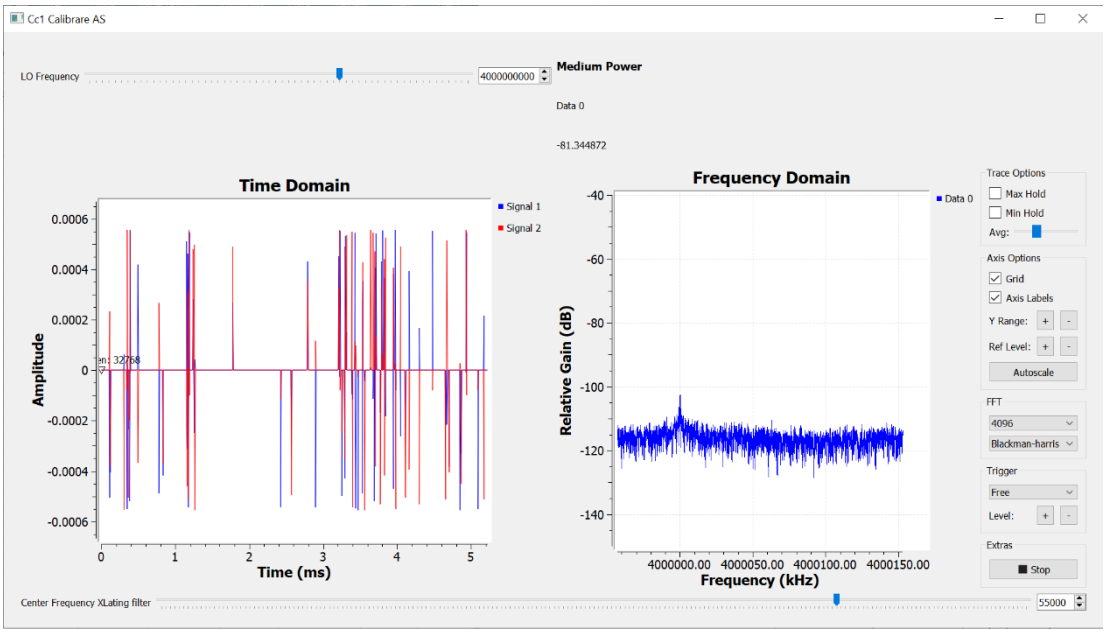


Figure 4. GUI with results for no signal (only internal noise)

of approximately 10 dB above the noise floor. These, together with the near-zero spike components but also with a large part of the noise floor value, must be compensated using the coefficient C_{noise}. If these components due to internal noise did not exist, the noise floor should be made only by the thermal noise components.

The results in Fig. 4 are obtained in the absence of an RF signal with the measurement system in Figure 1. The RF cable was disconnected and a 50 Ohms load was connected to the Pluto SDR receiving antenna.

For the bandwidth used in this application (192kHz), according to (11) the thermal noise contribution is -121 dBm. Comparing the results in Fig. 3 with those in Fig. 4 we find that a difference of approximately 40 dB is due to the other types of noise except thermal. Upon closer analysis we find that in Fig. 3, outside the spectral component of the measured signal positioned at approximately 4 GHz, a series of other spectral components appear visible arranged above the noise floor with levels of approximately 10 dB.

These, together with those that in Fig. 3 versus Fig. 4 raise the noise floor by a level of approximately 10 dB, are parasitic signals caused by the processing of the measurement signal on the reception path. They are easy to visualize because we know that the measured signal must contain only one spectral component because it is CW. If the bandwidth of the measured signal were larger, these could no longer be separated.

We conclude that, where it possible to estimate C_{noise} in the presence of signal, this could lead to an increase in the accuracy of the measurement.

3.4 Power measurements with internal noise compensation

In order to perform this operation dynamically without the need to modify the hardware setup before each measurement, we propose that the noise level estimation be performed in the immediate vicinity of the measured signal band, on the same bandwidth. The solution we have used in this study is: divide the received bandwidth into two equal parts with respect to the zero value, translate the entire spectrum of the measured signal into one of the two halves, measure the signal power on each of these and extract the noise power from the measured signal which contains both the useful signal and the noise.

If, after frequency tuning, the spectrum of the signal to be measured is at positive frequencies, the two Frequency Xlating filters will separate the input data so that the positive part of the spectrum (signal + noise) will be processed on the upper branch and the negative part of the spectrum (noise) on the lower branch.

If no signal is applied at the input, the difference between the power of the samples processed on the two paths is -380 dBm, so the two lines are balanced. When introducing a useful signal with a frequency of 4 GHz from the signal generator, for a dynamic range between -60 dBm and -10 dBm the average error is below 1 dB. At the other frequencies in the PlutoSDR working range the average error increases to differences of +/- 4,15 dB, an aspect due to the non-uniformity of the receiver's amplification.

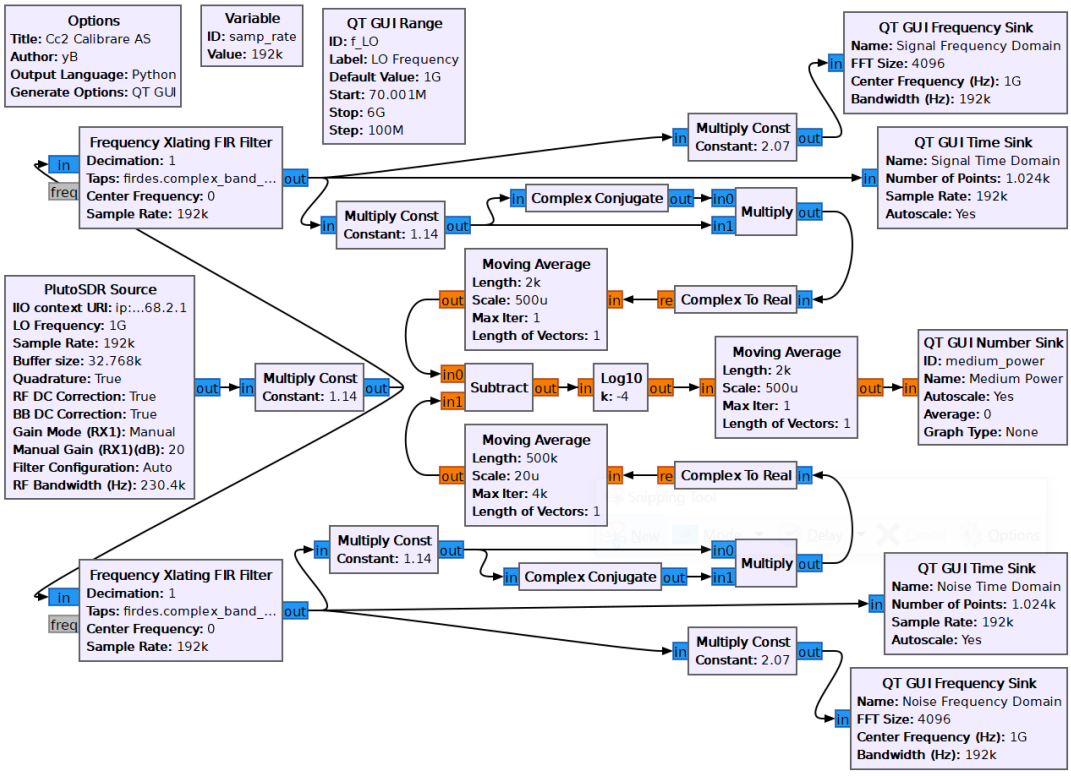


Figure 5. Flowchart for power measurements with internal noise self-compensation

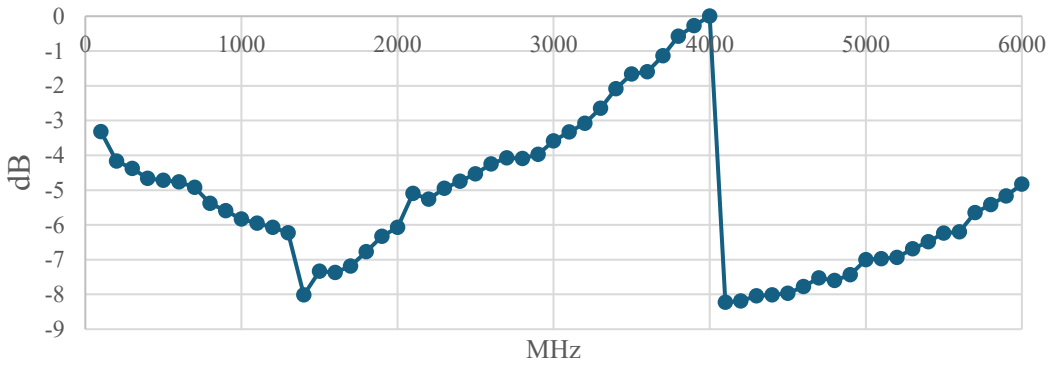


Figure 6. PlutoSDR radio receiver amplification characteristic (C_{nonlin})

To determine the nonlinearity compensation coefficient, C_{nonlin}, an experiment was performed using the measurement scheme in Fig. 5 and the following algorithm:

1. For each frequency, the previously determined measurement coefficient is taken into account.
2. For each frequency, a CW signal with $N_{generator} = N_{inSDR} + a_{cable}(f)$ is generated from a calibrated signal generator (AnaPico APSIN20G), where $a_{cable}(f)$ is the amount of attenuation introduced by the RF cable between the generator and PlutoSDR and $N_{inSDR} = -50dBm$
3. The level is measured with the Fig. 5 flowchart.
4. The value of the correction factor is computed as the difference between the real level and the measured level.

The results are presented in Fig. 6. The nonlinearity coefficient is added to the value of the Gain parameter in the PlutoSDR Source block so that the receiver compensates the gain depending on the working frequency.

3.5 Results

After modifying the flowchart so that the correction is performed dynamically, measurements were performed on all frequencies in the set with input signal values in the range - 60dBm ...- 10 dBm. A summary of these results is presented in Table 1.

Table 1. An excerpt of measurement results

N_{inSDR} [dBm]	N_{meas} [dBm]							$\lvert\Delta N_{max}\rvert$ [dB]
	0.1 GHz	**1 GHz**	**2 GHz**	**3 GHz**	**4 GHz**	**5 GHz**	**6 GHz**	
-10	-10.41	-10.55	-10.40	-10.28	-10.30	-10.02	-10.00	0.55
-20	-20.05	-20.19	-20.06	-19.98	-20.20	-20.13	-20.17	0.20
-30	-30.12	-30.34	-30.11	-29.75	-30.20	-30.24	-30.13	0.34
-40	-40.00	-40.09	-39.90	-39.60	-40.00	-39.88	-40.02	0.40
-50	-49.98	-50.02	-49.97	-49.50	-50.00	-49.99	-50.02	0.50
-60	-61.16	-61.30	-61.08	-60.51	61.16	-60.99	-61.26	1.30

3.6 Measurement uncertainty

The receiver power level (dBm) measurement uncertainty was evaluated based on an uncertainty budget that includes the contributions of the APSIN20G generator, the Withwave W701-SM1SM1-1m measurement cable and the receiver to generator calibration residual.

The measurements were performed in the range 100 MHz – 6 GHz, with a step of 100 MHz, for input levels appropriate to the receiver's range of use. At each frequency point the difference was determined (Table 1):

$$\Delta N = N_{inSDR} - N_{meas} \tag{13}$$

The observed experimental differences were |ΔP| < 0.55 dB over the entire frequency range.

When assessing the uncertainty, the following were taken into account:

- Declared uncertainty of the APSIN20G generator [2] in the 0.2–5.5 GHz band, $U_{APSIN} = \pm 0.6$ dB, for which the standard uncertainty $u_{APSIN} \approx 0.35$dB is obtained.
- An overall tolerance of +/-0.2 dB for the Withwave W701-SM1SM1-1m cable, for which the standard uncertainty $u_{cable} \approx 0.12$dB is obtained.
- Experimentally measured comparison residue, $U_{residue} = \pm 0.55$ dB, for which the standard uncertainty $u_{residue} \approx 0.32$dB is obtained.

Each contribution was treated as a rectangular distribution and the associated standard uncertainties were combined using the root-sum-of-squares (RSS) method.

The resulting combined standard uncertainty is $u_c \approx 0.49\, dB$ which leads to an expanded uncertainty:

$$U = \pm 1 \text{ dB (k=2)} \tag{12}$$

valid for the power level measured in dBm, in the frequency range 100 MHz – 6 GHz, under similar conditions of use and calibration.

3.7 Discussions

When measuring CW signals with levels in the range -50dBm...-10dBm, the measurement scheme responds very well, the maximum error being 0.55 dB. For lower levels the error increases to 1.3 dB for measured signal levels of -60 dB. For levels lower than -60 dB the error increases significantly so that at -70dB, the signal can no longer be measured.

From the visual analysis of the measured signal's spectra with the flow chart in Fig. 5 we found that at some frequencies of the measurement signal and at values higher than -40 dBm intermodulation products appear. Their number and value increase with the signal level. However, the measurement scheme manages to compensate for their contribution to the final result because these intermodulation products are also found in the lower spectrum. If at high levels the scheme performs surprisingly well, at levels lower than -60dBm the results are very poor, practically failing to measure signals with values lower than -65dBm.

With a level uncertainty of ±1 dB the instrument complies:

- with the typical range for EMC pre-compliance systems (±1…3 dB);
- with the lower (better) uncertainty range of commercial pre-compliance receivers (±1.5…2.5 dB);
- below the limits required by CISPR 16 for compliance laboratory equipment (±2 dB).

4 Conclusion

This paper presents a solution for implementing an electromagnetic field strength measurement receiver made with the PlutoSDR platform produced by Analog Device. In order to eliminate the contribution of internal noise to the measured signal value, a dynamic self-calibration solution was used. The solution compensates for the value of the internal noise level produced by the receiver. The principle used is that the level of internal noise in a frequency band in the immediate vicinity of the working band is approximately equal to that in the working band. As far as we know, our work is the first to use the idea of measuring the noise in an adjacent band to estimate the internal noise of the measuring instrument.

A second aspect addressed in the paper (in order to offer a complete measurement solution that does not require pre-measurement calibration) concerns with the compensation of the nonlinearities of the reception path. The compensation was automatically performed by modifying the amplification of the reception path in PlutoSDR depending on the working frequency in order to obtain a flat characteristic of the reception path throughout the working band.

For high levels of the measured CW signal (-10 dBm ... -50 dBm) this solution allows to obtain measurement errors smaller than 0.55 dB. Regarding measurement uncertainty, having a value of $U = \pm 1$ dB (k=2), the instrument (with the method used) fits very well into the requirements for pre-compliance EMC measurements. For levels lower than -60 dBm the error starts to increase rapidly, so the instrument can no longer be used. We assume that the lower limit of the measurement threshold is due to the low amplification used in the amplification chain (20 dB). This hypothesis will be the subject of a new study in which the quality of the solution will be verified also in the case of broadband signal measurements.

References

[1] Analog Devices, *ADALM-Pluto SDR Active Learning Module, Technical Documen*tation, 2023.
[2] AnaPico *APSIN20G Datasheet*, https://www.testequipmenthq.com/datasheets/ANAPICO-APSIN20G-Datasheet.pdf, accessed 03.11.2025
[3] Costouri, J. Nessel and G. Goussetis, "Validation of a Digital Noise Power Integration Technique for Radiometric Clear Sky Attenuation Estimation at Q-Band," in IEEE Transactions on Antennas and Propagation, vol. 68, no. 9, pp. 6743-6751, Sept. 2020, doi: 10.1109/TAP.2020.3001452.
[4] E. Blossom, *GNU Radio: Tools for Exploring the Radio Frequency Spectrum*, Linux Journal, 2004.
[5] Goldsmith, A., *Wireless Communications*, Cambridge University Press, 2005.
[6] H. Zhang et al., *Noise Reduction in SDR Measurement Systems Using Adaptive Filters*, IEEE Sensors Journal, 2023.

[7] IEEE, *Standard for Spectrum Management and Measurement*, IEEE Std. 1900.5-2019.
[8] ITU-R, *Radio Regulations - Frequency Allocations*, International Telecommunication Union, 2023.
[9] J. D. Kraus, *Electromagnetics*, McGraw-Hill, 2002.
[10] J. P. Darnell, *Noise Compensation Techniques in Software-Defined Radios*, IEEE Trans. on Instrumentation and Measurement, 2019.
[11] L. Peng, *Spectral Power Estimation Techniques for SDR Systems*, IEEE Trans. on Signal Processing, 2021.
[12] M. Rice, *Digital Communications: A Discrete-Time Approach*, Pearson, 2008.
[13] P. Kenington, *RF and Baseband Techniques for Software Defined Radio*, Artech House, 2005.
[14] R. Mitola, *The Software Radio Architecture*, IEEE Communications Magazine, vol. 33, no. 5, 1995.
[15] S. Haykin, *Communication Systems*, 5th ed., Wiley, 2009.
[16] T. S. Rappaport, *Wireless Communications: Principles and Practice*, Prentice Hall, 2010.
[17] V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*, 3rd ed., Prentice Hall, 2010.
[18] W. B. Cook, *Low-Cost SDR Solutions for EMF Measurement*, Sensors Journal, 2021.

DOI: 10.2478/ijasitels-2025-0004

Forecasting and reliability

Ioan Lucian Diodiu[1]

[1]*Computer Science and Electrical Engineering Department, Lucian Blaga University of Sibiu, Victoriei Street, no 10 , Sibiu*

Abstract

No precise law governs the design of electrical power systems, and due to the errors inherent in the results obtained, confidence in forecasting methods is low. In reality, a large number of possible solutions may arise, and one aims to determine the limits within which they will be found — more precisely, within what boundaries a certain proportion of them lies.

Keywords: reliability, probability, law

1 Introduction

1.1 Probability

No precise law governs the design of electrical power systems, and due to the errors inherent in the results obtained, confidence in forecasting methods is low. In reality, a large number of possible solutions may arise, and one aims to determine the limits within which they will be found — more precisely, within what boundaries a certain proportion of them lies.

Nothing can be predicted through a purely deductive process; it is recognized by the axioms of any logic that no deductive process can lead to the demonstration of a proportion related to forecasting.

No two events are identical, yet certain factors common to them must be selected and categorized as favorable or unfavorable.

The probability of success in an experiment is defined as the ratio between the number of possible cases in which a favorable result occurs and the total number of possible outcomes. Ultimately, nothing can occur twice in exactly the same way; this introduces the ratio of two infinite numbers. However, in practice, possibilities are grouped discretely, and the ratio is taken between the number of possible results in the favorable subgroup and the total number of possibilities. This statement cannot be considered a vicious circle.

Starting from this point, the mathematical procedure follows the path of measurement theory, defining an event as an element belonging to a set; however, this consideration is avoided here in order to ensure simplicity and speed.

In light of the above ideas, Bernoulli’s theorem becomes a tautology. It can thus be stated as follows: if the ratio between the number of favorable results and the number

of possible results in a single trial is p, then the ratio between the number of ways in which the above condition occurs in n trials and the total number of possible results tends toward 1 as n tends toward infinity. A careful examination of this theorem shows that, in reality, what has been stated indeed occurs, and the statement is self-evident or tautological.

1.2 Practical Implication of Definition of Probability

The probability of an event is, in fact, an important and the only scientific basis we have for a rational decision. In practical terms, the ratio of the number of ways—or the probability—represents the degree of confidence we have in a rational assumption or statement we hold, and it is the basis upon which we could rationally wager.

Nevertheless, one must respond to a criticism, since there exists a new logical sequence in the reasoning above. If it can be said that the probability of an event corresponds to the chance at which a reasonable person would be equally willing to propose and accept wagers on a trial, then we would have a method for putting the theory into practice [1]. However, there are no serious reasons to affirm this. There are many people who accept Macpherson's law. It must be noted that such convictions can remain fixed in a person's memory and may obscure many cases in which everything went well. It appears that Macpherson considers himself a conventional inductionist and may make the same claims as any other inductionist.

Probability means only the ratio between favorable cases and the total number of possible cases, and this definition says nothing about the events that have actually occurred. In fact, Macpherson could claim—and indeed does claim—that, from a logical standpoint, he is "probably" right.

Nevertheless, there is a factor that lends support to the reasoning of classical inductionists, at least in engineering practice. Macpherson's law leads to impractical results. Let us suppose, for example, that we have concluded that 90% of the total load values that a system was required to sustain a few years ago fall between 100 and 120 MW [2]. According to Macpherson, we should certainly expect values below or above these limits. But how much above or below? And which limits should be considered? What is preferable—to risk failing to supply power, or to oversize the generation capacity in an irrational manner?

2 Determining Probability from Frequency data

In practice, we rarely possess data that allow for a direct evaluation of the probability that an event will occur. This probability must therefore be assessed based on frequency data. For example, suppose that in ten draws, one black ball is obtained. Let us then consider the probability p_2, such that the probability of the outcomes is p_1. Naturally, we would wish to choose p_1 so that p_2 is maximized.

Thus, p_2 will be equal to the ratio between the number of ways in which one can draw ten balls (one of which is black) and the number of ways in which one can draw ten balls, assuming that everything indicates the proportion of black balls is p_1. It is evident that in our case p_2 is maximal when $p_1 = 1/10$.

This is the principle of maximum likelihood, based on which we select hypotheses that lead to the most probable observation.

Having clarified the nature of the concept of probability, it now remains to show how the principle of maximum likelihood provides a method for obtaining a law on the basis of an adequate set of data, and how an appropriate form of that law can be determined.

2.1 Regression and Method of Least Squares

Let us consider a random variable Y_i, assumed to have a normal distribution with a mean μ, a variance δ^2, and (legitimately, in this case) let us suppose that $Y_i \sim N(\mu, \delta^2)$.

Furthermore, let us assume that the relationship is homoscedastic, meaning that its variance is independent of other variables (that is, δ^2 is independent of Y).

Given a set of samples (y_i, x_i), an estimate of μ and β can be obtained as indicated below.

The probability that

$$x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}[y_i - \mu - \beta(x_i - x)]^2\right\} \delta y_i \quad (1)$$

The likelihood L

$$L = \prod_{i=1}^{n} \left\{\frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2\sigma^2}[y_i - \mu - \beta(x_i - x)]^2\right\} \quad (2)$$

Assuming

$$\sum_1^n [y_i - \mu - \beta(x_i - x)]^2 \quad (3)$$

Thus, a law of adjustment for a sequence of data is obtained, and we are provided with a rational method for refining or fitting even the law itself [3].

2.2 Choosung an appropriate law

The postulate of simplicity, as defined by Jeffreys, is demonstrated in [23]. The demonstration provided does not show why only N_0 possible laws exist (although this has been omitted as something self-evident), nor why the highest probability is associated with the simplest laws. These points are discussed below.

Moreover, Jeffreys seems to exclude differential equations from the class of laws; however, since $N_0 + N_0 = N_0$, any finite number of other applicable schemes may exist, and in practice, one of this type is used in what follows.

We consider that the best proof of this postulate is experience itself.

We shall demonstrate that, at any level of knowledge, among two or more possible laws derived from the available data that relate a certain number of factors, the simplest law is the most probable.

First, let us determine how many laws are possible. The apparent answer is that their number is infinite. In other words, the possibility that a law might be represented by a high-order polynomial or a higher-order differential equation cannot be entirely

eliminated; nevertheless, no matter how improbable a law may be, it will always have a finite probability, however small.

We now face the following problem: there appears to be an infinite number of laws, each with a finite probability. This infinity of laws covers all possible probabilities — that is, the sum of all their probabilities equals unity (certainty). We then ask: which infinite series of finite numbers can have a finite sum?

A certain answer is given by the terms of a convergent infinite series, for example:

$$1 = \frac{1}{2} + \frac{1}{2} + \frac{1}{8} + \frac{1}{16} \ldots \quad (4)$$

The number of terms in such infinite series is N_0, the first and smallest **transfinite number**. Therefore, it is easy to show that there exist only N_0 possible laws and that all these laws have **rational probabilities**. The proof is carried out as follows.

The **sum of the probabilities** of all possible laws is finite (equal to unity). Consequently, the number of laws with any given probability $P$is less than $1/P$ and is finite if the law is possible (that is, if it has a finite probability). Likewise, the number of laws with a probability greater than P is also finite.

Arrange the probabilities in a **progression** [4] as follows: place first the greatest probability, and then arrange the rest so that they form a decreasing progression. Each element in this progression is a finite number, and therefore the progression is uniquely determined. Thus, the probabilities can be assigned a **rank** — 1, 2, 3, and so on — and their total number may be finite or N_0, that is, the number of all integers.

From the consideration of the transformation

$$y = \frac{z}{1-z^2} \quad (5)$$

it is clear that the above demonstration does not depend on whether we take the probability of certainty as 1 or as infinity.

Hence, we have shown that the probabilities of all possible laws can be arranged as an **ordered and convergent series**.

It remains to show why the **highest probability** in this series belongs to the **simplest law**. At a given level of our knowledge, we have at our disposal a finite number of data. This allows for the determination of a finite number of coefficients in a law. For example, two pairs of readings allow us to draw a straight line, while three readings make it possible to draw a quadratic curve or to verify a straight line.

Thus, assuming that we have learned from experience, our knowledge is such that at an initial stage the probability of a **simple law** is greater than that of a more complex one.

Let us now consider successive sets of verifying data. This does not affect the prior probability of a more complex law, $P(q/h) = 1$; that is, the initial quantities of data are not sufficient to determine the coefficients of more complex laws, and thus the probability of these data under the proposed variability of the law is certainly equal.

Therefore, if we exclude the hypothesis that the highest probabilities are associated with the simplest laws, the **posterior probabilities** of a more complex law would remain greater than those of the simpler ones (although the most complex law might remain unverified), when we have a finite number of data — since the sum of the series of probabilities must equal unity. But this conclusion contradicts the result of the previous paragraph.

These findings can also be interpreted as follows: the **series of complexity** must correspond **biuniquely** to the first term, since $N_0 - n = N_0$.

This demonstrates that the probabilities of possible laws that can be used for forecasting form a **convergent series**, with the **highest probability** being associated with the **simplest law**.

3 Conclusions

It can be shown that the above demonstration of the **postulate of simplicity** is subject to **Kronecker's critique**, which—if carried to its ultimate consequences—could undermine all of analysis, including infinitesimal calculus. Nevertheless, the criticism that can be raised here is that any rational probability can be associated with any law, and that the entire problem lies merely in reducing the number of possible laws from $N_0$to a number greater than a given one. Therefore, for **practical purposes**, such criticism is of no significance.

Moreover, the conclusions are consistent with the results of **regression analysis**: the more complex a law is, the greater the **dispersion of its regression coefficients**. Hence, it is found that the more complex the law, the **wider the forecast limits**. And precisely in order to obtain results that are as accurate as possible, we seek to choose **the simplest possible laws**.

References

[1] J.W Woods and H. Stark *Probability and Random Processes*, Prentice Hall, New Jersey, ISBN 0-13-178457-9, 2001.
[2] S.M Ross *A Firs Course in Probability*, Prentice Hall, New Jersey, ISBN 13-978-0-13-603313-4, 2002.
[3] S.M Ross *A Firs Course in Probability*, Academic Press, San Diego, ISBN 978-0-12-375686-2, 2002
[4] A. Populis and S.U. Pilai. *Probability, Random Variables and Stochastic Processes*, Mc Gaw Hill, New York, ISBN 0-07-366011-6, 2001

DOI: 10.2478/ijasitels-2025-0006

GRADIS: A Comprehensive System for Academic Contribution Tracking and Evaluation at Lucian Blaga University of Sibiu

Pitic Antoniu Gabriel[1][0000-0003-0104-5430],
Mihu Cantemir[1][0000-0002-4272-3990],
Pitic Elena Alina[1]

[1] *Lucian Blaga University of Sibiu*

Abstract

Academic institutions face increasing pressure to systematically track and evaluate faculty contributions across teaching, research, and administrative activities. This paper presents GRADIS (Grila de evaluare a Activitatii Didactice si Stiintifice - Grid for Evaluation of Teaching and Scientific Activity), a comprehensive web-based system developed for Lucian Blaga University of Sibiu (ULBS) to address these challenges. GRADIS provides a unified platform for managing academic contributions across multiple dimensions including scientific publications, teaching activities, student guidance, and administrative responsibilities. The system integrates with external academic databases such as Crossref, Web of Science, and Scopus to automate data collection and validation, while supporting flexible reporting and evaluation workflows tailored to Romanian academic standards. Built using Spring Boot and Vue.js, GRADIS demonstrates how modern web technologies can be effectively applied to academic administration, achieving significant improvements in data quality, process efficiency, and compliance with institutional evaluation frameworks.

Keywords: Academic evaluation, contribution tracking, research management, Spring Boot, Vue.js, educational technology

1 Introduction

1.1 Context and motivation

Academic institutions worldwide are increasingly required to demonstrate the quality and impact of their teaching and research activities. In Romania, universities must comply with evaluation standards set by ARACIS (Romanian Agency for Quality Assurance in Higher Education) and CNATDCU (National Council for Attestation of University Degrees, Diplomas, and Certificates). These regulatory bodies require comprehensive documentation of faculty contributions across multiple dimensions: scientific research, teaching excellence, student mentorship, and institutional service.
At Lucian Blaga University of Sibiu, this evaluation process historically relied on manual data collection, spreadsheet management, and paper-based documentation. Faculty members spent considerable time compiling their annual reports, often entering the same information multiple times across different forms. The administrative burden was substantial, data quality was inconsistent, and the validation process was time-consuming and error-prone.

1.2 Genesis of GRADIS

The GRADIS project emerged from collaborative discussions between ULBS administration and the Faculty of Engineering in 2022. Initial requirements gathering revealed several critical pain points:

- Data redundancy: Faculty entered the same publication or activity multiple times for different reporting purposes
- Validation bottlenecks: Validators struggled to verify hundreds of submissions against external databases
- Limited transparency: Faculty had restricted visibility into the evaluation process and scoring methodology
- Integration challenges: No connection existed between internal systems and external academic databases
- Compliance complexity: Mapping contributions to specific GRADIS indicators (the standardized ULBS academic evaluation framework) required extensive manual effort

A development team was assembled comprising faculty members with software engineering expertise and external consultants. The decision was made to build a custom solution rather than adopt an existing commercial product, primarily because:

- Commercial solutions were prohibitively expensive for a mid-sized Romanian university
- Off-the-shelf products did not support the specific GRADIS indicator framework required by ULBS regulations
- Custom development would allow tight integration with existing ULBS infrastructure
- The project could serve as a learning opportunity for computer science students

Development began in early 2023, with the first production deployment occurring in September 2023. The system has since been actively tested by over 50 faculty members across multiple departments. However, users with imported or added data exceed 200.

1.3 Scope and objectives

GRADIS was designed with the following primary objectives:

- Centralized contribution management: Provide a single system of record for all faculty academic contributions
- Automated data enrichment: Integrate with external databases to automatically fetch publication metadata and citation counts
- Transparent evaluation: Implement clear, auditable scoring based on GRADIS indicators
- Efficient validation: Enable validators to quickly review and approve faculty submissions with supporting evidence
- Comprehensive reporting: Generate reports required by various stakeholders (faculty, departments, university administration, accreditation agencies)
- User-friendly interface: Minimize the learning curve and maximize adoption through intuitive design

This paper describes the business requirements, architectural decisions, technical implementation, and lessons learned from developing and deploying GRADIS at ULBS.

1.4 Similar applications

In the broader landscape of research information management and academic performance evaluation, GRADIS can be situated alongside several institutional systems developed at other universities, which illustrate different design choices and organizational priorities. At the University of Kent, the KRIMSON (Kent Research and Innovation System Online) platform represents a full-scale Current Research Information System (CRIS) built on top of the commercial Converis v5 product. McDonnell and Kerridge describe KRIMSON as an integrated research information management solution intended to consolidate data on projects, publications, funding, and outputs into a single authoritative source, streamline REF-related reporting, and support both managerial decision-making and compliance with external evaluation exercises [1]. While GRADIS and KRIMSON share the goal of providing a unified, institution-wide record of research activity, their design trajectories diverge: GRADIS has been developed in-house, tailored to the specific GRADIS indicator framework and Romanian ARACIS/CNATDCU requirements, whereas KRIMSON is a local configuration of a generic CRIS platform; this gives GRADIS more flexibility in encoding national evaluation rules, but KRIMSON benefits from the vendor's mature product ecosystem and broader interoperability within the CERIF/CRIS community.

A second relevant example is the research information management system implemented at the National University of Colombia – Manizales Campus. Bermon Angarita and Prieto Taborda describe a web-based system conceived as a knowledge repository for the Directorate of Research and Extension, designed to capture projects, research products, and associated documentation and to support institutional reporting to MinCiencias (the national science and technology council) [2]. Their system emphasizes knowledge-management principles (knowledge repositories, reusable metadata structures, and support for group-level evaluation of research units) and was validated in the context of a national call for measuring and recognizing research groups [3]. Compared with GRADIS, which operates at the granularity of individual faculty members and GRADIS indicators across teaching, research, and service, the Manizales system is more narrowly focused on research groups and project outputs; it does not appear to cover teaching activities, student supervision, or administrative roles, nor does it integrate directly with external bibliographic APIs such as Crossref, Web of Science, or Scopus. However, both systems share the goal of reducing manual reporting effort, improving data quality, and enabling more robust institutional reporting based on structured, centrally managed data.

Beyond research information management in the strict sense, several universities have built or adopted systems that concentrate on academic staff evaluation, often with an emphasis on quantitative and multi-criteria scoring. At Palacký University in Olomouc, the IS HAP (Information System for Academic Staff Performance Evaluation) is used as a core tool for human-resources quality management and institutional accreditation. Public documentation and related research reports describe IS HAP as an information system whose core is a multiple-criteria evaluation model based on linguistic fuzzy rule bases, used to aggregate various performance indicators into an overall staff evaluation [4]. The system supports the collection of data on teaching, research, and other activities, and has been extended over time as part of broader reform of research assessment at the university [5]. Conceptually, IS HAP is closer to the evaluation engine inside GRADIS than to its data-collection workflows: both systems encode institution-

specific scoring models and aim to make academic staff evaluation more transparent and consistent, but IS HAP is framed primarily as an HR/quality-assurance tool, whereas GRADIS also provides rich operational functionality for contribution entry, evidence upload, duplicate detection, and integration with external scholarly databases.
An additional class of related systems focuses on building centralized faculty information repositories, usually to support accreditation, promotion, and tenure processes, and external visibility rather than detailed indicator-based scoring. De La Salle University–Manila, for example, implemented a web-based faculty information system to replace email-based submissions of faculty activity forms. Franco's case study describes the system as a central repository of faculty-related accreditation requirements, covering personal data, teaching history, research outputs, and outreach engagements, with role-based access for faculty and administrators [6]. The emphasis there is on data completeness, reliability, and availability for accreditation audits, rather than on sophisticated evaluation logic. A similar orientation can be observed in generic Faculty Information Systems (FIS) deployed at universities such as the University of Arkansas, where FIS is described as a web-based solution allowing faculty to view, enter, and track teaching, research, and service activities in a central repository, streamlining annual reviews and personnel processes [7]. In these implementations, the system acts as a structured CV and reporting hub, while evaluation rules are often applied downstream by separate HR or promotion committees.
When these systems are viewed together, several comparative dimensions emerge. In terms of scope, KRIMSON and the Manizales research information system primarily target research inputs and outputs (projects, grants, publications, research products), with limited explicit support for teaching and administrative contributions, whereas GRADIS and IS HAP explicitly include teaching, supervision, and service indicators as first-class entities in their models. From an architectural standpoint, GRADIS and the Colombian system share a custom, in-house development approach, while KRIMSON is based on a configurable commercial CRIS (Converis) and many faculty information systems leverage vendor platforms such as Interfolio or similar enterprise solutions. Regarding evaluation models, IS HAP stands out through its fuzzy multi-criteria approach, whereas GRADIS uses a rule-based mapping to nationally defined GRADIS indicators with point weights aligned to ARACIS and CNATDCU standards; KRIMSON and the faculty information systems tend to focus more on data capture and reporting, leaving actual scoring schemes to external processes such as REF or internal academic committees. Finally, integration capabilities differ considerably: KRIMSON and GRADIS both emphasize interoperability with external scholarly information sources (e.g., Crossref, Web of Science, Scopus), while the Manizales system and many faculty information systems emphasize internal integration with institutional repositories and administrative systems.

Table 1. Comparative table

System	University	Scope	Evaluation	Technology	Integrations
GRADIS	ULBS	Research, teaching, supervision	Rule-based, GRADIS indicators	Spring Boot + Vue.js	Crossref, WoS, Scopus
KRIMSON	University of Kent (UK)	Research & innovation	REF-aligned (external)	Converis CRIS	CERIF/CRIS ecosystem
Manizales RIMS	National University of Colombia	Research groups	National MinCiencias framework	Custom web KM system	Internal focus
IS HAP	Palacký University (CZ)	Staff performance	Fuzzy multi-criteria	Institutional IS	Internal QA
FIS DLSU	De La Salle University (PH)	Faculty accreditation & data	Review-based, external committees	Web app	Internal admin systems

2 Business needs and requirements

2.1 The ULBS academic evaluation framework

Romanian higher education operates under a specific evaluation framework that categorizes academic contributions into numbered indicators and the ULBS's GRADIS comply with it. These indicators are divided into two main categories [8]:

- Scientific Activity Indicators (IC - Indicatori Cercetare): - IC01-IC17: Research publications, citations, patents, conference presentations, artistic activities, and sports achievements
- Teaching and Administrative Activity Indicators (ID - Indicatori Didactici): - ID01-ID18: Teaching materials, student guidance, program coordination, committee memberships, and other academic service

Each indicator has specific criteria and scoring rules. For example: - IC01: Articles published in Web of Science indexed journals with impact factor - IC04: Citations in Web of Science or Scopus databases - ID11: Coordination of bachelor's or master's degree final projects - ID13: Scientific guidance of doctoral students

Faculty members must accumulate a minimum number of points across both categories to qualify for academic advancement (lecturer to associate professor, or associate professor to full professor). Different faculties and disciplines may have varying point requirements.

2.2 User Roles and Workflows

GRADIS supports four primary user roles, each with distinct needs.

2.2.1 Professors

Professors are the primary data contributors. Their primary workflows are described in table 2.

Table 2. The professor role

Workflow	Description
Authentication	Login via institutional Google OAuth2 (all ULBS faculty have @ulbsibiu.ro email addresses)
Contribution entry	Add publications, teaching activities, or other contributions • Manual entry with form-based input • Bulk import from BibTeX files for publications • Import via DOI lookup (automatic metadata retrieval from Crossref)
Author management	Associate internal colleagues and external collaborators with contributions
Evidence upload	Attach proof documents (PDFs, certificates, official letters)
Report generation	Create annual reports for specific evaluation periods (calendar year for scientific indicators, academic year for teaching indicators)
Status monitoring	Track validation status of submitted reports

2.2.2 Secretaries

Department secretaries could assist with data management and administrative tasks. However, in the application the role includes basic user management, and the administration of important journals and publishers. Table 3 shows Some of the workflows:

Table 3. The secretary role

Workflow	Description
Bulk operations	Import multiple contributions on behalf of faculty
Report coordination	Ensure all department faculty submit required reports by deadlines
Document management	Organize and archive supporting documentation
Edit journals and publishers	Maintain lists of journals, publishers, and other master data

2.2.3 Validators

The role can be at the department, faculty, or university quality department level. Validators ensure quality and compliance (Table 4).

Table 4. The validator role

Workflow	Description
Review queue	Access lists of pending reports requiring validation
Evidence examination	Review uploaded proof documents
Citation validation	Verify citation counts against Web of Science, Scopus, or Google Scholar
Approval/rejection	Check that contributions meet criteria for claimed indicators. Mark reports as VALID, INVALID, or request additional information
Batch operations	Process multiple similar contributions efficiently

2.2.4 Administrators

System administrators manage high level permissions user accounts and system configuration (Table 5).

Table 5. The administrator role

Workflow	Description
CRUD operations on users	Manage the secretary and administrator users.
System monitoring	Track usage, performance, and error logs
Import journals	The UEFISCDI institution publishes yearly a list of WOS Journals which must be imported in the application. ERIH Plus, Scopus and prestigious publisher will also be imported.

2.3 Functional requirements

Based on stakeholder interviews and regulatory analysis, the following functional requirements were identified:

2.3.1 Contribution management

- Support for 24 contribution types (articles, books, patents, conferences, projects, teaching activities, etc.)
- Each contribution type must capture type-specific metadata (e.g., journal impact factor for articles, ISBN for books)
- Version history and audit trails for all changes
- Soft delete functionality (mark as deleted but retain in database for audit purposes)

2.3.2 Author management

- Distinction between internal authors (ULBS faculty) and external authors

- Support for author order and corresponding author designation
- Affiliation tracking (especially important for multi-institutional collaborations)
- ORCID integration for unique author identification

2.3.3 Data integration

- Crossref API: Retrieve publication metadata by DOI
- Web of Science API: Fetch citation counts and article details
- Scopus API: Alternative citation source (especially important for engineering and computer science)
- Google Scholar: Citation tracking for works not indexed in WoS/Scopus

2.3.4 Reporting and validation

- Annual report generation with automatic point calculation per GRADIS indicator
- Multi-stage validation workflow (unverified → valid/invalid)
- Commenting system for validators to request clarifications
- Excel export functionality for external reporting to accreditation bodies

2.3.5 Evidence management

- File upload for proof documents (PDFs, images)
- Multiple files per contribution (article PDF, acceptance letter, certificate, etc.)
- File size limits (10MB per file) and format restrictions
- Secure storage with access control

2.4 Non-Functional requirements

2.4.1 Security and privacy

- Authentication: OAuth2 integration with institutional Google Workspace
- Authorization: Role-based access control (RBAC)
- Data protection: Faculty can only view their own data unless granted specific permissions
- Audit logging: All data modifications tracked with user and timestamp
- GDPR compliance: Ability to export and delete personal data upon request

2.4.2 Performance and scalability

- Support 300+ concurrent users during peak periods (report submission deadlines)
- Response time < 2 seconds for typical page loads
- Database queries optimized for large datasets (10,000+ contributions)
- Batch operations for validators processing dozens of reports

2.4.3 Availability and reliability

- 99% uptime during academic year (September-July)
- Daily automated backups
- Disaster recovery procedures
- Graceful degradation when external APIs are unavailable

2.4.4 Usability

- Intuitive interface requiring minimal training
- Consistent design language across all modules
- Responsive design supporting desktop and tablet devices
- Romanian and English language support

3 System architecture

3.1 High-Level architecture

GRADIS follows a three-tier architecture pattern common in modern web applications (Figure 1).

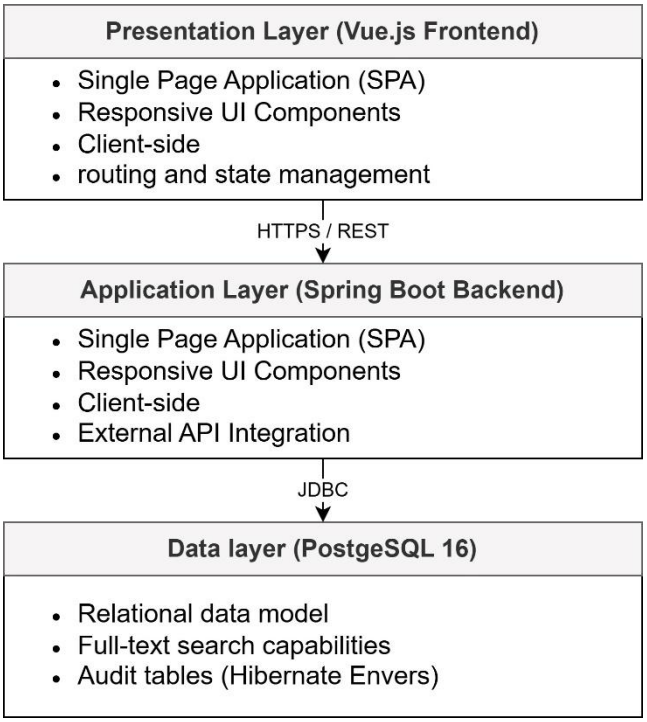


Figure 1 – High level architecture

The system architecture is organized into three primary layers: Presentation Layer, Application Layer, and Data Layer, each responsible for distinct logical and operational concerns.

The **Presentation Layer**, implemented using a Vue.js Single Page Application (SPA), delivers a modern, responsive user interface composed of reusable components and

client-side routing supported by state management mechanisms. Communication between the client and the server is conducted securely via HTTPS/REST endpoints.
The **Application Layer**, developed using the Spring Boot framework, exposes REST API controllers that process incoming requests and return standardized responses. Core institutional rules and workflows are encapsulated within dedicated business logic services, while system security is enforced through robust authentication and authorization mechanisms. Additionally, this tier manages external API integrations for data synchronization, enrichment, and validation.
The **Data Layer**, implemented using PostgreSQL 16, provides the underlying persistent storage through a relational data model. It incorporates full-text search capabilities to support advanced query scenarios and employs audit logging via Hibernate Envers to ensure traceability, historical record management, and compliance with accreditation and reporting requirements.
The GRADIS backend integrates with three external scholarly data providers (Crossref, Web of Science, and Scopus) to support automated metadata retrieval, citation synchronization, and publication verification (Figure 2). These services are accessed via dedicated API client modules implemented in the backend. Each external service communicates directly with the GRADIS application, which aggregates, validates, and harmonizes retrieved metadata before storing it in the institutional database.

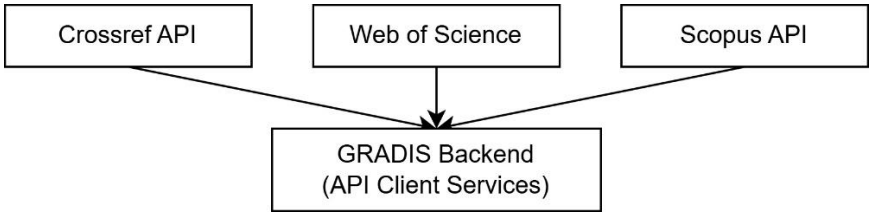


Figure 2 – External integrations

3.2 Technology stack

3.2.1 Backend (gradis-backend)

Core framework
The backend of the GRADIS system is built on Java 21, a modern Long-Term Support (LTS) release that offers improved performance, enhanced memory management, and advanced language constructs beneficial for large-scale enterprise applications. The application is developed using Spring Boot 3.4.5, which accelerates development through auto-configuration, opinionated defaults, and an embedded server architecture that simplifies deployment across environments. Data persistence is handled through Spring Data JPA, which provides an abstraction layer based on the repository pattern and enables clean, maintainable data-access logic. Underlying the persistence layer, Hibernate 6.2 serves as the JPA implementation, offering advanced query optimization and automatic dirty-checking mechanisms. For auditability requirements, Hibernate Envers is integrated to provide automatic entity versioning and complete change history tracking at the database level.

Database technologies
Data storage and transactional operations are managed using PostgreSQL 16, chosen for its robustness, ACID-compliance, and advanced features such as native JSON handling and built-in full-text search capabilities, which are useful in publication-metadata indexing scenarios. Database schema evolution is controlled through Flyway, enabling reliable versioning, reproducible migrations, and environment-agnostic deployment pipelines.

Security infrastructure
Security is implemented using Spring Security, which provides comprehensive authentication, authorization, and request-filtering capabilities. Integration with institutional Google Workspace accounts is achieved via OAuth 2.0 Client support, enabling secure single sign-on (SSO) without storing local user credentials. For distributed deployments and session replication, server-side session persistence is configured through JDBC-based session management, which supports horizontal scalability.

API integration layer
External data acquisition and metadata synchronization are handled via HTTP communication components based on Spring RestTemplate and WebClient. The system integrates with scholarly databases and indexing services using a combination of official and custom-developed clients, including the Clarivate Web of Science SDK and bespoke connectors for Crossref and Scopus, enabling automated retrieval, parsing, and enrichment of bibliographic data.

Document processing and reporting
The GRADIS system supports document import/export workflows through Apache POI, enabling Excel-based bulk operations for contribution submission and validation. Printable academic reports and summaries are produced using iText PDF, which provides fine-grained control over document layout and typography. For structured publication import, JBibTeX is integrated to parse BibTeX sources and map them to internal contribution entities.

Development and DevOps tooling
The system build process and dependency management rely on Gradle 8.8, providing efficient incremental builds and Kotlin-DSL extensibility. To reduce boilerplate code and improve maintainability, Lombok is used to generate accessor methods, constructors, and data-transfer scaffolding at compile time. REST API documentation is automatically generated using SpringDoc OpenAPI, enabling user-friendly inspection via Swagger UI. Application health, metrics, and runtime visibility are provided through Spring Boot Admin, supporting proactive monitoring and operational management.

3.2.2 Frontend (gradis-frontend-vuejs)

Frontend core framework
The GRADIS frontend is developed using Vue.js 3, a modern progressive JavaScript framework that leverages the Composition API to improve component reusability,

maintainability, and type-safety. Application development and bundling are handled by Vite, which provides fast hot-module reloading and optimized build performance suitable for large-scale SPAs. Client-side state is managed using Pinia, the successor to Vuex, which offers a simpler and more modular store architecture. Routing and view transitions across the single-page interface are implemented with Vue Router, allowing dynamic navigation without full-page reloads.

UI components and visualization
The user interface integrates Vuetify, a feature-rich component library that provides a wide range of pre-built elements such as data tables, form controls, wizards, and dialog windows, allowing for rapid UI development with consistent visual standards. Layout styling and responsive behaviour are also supported. For dashboard-level analytics and graphical reporting, the system incorporates Chart.js, enabling interactive chart visualizations directly within the SPA.

HTTP communication layer
Communication between the SPA and the backend server is implemented using Axios, a promise-based HTTP client that supports request/response interception, automatic token injection, and centralized error handling. Custom API interceptors are used to ensure that authentication and session-handling logic remain transparent to the UI layer, reducing code duplication and improving reliability and security across all HTTP requests.

Development and build tools
The development environment is powered by Node.js 18, which serves as the runtime for build tooling, script execution, and dependency resolution. Package management is handled through npm, while ESLint is used to enforce coding standards and maintain a consistent project style. TypeScript is being gradually integrated to introduce optional static typing, improving refactoring safety and enabling more robust IDE assistance.

Key frontend features
The GRADIS frontend offers several advanced interface features, including customizable form-validation mechanisms, drag-and-drop file upload components, and responsive grid-based layouts optimized for both desktop and tablet usage. The system currently supports near real-time content updates via short-interval polling and is planned to transition to WebSocket-based live updates in a future release. To support multilingual usage and international collaboration, the application includes internationalization (i18n) features with current availability in Romanian and English.

3.3 Data model

The GRADIS data model focuses on three main entities.

3.3.1 Contribution entity hierarchy

The chosen data-model design is based on single-table inheritance (Figure 3), which enables polymorphic behavior across all contribution types and allows unified querying without additional joins or complex discriminator logic. This structure also simplifies

filtering, aggregation, and reporting operations across heterogeneous contribution records. To optimize data retrieval, author information is eager-loaded by default, since this represents the dominant access pattern and helps prevent N+1 query overhead during listing, search, and validation workflows. Contribution records employ soft-delete semantics, ensuring that entries removed from active use remain available for historical inspection, audit-trail preservation, and compliance-driven traceability. Finally, the implementation defines abstract methods within the shared base class to enforce uniform DTO conversion and guarantee that all contribution types provide a consistent and extensible serialization contract.

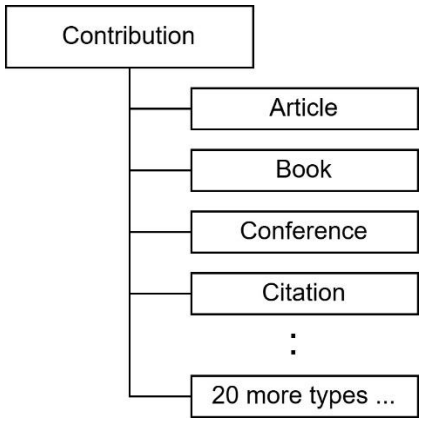


Figure 3 – Contribution entity hierarchy

3.3.2 Report entity

Reports link contributions to specific users and evaluation periods. The data model supports several key relationships that reflect real academic reporting workflows. A single contribution may be associated with multiple reports, since the same publication or activity can be claimed by different co-authors or may be relevant across multiple reporting periods, such as annual evaluations. Each user maintains multiple reports over time, forming a cumulative academic activity history that grows with each evaluation cycle rather than replacing previous submissions. In addition, report entities are fully audited, allowing the system to track changes in validation status, assigned points, evaluator remarks, and system-generated recalculations, thereby preserving transparency, accountability, and compliance with institutional and accreditation standards.

3.3.3 User and authorization

The system implements a role-based authorization model in which each user may hold multiple roles through a dedicated UserRole join entity, allowing flexible assignment and combination of responsibilities. The available roles include PROFESSOR, SECRETARY, VALIDATOR, and ADMIN, each associated with distinct permission sets and operational scopes. To support contextual and time-bounded authorization, role assignments can be further scoped to specific departments and validity periods using the UserRoleDepartmentPeriod structure, enabling fine-grained access control that

reflects real institutional governance, organizational hierarchies, and rotating academic responsibilities.

3.3.4 Supporting entities

The Journal entity is used to store bibliographic and indexing metadata, including annual impact factor values, which enables longitudinal citation-impact analysis and automated scoring rules. The ExternalAuthor entity represents contributors who are not affiliated with ULBS, allowing the system to handle mixed-affiliation publications without compromising institutional reporting accuracy. Authorship relationships are modeled using ContributionInternalAuthor and ContributionExternalAuthor join entities, both of which represent many-to-many associations enriched with additional attributes such as author order, institutional affiliation, and corresponding-author status, ensuring accurate representation of authorship semantics and evaluation-relevant distinctions.

3.4 Deployment architecture

3.4.1 Production environment (ULBS on-premises)

The GRADIS system runs in the ULBS on-premises production environment, hosted on university-managed servers that provide controlled access, secure data handling, and compliance with institutional IT and academic data-protection policies (Figure 4).

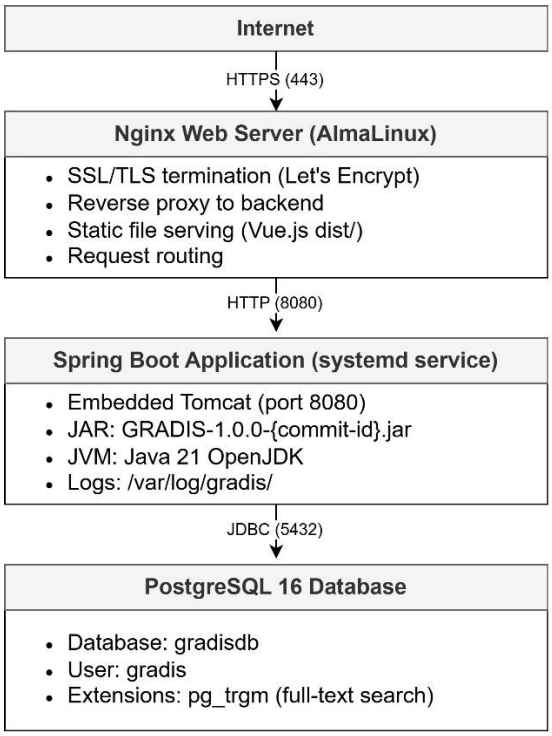


Figure 4 – Production environment

3.4.2 Deployment process

The deployment workflow is fully automated through a GitLab-based CI/CD process. When new code is pushed to the ulbs branch, the pipeline is triggered and executed by a GitLab Runner installed on the application server. The deployment script retrieves the latest codebase, compiles the backend using Gradle to produce the executable JAR, and builds the frontend using npm run build inside the Vue.js project directory. The existing gradis.service systemd process is then cleanly stopped, after which the new JAR file is placed in /opt/gradis/ and the newly generated dist/ directory from the frontend build is copied to the Nginx web-serving directory. Once the artifacts are replaced, the gradis.service is restarted, and the system performs an automated health check against the /actuator/health endpoint to confirm successful startup. In case of any failure during the process, an automated email notification is sent to the administrative team to initiate manual intervention.

4 Conclusions and future work

4.1 Achievements and impact

The GRADIS system has been available at ULBS since September 2023, where it is currently deployed in a controlled institutional beta-testing and validation phase. During this period, it has reached institution-wide adoption levels and has demonstrated clear operational and strategic value, while continuing to undergo iterative refinement in alignment with stakeholder feedback and institutional requirements. Since deployment on university servers, measurable usage data confirms both engagement and efficiency gains. As of the testing phase, the platform has been actively used by faculty members across eight faculties, resulting in a recorded total of 3,847 contributions, including 2,156 journal articles, 487 books, 341 conference outputs, and 863 teaching or administrative activities. Notably, 67% of the indexed research publications were imported automatically using DOI or BibTeX metadata, significantly reducing manual data-entry effort. The system generated a total of 856 annual evaluation reports during the 2023–2024 assessment period.

Beyond quantitative indicators, several qualitative benefits have also been observed. Automatic metadata retrieval, including citation data from trusted sources such as Web of Science and Scopus, has improved data accuracy and completeness. Moreover, the platform increases transparency by enabling faculty members to visualize how their contributions are scored against GRADIS indicators and institutional evaluation rules. From an institutional compliance perspective, complete digital audit trails support external accreditation requirements, while improved visibility into research activity has also facilitated collaboration by allowing users to identify colleagues with overlapping research interests.

Stakeholder feedback further confirms these benefits. Faculty users particularly value the automated publication import functionality, describing the ability to register a publication "in approximately 30 seconds instead of 10 minutes." Validators highlighted that the duplicate-detection feature prevented at least 50 redundant submissions during the first semester of usage, reducing both validation workload and system noise. Administrative offices also reported major efficiency gains, noting that

evaluation and accreditation reports required by bodies such as ARACIS can now be produced in minutes rather than days.

4.2 Future work and improvements

4.2.1 Short-term enhancements (next 6 Months)

In the short term, development efforts will focus on usability improvements, real-time communication features, data analytics capabilities, and expansion of reporting options. A priority is the introduction of real-time notifications, which will rely on WebSocket-based event delivery to inform users of validation status updates, report approvals, and approaching deadlines. These will be complemented by institutional email alerts and a mobile-friendly notification dashboard to ensure timely awareness regardless of device.

A second enhancement will target advanced analytics and visualization features, allowing departments to explore publication and contribution trends, monitor citation impact longitudinally, and compare anonymized performance indicators across units. In parallel, mobile accessibility will be strengthened through improvements to responsive design, and the development of a Progressive Web App (PWA) to enable partial offline usage. Finally, the reporting component will be expanded to include customizable templates aligned with the requirements of different accreditation bodies, automated report generation for program-level evaluations, and exports to formats beyond PDF and spreadsheet outputs, including Word and LaTeX.

A new capability will also be introduced to define and maintain point-calculation rules for contribution types using Google CEL (Common Expression Language) expressions. This will allow administrators to configure scoring logic declaratively, without modifying the application code, and adapt the evaluation model quickly in response to institutional or national policy changes.

4.2.2 Medium-term goals (6–12 Months)

Over the medium term, efforts will shift toward intelligent automation, deeper interoperability with external scholarly ecosystems, and collaborative validation workflows. Planned AI-based enhancements include natural-language processing to classify contributions automatically within the GRADIS indicator structure, semantic-level duplicate detection to improve accuracy beyond string matching, and a conversational assistant capable of explaining evaluation rules and criteria to faculty members. In addition, integration with external research identity and indexing systems is a priority, particularly through ORCID-based publication synchronization, Google Scholar import capabilities, and interoperability with the institutional DSpace repository.

Collaboration-oriented enhancements are also targeted, including co-author confirmation workflows for shared publications, pre-validation departmental review mechanisms, and support for inline comments and annotations on reported contributions. Finally, performance optimizations are planned through the introduction of Redis caching for frequently accessed data, systematic database indexing and query improvements, and frontend optimization strategies such as lazy loading and code splitting.

4.2.3 Long-term vision (1–2 Years)

From a long-term strategic perspective, the objective is to position the system as a scalable, flexible platform capable of supporting multiple institutions, international evaluation models, and research-oriented analytics. This includes developing a multi-tenancy architecture to allow configuration for other Romanian universities and potentially for institutions using evaluation frameworks distinct from GRADIS. A gradual open-source release is also envisioned, which will require removal of proprietary components, development of comprehensive technical documentation, and publication of community contribution guidelines, potentially followed by listing the project in recognized academic open-source repositories.
In addition, the long-term roadmap includes the incorporation of advanced research impact analytics, not only through altmetrics and collaboration-network visualization, but also through predictive models capable of estimating research trajectories and identifying potential institutional strengths.

Acknowledgment
This paper reflects the collaborative effort of many individuals committed to advancing academic management excellence at Lucian Blaga University of Sibiu.
We extend our deepest appreciation to Andrada Pitic, whose pioneering work in requirements analysis, user experience design, and initial development established the foundation upon which GRADIS was built.
We gratefully acknowledge the partnership between Lucian Blaga University of Sibiu and NTT DATA Romania S.A., which provided the professional development expertise and resources essential to transforming academic requirements into production-ready software.
Special thanks to the faculty members, validators, and administrative staff who participated in pilot programs, provided invaluable feedback, and demonstrated patience during the inevitable challenges of organizational change. Their willingness to embrace new approaches and contribute to continuous improvement has been fundamental to GRADIS's success.
We acknowledge the university administration for their vision in supporting digital transformation initiatives and their commitment to evidence-based management practices that benefit the entire academic community.

References

[1] R. McDonnell & S. Kerridge, *Research Information Management System (KRIMSON) at Kent*", Procedia Computer Science, Vol. 106, pp. 160–167, 2017. DOI: 10.1016/j.procs.2017.03.012. Available: https://www.sciencedirect.com/science/article/pii/S1877050917302806
[2] L. Bermon Angarita & M. A. Prieto Taborda, *Development of a research information management system based on knowledge repositories: Case of the Directorate of Research and Extension at the National University of Colombia – Manizales Campus*, RCTA, 2025. Available: https://ojs.unipamplona.edu.co/index.php/rcta/article/download/3129/7674/16799
[3] MinCiencias (Colombia), *National Research-group evaluation framework documentation*. https://researchonresearch.org/observatory/colombia/
[4] J. Talašová et al., *Information System for Academic Staff Performance Evaluation (IS HAP)*, Palacký University, Olomouc. Available: https://fuzzymcdm.upol.cz/ishap.html
[5] Palacký University Olomouc, *Quality Assurance: Evaluation models and IS HAP reform*. Available: https://www.upol.cz/en/university/quality-assurance/

[6] G. R. L. Franco, *Design and Implementation of a Web-Based Faculty Information System*, 2016, IEEE Region 10 Conference. https://ieeexplore.ieee.org/document/7848535
[7] University of Arkansas, *Faculty Information System (FIS) documentation*. Available: https://its.uark.edu/administrative-services/faculty-information-systems/
[8] https://senat.ulbsibiu.ro/wp-content/uploads/250925/30_Grila%20evaluare%20GRADIS_Senat_sept2025.pdf

DOI: 10.2478/ijasitels-2025-0010

IoT System for Milk Dispenser Monitoring and Management

Claudia Banciu, Teodora-Artemisa Menchiu, Adrian Florea
Department of Computer Science and Electrical Engineering, "Lucian Blaga" University of Sibiu, Sibiu, Romania
Corresponding author: Claudia Banciu, e-mail: claudia.banciu@ulbsibiu.ro
Teodora Menchiu, e-mail: teodora.menchiu@ulbsibiu.ro
Adrian Florea, e-mail: adrian.florea@ulbsibiu.ro

Abstract
An Internet of Things (IoT)-like application is an advanced system for automation and data analytics that leverages communication networks, sensors, and artificial intelligence to provide a comprehensive solution for a product or service. The development of IoT technologies has facilitated advancements in automated systems across various industries like healthcare, agriculture, smart cities, smart homes, and transportation. This paper presents an IoT-based system designed for the real-time monitoring and management of a milk dispenser. The system integrates an ESP32 microcontroller, sensors for temperature and milk level measurement, and cloud storage for data analysis and visualization. By using machine learning algorithms, the system predicts milk consumption trends and detects potential dispenser blockages, enhancing operational efficiency. Experimental evaluation showed that the Random Forest and Gradient Boosting regression models achieved high predictive accuracy, with the best configurations reaching a coefficient of determination (R2) of 0.95, a Mean Absolute Error (MAE) as low as 1.05 liters, and a Root Mean Squared Error (RMSE) of 1.32 liters, confirming the strong reliability of the proposed approach.

Keywords: IoT system, milk-dispenser, machine learning algorithms

1 Introduction

Milk dispensers are specialized machines that are designed to store and dispense milk efficiently while maintaining optimal freshness. Despite their practicality, there are also challenges in managing milk dispensers. Some challenges are ensuring consistent milk availability, maintaining the correct storage temperature, avoiding blockages or any other failures.

Beyond their basic functionality, milk dispensers bring several practical and economic benefits that contribute to their growing adoption. First, they significantly reduce packaging waste, eliminating the need for single-use cartons or bottles, which makes them a more environmentally sustainable solution. Second, dispensers support cost efficiency, since purchasing milk in bulk is generally cheaper for institutions and businesses compared to buying multiple individual containers. Third, they increase convenience by providing quick, portion-controlled access to milk, which is particularly valuable in high-consumption environments such cafeterias, farmers' markets or grocery stores. Dispensers ensure freshness and hygiene, as they maintain

milk at a safe chilled temperature and reduce the risk of contamination that could occur when repeatedly opening bottles or cartons.

1.1. Challenges in Milk Dispenser Adoption

Despite their advantages, milk dispensers also face several challenges that can limit widespread adoption. First, the initial cost of purchasing and installing the equipment can be relatively high. However, in our study, this limitation is less significant, as the system is already installed and provides real operational data for analysis. Second, dispensers require regular maintenance and cleaning to prevent bacterial growth and ensure hygiene standards, which can be time-consuming. Third, they involve continuous electricity use, since the milk must remain chilled at a constant temperature, resulting in additional energy costs. Finally, dispensers have limited portability compared to bottles or cartons, as their size and bulk make them unsuitable for milk transport outside of fixed locations.

1.2. Market Context and Adoption Trends

The adoption of milk dispensers is steadily increasing both in Europe and worldwide. According to [1] the global market for milk vending machines was valued at approximately USD 0.9 billion in 2023 and is projected to nearly double to USD 1.8 billion by 2032, with a compound annual growth rate (CAGR) of around 8.1%. Europe represents the largest regional market, driven by consumer demand for fresh milk and strong traditions in dairy farming. Countries such as Germany, Italy, and France already report widespread use of milk dispensers in supermarkets, schools, and farmers' markets.
In Romania, milk dispensers are also present, though their adoption is more recent. The Romanian company FERMAT has reported the sale of more than 1000 milk vending machines across the country. Furthermore, consumer studies indicate that Romanian buyers associate these machines with product freshness, sustainability (through reusable bottles), and direct access to local dairy producers.

1.3. IoT and milk dispensers

The IoT presents new possibilities for automating and optimizing these systems. An IoT system typically consists of three essential layers: the sensor layer (physical objects), the connectivity layer (networks such as Wi-Fi), and the cloud layer (data storage and algorithms). IoT enables the integration of sensors and microcontrollers to monitor milk levels and temperature in real-time, offering predictive tools for maintenance. The proposed system employs an ESP32 microcontroller, temperature and weight sensors, and cloud-based data analytics to improve operational efficiency and enhance user experience. The appearance of the IoT has opened new possibilities for optimizing and automating these devices. By leveraging IoT, sensors and microcontrollers can track milk levels and temperatures, while cloud platforms provide users with real-time data access. This connectivity aids in predictive maintenance, reduces downtime, and ensures a smooth user experience.

Through smart devices and data analysis, IoT makes it possible to monitor milk levels, track temperature, and even predict issue. Incorporating IoT into milk dispensers provides several significant advantages. For instance, real-time monitoring of milk levels and temperature ensures that the milk remains fresh and available when needed. Cloud platforms offer centralized data storage and analysis, enabling users to remotely access data through intuitive interfaces. Additionally, IoT systems can use machine learning algorithms to predict milk consumption patterns and detect potential issues, such as blockages, before they disrupt service.

This paper aims to present an IoT-based solution for milk dispenser management. The proposed system integrates an ESP32 microcontroller, temperature and weight sensors, and cloud-based analytics to offer real-time insights. By addressing common challenges and introducing predictive capabilities, the system seeks to improve the efficiency, reliability, and usability of milk dispensers.

The rest of this paper is structured as follows. Section 2 reviews the state of the art and related work in IoT-enabled applications targeting different fields such as agriculture, food industry and dairy supply chain. Section 3 describes the methodology used, presenting the hardware and software architecture of the IoT system. Section 4 presents and compares results obtained by different prediction techniques. Section 5 highlights the conclusions and presents future research directions.

2 Related Work

Over recent times, the adoption of this kind of technologies has become increasingly prevalent in agriculture and food monitoring systems, and many other fields like healthcare, smart homes, and even smart cities. There are numerous studies which emphasize the critical role that the intelligent monitoring and automation have in providing efficiency, sustainability, also enabling the real-time decision-making.

In [2] the authors provide a comprehensive review of Agriculture 4.0, highlighting the importance of integrating the IoT, the artificial intelligence, and automation in modern farming practices. The authors highlight several key benefits including optimizing the resource management, increasing the crop yields, monitoring the soil and weather in real time or even reducing the costs of labor.

Similarly, the authors of [3] demonstrate how IoT-enabled smart agriculture platforms improve crop yield and environmental monitoring through the deployment of sensor networks. In this study the authors discussed about the importance of using the WSNs (wireless sensor networks) which can advise the farmers about different aspects like the status of the agricultural machinery and about all the factors that can influence the crop growth. They also highlighted the significance of using machine learning algorithms which can help to detect or even prevent plant diseases. These techniques can be used to analyze the collected data and identify the initial signs of pest infestations or other diseases, which will allow for further effective actions.

In [4], the authors have developed an IoT platform using LoRaWAN (‘Long-range wide-area network’) technology specifically designed for smart irrigation. The study implements LoRaWAN within an IoT framework and it is used for precision

irrigation. This system was tested in a plasticulture field for fresh-market tomato production and the obtained results indicate the strong potential in aspects of precision irrigation, yield improvement and water conservation.

Beyond the agricultural context, IoT and AI technologies are increasingly being adopted in the food industry and dairy supply chain, where they contribute to real-time monitoring, predictive analytics, and quality assurance. Recent studies demonstrate that IoT-based systems integrated with machine learning models can significantly enhance milk production management and safety. For example, [5] introduced an inexpensive AI-powered IoT sensor for continuous farm-to-factory milk quality monitoring. This system combines spectral analysis with machine learning models such as Random Forest to estimate protein and fat content, achieving an impressive accuracy of ±0.14% for fat and ±0.07% for protein. Similarly, [6] presented predictive modeling for milk production using artificial intelligence and machine learning techniques, emphasizing the ability of these models to capture complex nonlinear relationships between environmental, biological, and management factors that influence milk yield. The study compared several approaches, including Random Forest, Gradient Boosting, Support Vector Machines, and Artificial Neural Networks, demonstrating that ensemble and deep learning models consistently provided more reliable and accurate predictions than traditional statistical methods. The authors concluded that integrating AI and ML within dairy management systems can significantly improve forecasting capabilities, resource allocation, and decision-making efficiency in milk production.

In addition, [7] developed an IoT-based monitoring framework designed to preserve the quality and safety of dairy products throughout the cold supply chain. The proposed system, named i-EAT, integrates sensor-equipped crates (i-EAT Tags) and vehicle-level modules (i-EAT Trucks) capable of collecting temperature, humidity, vibration, pressure, and GPS location data during transportation. Using Bluetooth Low Energy (BLE) for local communication and cloud-based data storage, the system allows real-time tracking and early detection of anomalies, such as temperature fluctuations or excessive vibration during loading and delivery. The study demonstrated that this IoT infrastructure significantly enhances traceability, reduces waste, and increases consumer confidence by ensuring product quality and freshness from producer to retailer.

Complementing these efforts, [8] presented a digital IoT-enabled platform for milk collection and quality assessment using machine learning algorithms to detect adulteration in milk. Their model achieved an accuracy rate above 98%, demonstrating that ML-based IoT systems can effectively ensure product integrity and transparency across the dairy value chain.

Taken together, these studies provide strong evidence that combining IoT infrastructures with advanced machine learning algorithms can lead to substantial improvements in predictive performance, real-time control, and process optimization. This growing body of literature validates the approach presented in this work, which integrates IoT technologies with ensemble regression models such as Random Forest and Gradient Boosting [9, 10, 11] to forecast milk consumption patterns. The use of multiple temporal and contextual features, as shown in the present study, further

enhances prediction accuracy, aligning with the best practices and quantitative gains reported in recent research.

3 Methodology

The primary objective of this project is to develop an IoT system for real-time monitoring and management of a milk dispenser. The project aims to create a sensor infrastructure capable of continuously monitoring milk levels and the internal temperature of the dispenser. To achieve this, the system integrates the following sensors: a DHT11 temperature sensor, selected for its simplicity, low cost, and adequate precision for monitoring storage conditions, and a VL53L1X time-of-flight (ToF) distance sensor, which measures the distance from the sensor to the milk surface. This non-contact optical method allows for accurate volume estimation by comparing the measured distance with the known internal height of the dispenser, avoiding the mechanical complexity and calibration issues associated with load cells.

The final objective is to implement machine learning algorithms capable of analyzing sensor data to forecast milk consumption and detect potential anomalies, such as irregular patterns that could indicate blockages or malfunctions, thereby enabling timely maintenance and uninterrupted operation.

This IoT system is designed to provide real-time tracking of milk levels and temperature, detect potential blockages, and predict consumption patterns. The architecture of this system consists of three components:

- Hardware: used for the data collection.
- Cloud storage: used for storing the data.
- Desktop application: used for user interaction.

The system architecture can be seen in Fig. 1 above, which illustrates the integration of sensors with the ESP32 microcontroller, cloud communication, and user visualization. In addition to the sensors implemented in the prototype, the diagram also includes a pH sensor and a weight sensor, which were added to reflect the potential future extension of the application and the possibility to integrating additional measurement capabilities in later development stages.

The hardware layer acts as the core processing unit and is connected to the necessary sensors to collect data: the distance sensor measures the level of milk in the dispenser and the temperature sensor monitors the temperature inside the dispenser. The sensors are linked to the microcontroller, which manages raw data and prepares it for transmission. The microcontroller processes the collected data, then transmit it to the cloud. After the data is gathered using the system, it is recorded inside Firebase platform using WI-FI module. The desktop application retrieves real-time sensor data from Firebase and allows the user to visualize the output in an intuitive interface.

3.1.Implementation

The electrical circuit presented in Fig. 2 illustrates the integration of the DHT11 temperature and the VL53L1X time-of-flight distance with the ESP32 development board.

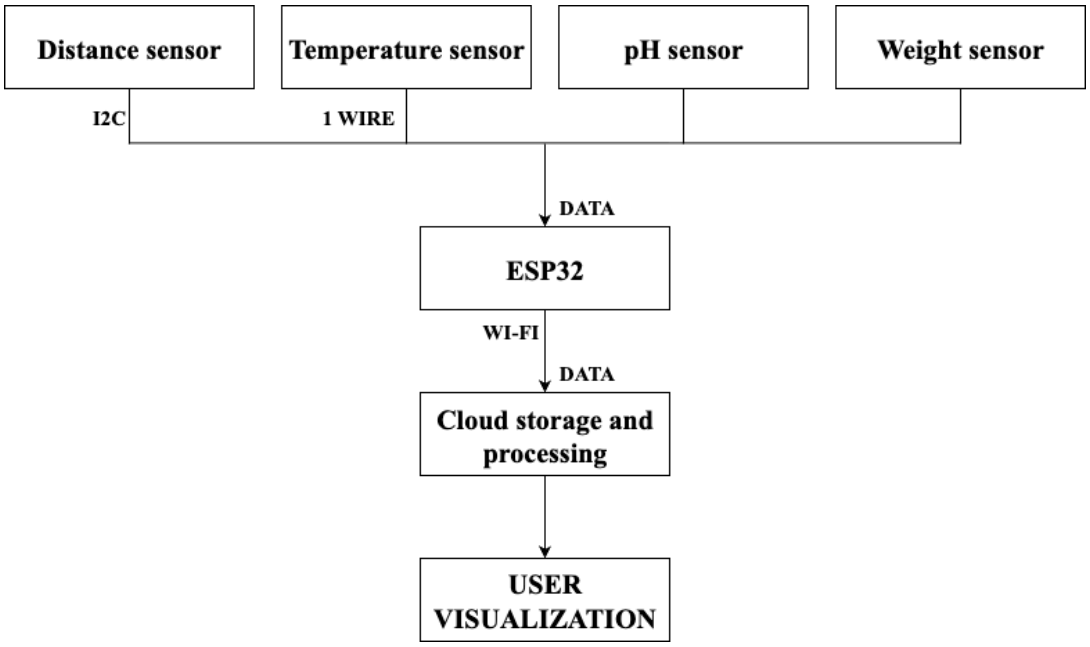


Figure 1. System architecture

The DHT11 sensor utilizes a single-wire digital communication protocol, enabling reliable temperature and humidity data acquisition with minimal wiring complexity. In contrast, the VL53L1X operates via the Inter-Integrated Circuit (I²C) communication protocol, which allows efficient data transfer through its Serial Data (SDA) and Serial Clock (SCL) lines.

The ESP32 serves as the central processing unit, providing the necessary voltage supply, ground reference, and data communication lines for both sensors. This configuration ensures accurate environmental sensing while maintaining compatibility with the microcontroller's digital and I²C interfaces.

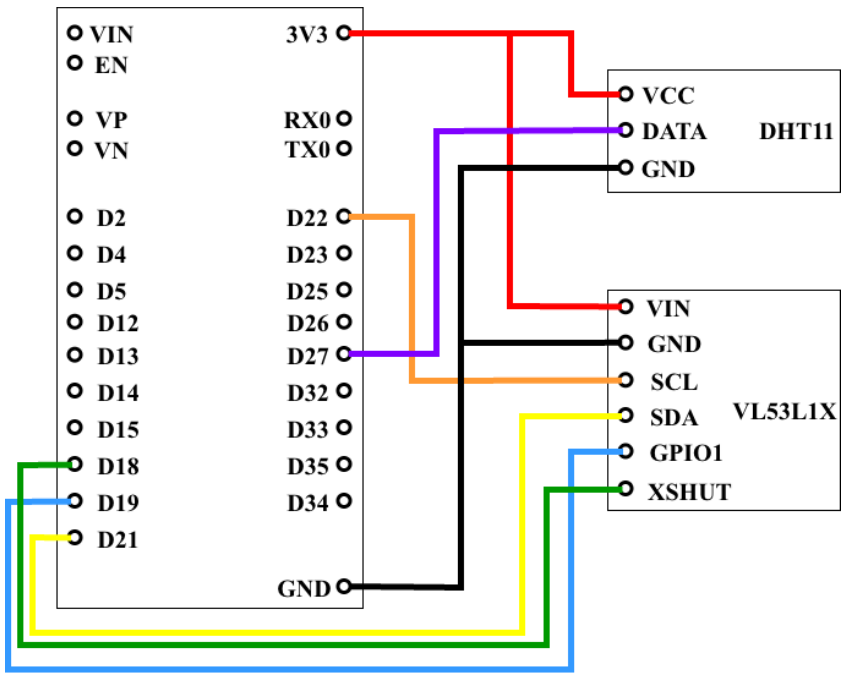


Figure 2. Electrical circuit

The real hardware implementation of the proposed system is illustrated in Fig. 3 below.

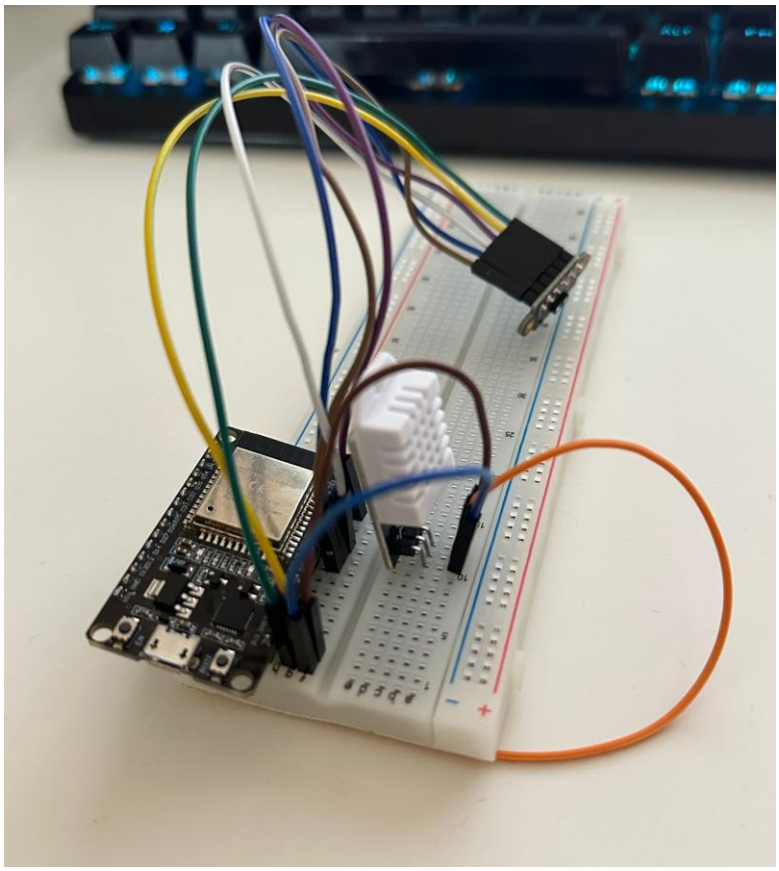


Figure 3. Real-life hardware setup with ESP32 and sensors

Fig. 4 illustrates the complete workflow for predicting milk consumption using machine learning. The process begins with data collection from Firebase, followed by preprocessing, where only events labeled “Buy” are retained and subsequently grouped by day to compute daily quantities. In cases where a date contains no data, a value of zero is assigned to preserve the continuity of the time series. Feature extraction is then performed, incorporating attributes such as the day of the week and weekend indicators. Based on these features, a regression model is selected. When Random Forest is applied, multiple trees are trained in parallel and their outputs are averaged, while Gradient Boosting trains trees sequentially, each improving upon the previous iteration. The final predictions are stored and subsequently evaluated using standard performance metrics.

Random Forest is an ensemble learning method that constructs multiple independent decision trees using the bagging technique. These trees are trained in parallel and subsequently combined, a strategy that reduces overfitting and enhances prediction accuracy, while also providing faster training. In contrast, Gradient Boosting builds trees sequentially, where each new tree corrects the errors of its predecessors. This approach is more effective in capturing complex patterns but generally incurs higher computational costs. In the context of milk consumption forecasting, both algorithms are applied to historical data in order to improve demand prediction and system reliability.

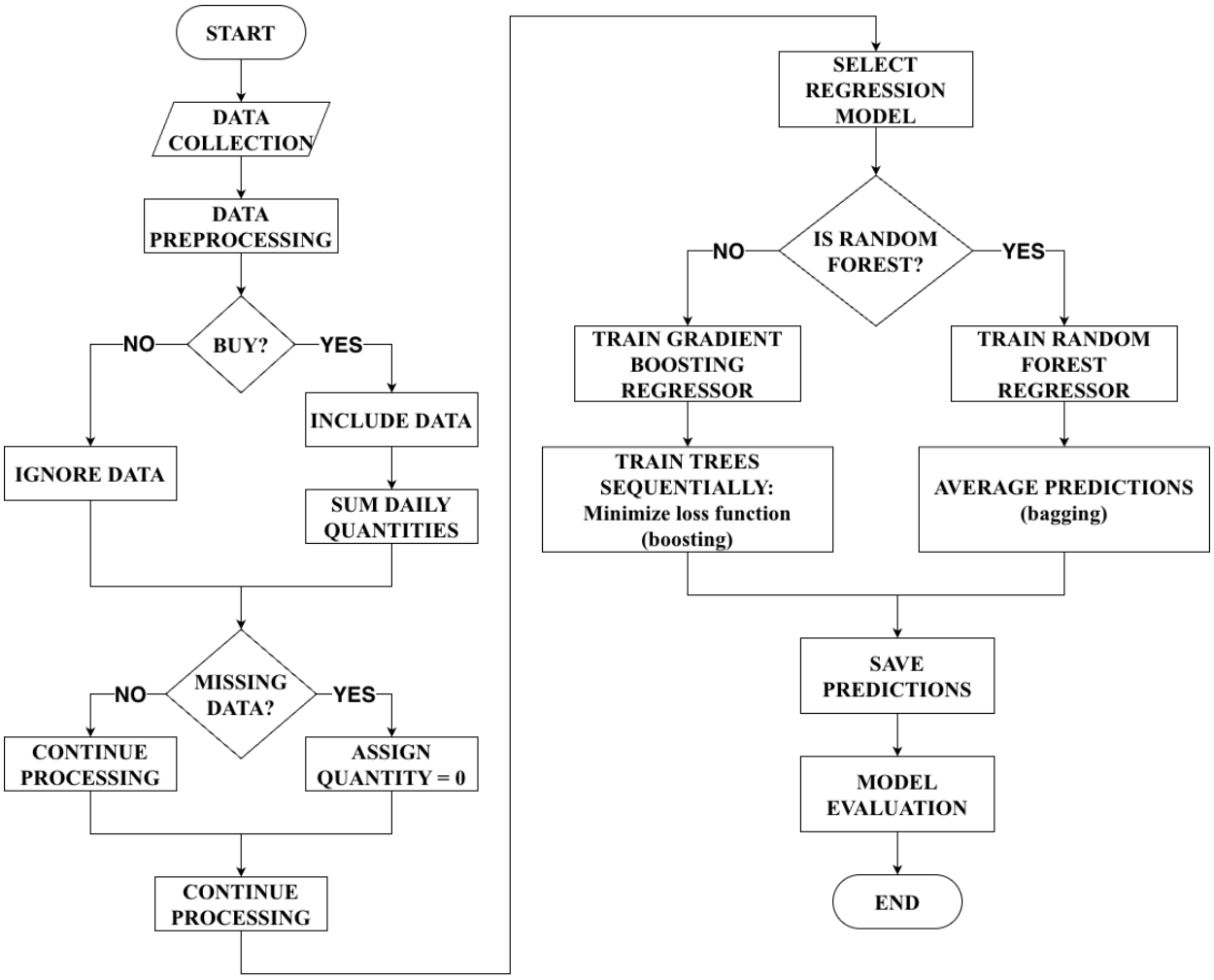


Figure 4. Regression model flow

4 Results

The prediction task was conducted using two distinct feature sets. The first set consisted of the day of the week, encoded as an integer ranging from 0 (Monday) to 6 (Sunday), and a Boolean variable indicating whether the day corresponded to a weekend or a weekday. The performance of both algorithms based on these features is illustrated in the presented results. The findings indicate that both models achieve reasonably accurate forecasts, successfully capturing the general trend of the actual values even with a minimal feature set. The discrepancies between predicted and observed values remain relatively small; however, Gradient Boosting demonstrates superior performance in most cases, exhibiting a stronger capacity for generalization.

Table 1 compares actual recorded daily milk consumption to the corresponding predictions given by the Random Forest and Gradient Boosting models over a one-week period. The table demonstrates the models' ability to estimate actual consumption trends using historical data and specified temporal variables. By comparing the genuine values to the anticipated outputs, it is able to visually analyze each algorithm's accuracy and consistency, as well as find tiny differences between

observed and estimated values. This comparison provides an early assessment of the prediction performance of both models.

Table 1. Prediction - two features

Date	True value(liters)	Random Forest (liters)	Gradient Boosting (liters)
5/29/2025	75	72.15	72.14
5/30/2025	61	70.32	70.23
5/31/2025	79	76.31	76.43
6/1/2025	68	72.24	72.00
6/2/2025	75	72.94	73.00
6/3/2025	70	71.94	71.93
6/4/2025	78	70.85	70.86

The graphical representation from Fig. 5, emphasizes the close alignment between predicted and actual values across the tested days, although minor deviations are still present.

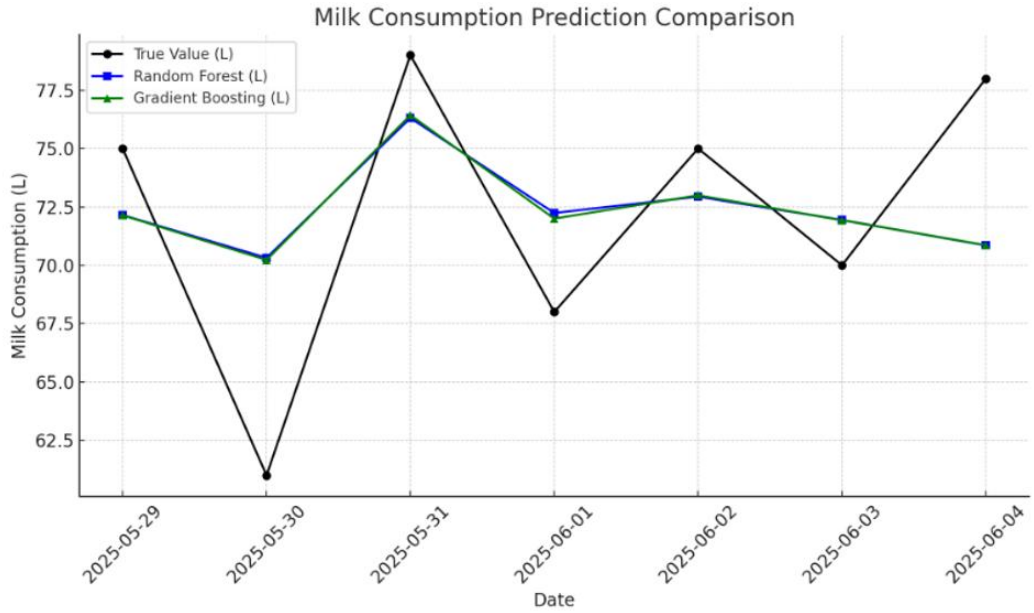


Figure 5. True vs predicted values - two features

To evaluate the performance of the models, several error metrics were computed, as summarized in Table 2.

Table 2. Metrics for two features

	MAE	RMSE	R^2
Random Forest	4.32	5.06	0.26
Gradient Boosting	4.27	5.01	0.28

The Mean Absolute Error (MAE) obtained by both algorithms is approximately 4.3, indicating relatively small deviations between predicted and actual values. The Root Mean Squared Error (RMSE) is slightly higher than the MAE, reflecting the presence of occasional larger prediction errors. Gradient Boosting achieves a marginally lower RMSE compared to Random Forest, suggesting a reduced frequency of large errors. Although the coefficient of determination (R^2) values is relatively low, they still

indicate that the models are able to capture part of the variance in the data, with Gradient Boosting providing a modest improvement in explanatory power.

To enhance model performance, a third feature, day of the month, was incorporated to provide additional temporal context. This feature enables the models to better capture periodic patterns, resulting in a noticeable improvement in prediction accuracy. Gradient Boosting continues to slightly outperform Random Forest, demonstrating a stronger ability to exploit the additional information. The visual comparison highlights that including the day of the month allows both models to adapt more effectively to daily variations. Table 3 compares the actual daily milk consumption values with the predictions generated by both models using the enriched three-feature configuration.

Table 3. Prediction - three features

Date	True value (liters)	Random Forest (liters)	Gradient Boosting (liters)
5/29/2025	75	73.59	75.94
5/30/2025	61	64.58	62.09
5/31/2025	79	74.13	77.43
6/1/2025	68	66.65	68.87
6/2/2025	75	75.09	74.99
6/3/2025	70	71.97	72.61
6/4/2025	78	77.06	77.72

Including this feature results in a noticeable improvement in prediction accuracy across most cases. Furthermore, the slight performance advantage observed for Gradient Boosting indicates that this algorithm is more effective in leveraging the additional information provided by new features, thereby producing more precise predictions.

Figure 6 presents the comparison between the observed values and the predictions generated by the two algorithms when three features are included. The results clearly indicate that adding the day of the month as a feature enables the models to better capture daily variations compared to the two-feature configuration. This observation is further supported by the evaluation metrics summarized in Table 4.

Gradient Boosting consistently outperforms Random Forest across all metrics, achieving a substantially higher R^2 and significantly lower error values. These results highlight the superior capacity of Gradient Boosting to exploit the richer feature set, thereby generating more accurate and reliable predictions.

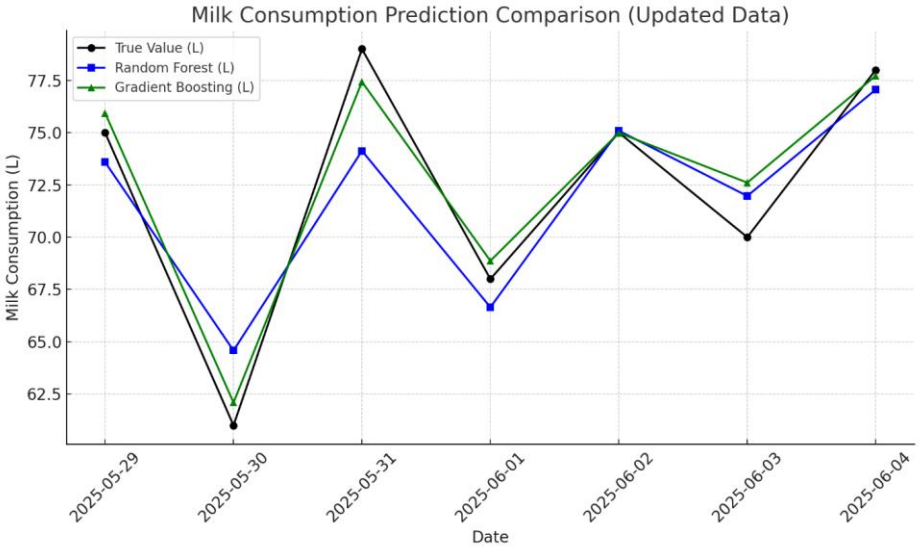


Figure 6. True vs predicted values - three features

Table 4 Metrics - three features

	MAE	RMSE	R^2
Random Forest	2.03	2.54	0.81
Gradient Boosting	1.05	1.32	0.95

To investigate the performance of the model and evaluate the relevance of the selected features, a new dataset covering the period January 2024 – June 2025 was generated. Since the prototype was not yet connected to a real milk dispenser, the training data were synthetically generated within realistic value ranges, allowing the analysis to be carried out in a controlled environment. In future work, these models will be retrained on real world operational data once the system is deployed on an actual dispenser. This extended dataset allowed the introduction of two additional features, public holiday and season, which were used to encapsulate temporal patterns that may influence milk consumption patterns. The four-feature set includes a binary variable indicating whether a given day is a public holiday, reflecting potential changes in consumer behavior on special days. The five-feature set further incorporates the season, providing a broader temporal context that may affect consumption trends. By progressively enriching the feature set, the models can better learn periodic and contextual variations, which is expected to improve predictive accuracy for both Random Forest and Gradient Boosting algorithms.

Table 5 presents the predictions obtained by the Random Forest model for a one-week period, from April 2 to April 8, 2025. The results are shown for configurations using two to five features. When only two features are used, the discrepancies between predicted and actual values are evident. For example, on April 2, the observed value was 85.99 liters, while the model predicted 72.23 liters, indicating a substantial underestimation and highlighting the limitations of a basic temporal structure. In contrast, the five-feature set produces a prediction of 85.42 liters for the same day, demonstrating a significant improvement in accuracy and a stronger ability to capture daily consumption patterns.

Table 5. Random Forest

Date	True value (liters)	Two features (liters)	Three features (liters)	Four features (liters)	Five features (liters)
4/2/2025	85.99	72.23	87.02	86.99	85.42
4/3/2025	77.17	68.18	80.13	80.35	79.73
4/4/2025	77.9	67.06	81.82	81.83	79.8
4/5/2025	81.92	77.92	76.14	76.29	80.18
4/6/2025	82.89	77.45	77.18	77.16	79.89
4/7/2025	63.76	68.34	65.2	65.23	63.69
4/8/2025	64.02	67.96	65.53	65.54	63.55

Table 6 summarizes the evaluation metrics for the Random Forest model across the different feature sets. When only two features are used, the model performs poorly, with a low R^2 value indicating its inability to capture meaningful patterns in the data. Adding a third feature leads to a substantial improvement: MAE decreases from 7.36 to 3.19, and R^2 increases from 0.02 to 0.8. Adding the holiday feature (fourth feature) yields similar performance to the three-feature configurations, which confirms the fact that holidays may not influence in a drastic way the weekly predictions. The best performance is achieved with the five-feature set, which underline the importance of including seasonal information for more accurate forecasting.

Table 6. Metrics – Random Forest

Number of features	MAE	RMSE	R^2
Two	7.36	8.19	0.02
Three	3.19	3.69	0.80
Four	3.21	3.69	0.80
Five	1.47	1.80	0.95

Table 7 presents the corresponding predictions obtained using the Gradient Boosting model over the same period and feature configurations. As with Random Forest, the use of only two features results in underperformance, particularly on April 2, where the predicted value significantly deviates from the observed one. The inclusion of a third feature improves the alignment between predicted and actual values, while the addition of holiday and seasonal features leads to further refinements in accuracy. The five-feature configuration produces the most balanced and accurate predictions, with values closely matching the true consumption data on multiple days.

Table 7. Gradient Boosting

Date	True value (liters)	Two features (liters)	Three features (liters)	Four features (liters)	Five features (liters)
4/2/2025	85.99	72.23	84.74	84.52	84.77
4/3/2025	77.17	68.22	81.52	81.66	81.55
4/4/2025	77.9	66.99	81.69	81.68	81.54
4/5/2025	81.92	77.93	76.92	77.58	77.22
4/6/2025	82.89	77.42	76.41	76.46	77.41
4/7/2025	63.76	68.37	64.86	65.0	64.88
4/8/2025	64.02	67.94	64.87	64.65	64.79

Table 8 shows the evaluation metrics for the Gradient Boosting model. Similar to Random Forest, performance with two features is weak, reflected by high error values and a low R^2. Adding the third feature leads to a substantial improvement in both MAE and R^2, while the inclusion of holiday and seasonal features further stabilizes the model's performance, yielding the best overall results with the five-feature configuration.

Table 8 Metrics – Gradient Boosting

Number of features	MAE	RMSE	R^2
Two	7.35	8.20	0.02
Three	3.26	3.85	0.78
Four	3.20	3.76	0.79
Five	3.04	3.54	0.82

In conclusion, both Random Forest and Gradient Boosting models benefit significantly from the inclusion of additional temporal and contextual features. While Random Forest achieves slightly better performance with the full feature set, Gradient Boosting demonstrates stable behavior even with the fewer features, confirming the effectiveness of ensemble learning techniques for forecasting consumption patterns in time-series data.

Figure 7 compares the evaluation metrics of both models across progressively richer feature sets. The results highlight that the inclusion of the day-of-month and seasonal features substantially reduces MAE and RMSE, while simultaneously increasing R^2. Although Gradient Boosting demonstrates stable performance with fewer features, Random Forest slightly outperforms it when the complete feature set is employed, achieving the most accurate predictions overall.

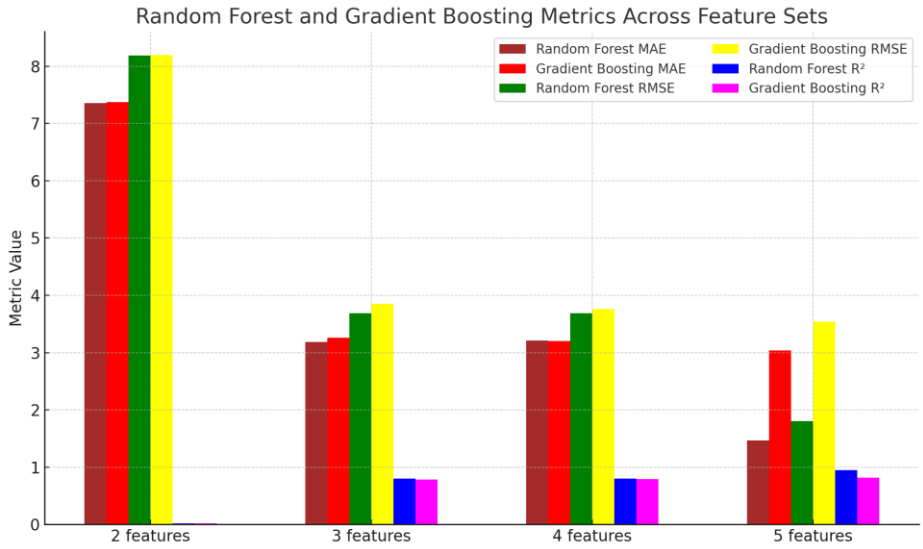


Figure 7. Performance metrics comparison

4.1.Prototype Implementation

To validate and test the functionality of the implemented system, a basic prototype was built using a container filled with water and the sensors mounted above it. The purpose was to simulate variations in liquid volume and confirm that the system accurately detects and displays the quantity in the monitoring application.

The prototype includes the following elements and could be seen in the Fig. 8 below:

- A container partially filled with water used to simulate the milk tank.
- VL53L1X sensor placed above the container connected to the microcontroller.

It is important to note that although the system architecture includes both a temperature and a distance sensor, the prototype used in this testing stage implements only the distance sensor. The temperature sensor was planned to be tested on an actual milk dispenser, but this step could not be implemented in time; therefore, the displayed temperature value represents an approximative reference corresponding to typical dispenser conditions.

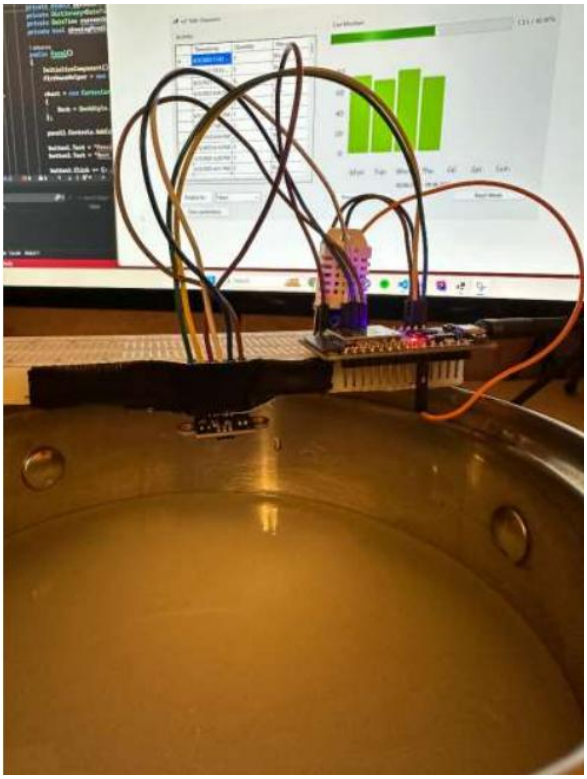


Figure 8. Initial setup

To observe the changes in quantity, measurements were taken at two stages: before and after consumption. Initially, the container was filled with 1.5 liters of water. The sensor measured the distances from its position to the water surface. By using the dimensions of the container, the measured distance was converted into an estimated volume of liquid.

The images below illustrate the experimental technique used to validate the system's ability to detect and display liquid level changes in real time. Figure 9 shows the initial state of the container before any modifications were made. To simulate real usage conditions, 0.5 liters of water were manually removed from the container. After this change, the sensor immediately detected the new distance between the sensor and the liquid surface, and the application updated the displayed value accordingly, as illustrated in Figure 10. This experiment confirmed that the system can accurately register and reflect even small variations in liquid level, demonstrating both the precision of the sensor and the responsiveness of the data transmission and visualization process.

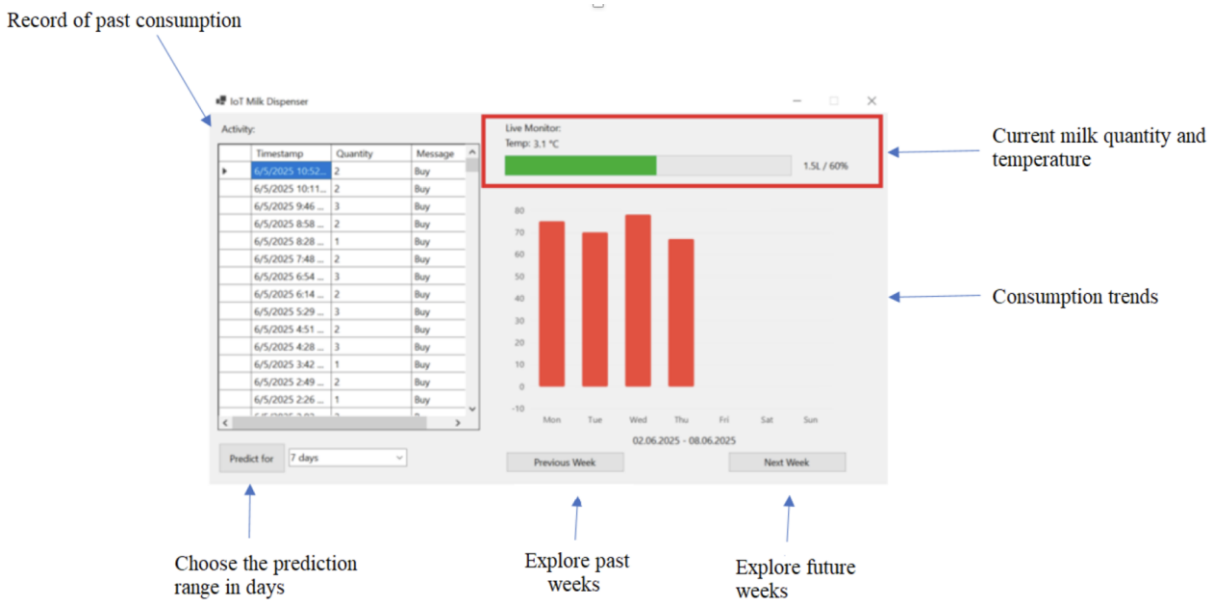


Figure 9. Initial state of the container

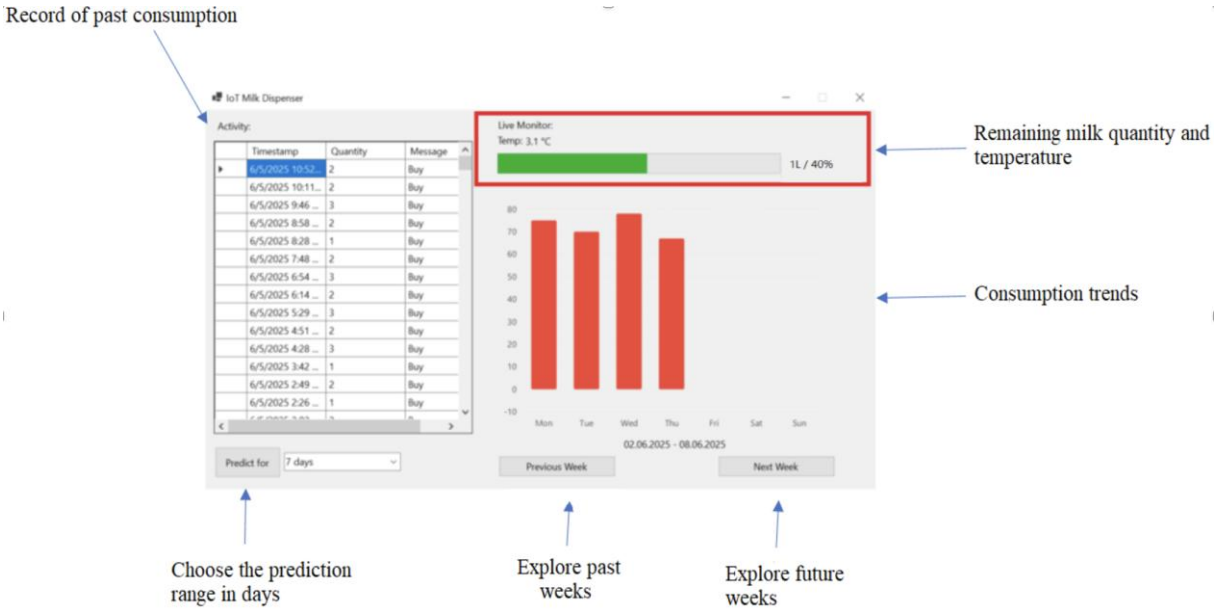


Figure 10. Final state of the container

5 Conclusions

This article explored the design and prototyping of an IoT-based system for monitoring the level of milk in a dispenser. The developed system successfully integrated multiple sensors with an ESP32 microcontroller, Firebase for cloud-based synchronization, and a dedicated application for real-time visualization. The experimental evaluation of the prediction models demonstrated that the inclusion of more than three features significantly improved the accuracy of the forecasts. Both Random Forest and Gradient Boosting benefited from the richer feature set, but Gradient Boosting consistently outperformed Random Forest, achieving higher R^2 values and lower error metrics. These results confirm that the integration of additional temporal and contextual information enhances the capacity of the models to anticipate milk consumption patterns with higher reliability.

Acknowledgment

This research was conducted as part of the collaborative work of three authors within the Embedded Systems master's program at the Faculty of Engineering, "Lucian Blaga" University of Sibiu, Romania. The authors gratefully acknowledge the support and guidance provided throughout the development of this study.

References

[1] Dataintelo. (2024). Milk Vending Machines Market – Global Industry Analysis, Size, Share, Growth, Trends, and Forecast 2024–2032. Dataintelo

[2] Rabiya Abbasi, Pablo Martinez, Rafiq Ahmad, The digitization of agricultural industry – a systematic literature review on agriculture 4.0, Smart Agricultural Technology, Volume 2, 2022, 100042, ISSN 2772-3755, https://doi.org/10.1016/j.atech.2022.100042.

[3] M. Ayaz, M. Ammad-Uddin, Z. Sharif, A. Mansour and E. -H. M. Aggoune, "Internet-of-Things (IoT)-Based Smart Agriculture: Toward Making the Fields Talk," in IEEE Access, vol. 7, pp. 129551-129583, 2019, doi: 10.1109/ACCESS.2019.2932609.

[4] Haozhe Zhang, Long He, Francesco Di Gioia, Daeun Choi, Antonio Elia, Paul Heinemann, LoRaWAN based internet of things (IoT) system for precision irrigation in plasticulture fresh-market tomato, Smart Agricultural Technology, Volume 2, 2022, 100053, ISSN 2772-3755, https://doi.org/10.1016/j.atech.2022.100053.

[5] Kaneez Fizza, Abhik Banerjee, Dimitrios Georgakopoulos, Prem Prakash Jayaraman, Ali Yavari, nas Dawod, An Inexpensive AI-Powered IoT Sensor for Continuous Farm-to-Factory Milk Quality Monitoring, Sensors 2025, 25, 4439, https://doi.org/10.3390/s25144439.

[6] V. Manibabu and M. Gomathy, "Predictive Modeling of Milk Production Using Artificial Intelligence and Machine Learning Techniques," International Journal of Computer Techniques, vol. 12, no. 3, May-June 2025.

[7] Syrmos, E. et al. (2023). An IoT Monitoring System for Dairy Products. In: Rabie, K., Kharel, R., Mohjazi, L., Elganimi, T., Khan, W.U. (eds) IoT as a Service. IoTaaS 2022. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 506. Springer, Cham. https://doi.org/10.1007/978-3-031-37139-4_2.

[8] R. D. Prasad and S. Chidambaranathan, "Digitizing Milk Collection Stage with IoT-Enabled System for Quality Assessment using ML," Journal of Information Systems Engineering and Management, vol. 10, no. 29s, 2025, https://doi.org/10.52783/jisem.v10i29s.4604.

[9] Henzel, J., & Sikora, M. (2020). Gradient Boosting Application in Forecasting of Performance Indicators Values for Measuring the Efficiency of Promotions in FMCG Retail. Silesian University of Technology, Faculty of Automatic Control, Electronics and Computer Science.

[10] Ayaru, L., Ypsilantis, P.-P., Nanapragasam, A., Choi, R. C.-H., Thillanathan, A., Min-Ho, L., & Montana, G. (2021). Prediction of Outcome in Acute Lower Gastrointestinal Bleeding Using Gradient Boosting.

[11] Boldini, D., Grisoni, F., Kuhn, D., Friedrich, L., & Sieber, S. A. (2023). Practical guidelines for the use of gradient boosting for molecular property prediction. Journal of Cheminformatics, 15(1), Article 35.

DOI: 10.2478/ijasitels-2025-0011

Integrating Multi-View features into diagrammatic languages

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 define diagrammatic languages by metamodels represented by categorical sketches. Then, a model, corresponding to a categorical sketch, is a Set-value functor M:$\mathcal{G}$→Set, where $\mathcal{G}$ is the sketch graph, and which respects the constraints imposed by the categorical sketch. A view of a model M, is another model $\mathcal{V}$ in the diagrammatic language that maps to M by a natural injective transformation. This is a theoretically satisfactory definition but does not provide us with solutions to actually find these views. In this paper, we will present a practical way to construct a satisfactory set of views for a given model. Moreover, this mechanism can be implemented generically, at the modeling language level, and then used for each model separately. The categorical mechanisms that we use to solve this problem are representable functors and the Yoneda lemma. This is also the main contribution of this paper to solving problems related to multi-view modeling.

Keywords: category, functor, model, categorical sketch, representable functors, multi-view, meta-model

1 Introduction

The high complexity of the systems that modeling activity currently faces requires the decomposition of models according to various complementary points of view. A conceptual and methodological approach [9], which allows modeling a complex system by decomposing it into complementary sub-models, is multi-view modeling [10]. In this approach, each view captures, partially or totally, a certain structural, behavioral or functional aspect of the system. These sub-models can be assembled into a whole that provides a complete and coherent picture of the modeled system.

Most modern complex systems such as cyber-physical or socio-technical systems involve multiple dimensions and actors with different concerns, objectives and perspectives. Under these conditions, a single global model rarely manages to capture all the necessary requirements. Multi-view modeling can respond to these challenges by conceptually decomposing the system into several sub-models, depending on these requirements.

In the Model-Driven Engineering (MDE) paradigm, views are defined as independent models, which are connected through mappings, transformations or synchronization rules that make up the concept of Model Integration. In the context of multi-level

modeling, hierarchical levels of abstraction can be detailed through views, thus maintaining coherence between levels and views.

Significant conceptual flexibility offered by multi-view modeling, however, also comes with a series of challenges related, first of all, to maintaining inter-model consistency, especially in large, dynamic systems and to automatically synchronizing changes between views. In the case of linguistically heterogeneous models, semantic compatibility issues also arise. In general, the solutions proposed by several recent researches to solve these problems are based on model transformation [4], ontology-based integration and co-simulation [15][8].

In the MDE paradigm, the development of domain-specific diagrammatic modeling languages [14] is an integral part of the modeling process [11]. Obviously, in the context of multi-view modeling, these languages must be endowed with special facilities that simplify the multi-view modeling process [10].

In this paper, we define diagrammatic languages by metamodels represented by categorical sketches. This approach allows the use of categorical concepts such as representable functors and Yoneda's lemma to endow diagrammatic languages with important facilities for specifying the views of a model. Thus, the endowment of diagrammatic modeling languages with multi-view modeling facilities based on representable functors and Yoneda's lemma represents the main contribution of this paper to the field of multi-view modeling.

Section 2 of the paper contains background notions, section 3 defines diagrammatic languages based on a meta-model represented by a categorical sketch, section 4 presents the use of representable functors in integrating multi-view modeling facilities into diagrammatic languages. In the last section, some conclusions and observations are presented.

2 Background Notions

Graph theory and category theory are the fundamental theories that support modeling in general and multi-view modeling in particular. In this section, we will briefly present some notions in this field and some notations used in this paper.

A graph is a tuple $\boldsymbol{G}=(X, \Gamma, \sigma, \theta)$ composed of the set of nodes X, the set of relations Γ and two functions $\sigma,\theta:\Gamma\rightarrow X$ indicating the source and, respectively, the target of each arc of the graph. Intuitively, a graph is a collection of nodes connected by directed arcs, but without composition rules. In its most intuitive form, a category is a graph with rules for the composition of arrows and with identity arrows for each node.

Thus, a category $\boldsymbol{C}$ consists of a set of objects $ob(\boldsymbol{C})$ and for each pair of objects A,B, a set of morphisms (arrows) Hom(A,B). The set of morphisms is endowed with a partial composition operation: $o:Hom(B,C)\times Hom(A,B)\rightarrow Hom(A,C)$ which is associative and for each object A, there is an identity morphism $id_A \in Hom(A,A)$ [1][5].

As we see, a category is endowed with a structure, given by the composition operation, unlike graphs. This structure also leads to the difference between a graph homomorphism and a functor. Thus, a functor is an application between categories. Which takes objects into objects and arrows into arrows and which preserves the categorical structure.

Among all functors, Set-value functors are distinguished, which are useful in modeling. A Set-value functor is a functor that maps an arbitrary category $\mathcal{C}$ into the category Set [1] [5].

Between two functors ϕ and ψ, from the category $\mathcal{C}$ to the category $\mathcal{D}$, we can define a map τ, which maps the image of the functor ϕ onto the image of the functor ψ, in the category $\mathcal{D}$, and which if it satisfies a condition, called the naturalness condition, is called a natural transformation. Natural transformations are defined at the level of objects of the category $\mathcal{C}$. The naturalness condition requires that for any arrow $f:A\rightarrow B$ in $\mathcal{C}$, we have the equality $(\psi f) \circ \tau_A=\tau_B \circ (\phi f)$, where τ_A and τ_B are the components of τ for objects A and B [1] [2].

Natural transformations are maps that express a natural way of mapping one functor to another. Obviously, natural transformations can be composed by the classical operation of composition of functions. Therefore, if we have two categories $\mathcal{C}$ and $\mathcal{D}$, we can form a category of functors that has as objects: functors defined on $\mathcal{C}$ with values in $\mathcal{D}$, as arrows the natural transformations between functors and as composition operation, the composition of natural transformations on components.

In this paper we specify diagrammatic languages by metamodels represented by categorical sketches. A categorical sketch [13], is a tuple formed by a graph $\mathcal{G}$, and a set of constraints $C(\mathcal{G})$, on models expressed by Set-value functors.

Constraints can be imposed by commutative diagrams, categorical limits and colimits or by predicates expressed on the components of some diagrams, called graph forms [7] [12] [13]. Graph forms are used as diagrams on which conditions are imposed by commutativity, limits or colimits, or as shape graph arity for predicates that specify constraints.

3 Diagrammatic languages

We define diagrammatic languages by meta-models represented by categorical sketches. A categorical sketch is composed of a typed graph $\mathcal{G}$ and a set of constraints $C(\mathcal{G})$ imposed on the image of the graph $\mathcal{G}$, by a functor M in the category Set, that is, on a model represented by the sketch $\mathcal{S}=(\mathcal{G}, \mathcal{C}(\mathcal{G}))$ [6] [7].

The typed graph $\mathcal{G}=(T, \Gamma, \sigma, \theta)$ is composed of the set of node types T, the set of relationship types Γ and two functions $\sigma,\theta:\Gamma\rightarrow T$ indicating the source and, respectively, the target of each arc of the graph. Obviously, the sets T and Γ can represent both concept types from the modeling domain and attribute types attached to the concepts.

A model $M:\mathcal{S}\rightarrow S$, is a functor $M:\mathcal{G}\rightarrow Set$, which maps each type of object $t\in T$ into a set of objects of type t and each type of arc $\gamma\in\Gamma$ into a function of the same type as the corresponding arc [6] [7].

Example 1. We will consider a very simple example of a categorical sketch to illustrate our approach to the problems treated in this paper. First, we define the types of concepts in the modeling domain and the graphical atoms of the modeling language that we want to specify. We therefore have the following concepts:

A type T_1 that represents buffers with limited capacity, which can store a single type of elements, and for which we use as graphical notation a red circle.

The type T_1 will also have two attributes, namely: stock and capacity.

A type T_2 that represents buffers with limited capacity, which can store two types of elements, and for which we use as graphical notation a blue rectangle.

The type T_2 will also have four attributes, namely: stock1, stock2, capacity1 and capacity2.

A type T_3 that contains an integer from a range of integers, representing values for the attributes of types T_1 and T_2, and for which we use a white circle as a graphic notation.

The type T_3 will have an attribute, namely: value.

We will also use oriented arrows to represent the relationships and functions involved in the diagrams. For our example, we will consider that the graph $\boldsymbol{\mathcal{G}}$, of the sketch $\boldsymbol{\mathcal{S}}$ is the one in Figure 1. To simplify the approach, we will consider that the sketch $\boldsymbol{\mathcal{S}}$, which is our metamodel, does not contain constraints. We also mention that the sketch graph could represent the attributes and even the algebra generated by the attributes [3] [4].

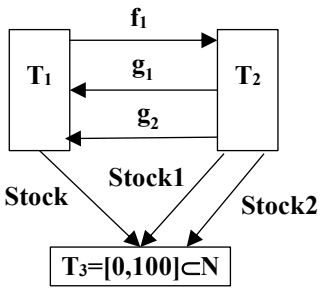


Figure 1. The graph of the sketch

A model specified by the sketch $\boldsymbol{\mathcal{S}}$ is the following: $M(T_1)=\{D_{11},D_{12},D_{13},D_{14}\}$; $M(T_2)=\{D_{21},D_{22}\}$; $M(T_3)=\{10,20,30\}$;

Although the arcs of the graph become functions, we will write them with the same symbols, that is, we consider that: $M(f_1)=f_1$; $M(g_1)=g_1$; M(Stock)=Stock; M(Stock1)=Stock1; M(Stock2)=Stock2.

These are defined as follows:

f_1: $M(T_1)\rightarrow M(T_2)$; $f_1(D_{11})=D_{22}$, $f_1(D_{12})=D_{22}$, $f_1(D_{13})=$Null, $f_1(D_{14})=$Null,

Stock: $M(T_1)\rightarrow M(T_3)$; $Stock(D_{11})=20$, $Stock(D_{12})=10$, $Stock(D_{13})=30$, $Stock(D_{14})=30$,

g_1: $M(T_2)\rightarrow M(T_1)$; $g_1(D_{21})=D_{11}$, $g_1(D_{22})=D_{13}$,

g_2: $M(T_2)\rightarrow M(T_1)$; $g_2(D_{21})=D_{12}$, $g_2(D_{22})=D_{14}$,

Stock1: $M(T_2)\rightarrow M(T_3)$; $Stock1(D_{21})=20$, $Stock1(D_{22})=10$,

Stock2: $M(T_2)\rightarrow M(T_3)$; $Stock2(D_{21})=10$, $Stock2(D_{22})=20$,

This model can be seen in diagrammatic form in Figure 2.

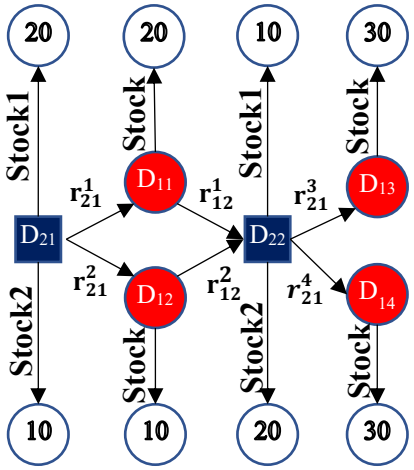


Figure 2. An instance of the sketch in Figure 1.

The diagrammatic language DL, represented by the sketch $\boldsymbol{\mathcal{S}}$, consists of all functors of the form M:$\boldsymbol{\mathcal{S}}$→Set and we denote it by DL($\boldsymbol{\mathcal{S}}$).

If $\boldsymbol{\mathcal{S}}$ is a categorical sketch and M_1, M_2 are two models, that is, two functors M_1,M_2:$\boldsymbol{\mathcal{S}}$→Set, then we can define a natural transformation τ:M_1,→M_2. The set of all functors M:$\boldsymbol{\mathcal{S}}$→Set together with the natural transformations between them form a category that we denote by $\boldsymbol{\mathcal{M}}(\boldsymbol{\mathcal{S}})$ and we call it the category of models of the language DL($\boldsymbol{\mathcal{S}}$).

As we mentioned, in a diagrammatic language it is important to have mechanisms to specify a variety of views. Obviously, these views are components of the language, that is, they are a specific part of a diagrammatic language.

In our approach, a view of the diagrammatic language DL($\boldsymbol{\mathcal{S}}$) over a model M∈DL($\boldsymbol{\mathcal{S}}$) is a model $\boldsymbol{\mathcal{V}}$∈DL($\boldsymbol{\mathcal{S}}$) with the property that there is a natural injective transformation τ:$\boldsymbol{\mathcal{V}}$→M. It is obvious that the model $\boldsymbol{\mathcal{V}}$ represents a coherent part of the model M.

To manage and manipulate views, we need two very important operations with views, namely, the intersection of two views that allows the visualization of their common part and the union of two views that allows the visualization of the content that cumulates the two views.

Example 2. In Figure 3 and Figure 4 we have two different views of the model in Figure 2.

In our approach we will use pullback for the intersection operation and pushout for the union operation. Pullback is the limit of a co-span in category theory and represents the generalized intersection of two objects over a common object [1]. Pushout is the co-limit of a span in category theory and represents the generalized joining of two objects over a common sub-object [1].

Let M∈DL($\boldsymbol{\mathcal{S}}$) and $\boldsymbol{\mathcal{V}}_1$, $\boldsymbol{\mathcal{V}}_2$ ∈DL($\boldsymbol{\mathcal{S}}$) be two views and τ_1,τ_2 be the two natural injective transformations τ_1: $\boldsymbol{\mathcal{V}}_1$ →M and τ_2: $\boldsymbol{\mathcal{V}}_2$→M. The pullback of $\boldsymbol{\mathcal{V}}_1$ with $\boldsymbol{\mathcal{V}}_2$ is the categorical limit of the co-span, $\boldsymbol{\mathcal{V}}_1 \xrightarrow{\tau_1}M\xleftarrow{\tau_2}$ $\boldsymbol{\mathcal{V}}_2$, which is the view of $\boldsymbol{\mathcal{V}}$, which makes the diagram in Figure 5 commutative.

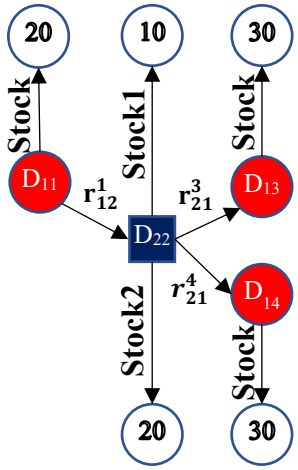


Figure 3. A view instance of the model in Figure 2.

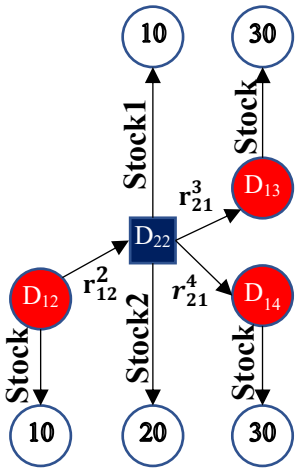


Figure 4. A view instance of the model in Figure 2.

It is worth noting that the pullback preserves injective homomorphisms [1], that is, in our case, it transmits the injectivity property to the resulting homomorphism $\tau:\boldsymbol{\mathcal{V}}\rightarrow M$, where $\tau = \tau_1 o \iota_1 = \tau_2 o \iota_2$.

Let $M \in DL((\boldsymbol{\mathcal{S}})$ and $\boldsymbol{\mathcal{V}}_0, \boldsymbol{\mathcal{V}}_1, \boldsymbol{\mathcal{V}}_2 \in DL(\boldsymbol{\mathcal{S}})$ be three views and τ_0, τ_1, τ_2 be the three natural injective transformations τ_0: $\boldsymbol{\mathcal{V}}_0 \rightarrow M$, τ_1: $\boldsymbol{\mathcal{V}}_1 \rightarrow M$ and τ_2: $\boldsymbol{\mathcal{V}}_2 \rightarrow M$. If the view $\boldsymbol{\mathcal{V}}_0$ is a common subobject of the views $\boldsymbol{\mathcal{V}}_1$ and $\boldsymbol{\mathcal{V}}_2$, then the pushout of $\boldsymbol{\mathcal{V}}_1$ with $\boldsymbol{\mathcal{V}}_2$ is the categorical co-limit of the span, $\boldsymbol{\mathcal{V}}_1 \xleftarrow{\iota_1} \boldsymbol{\mathcal{V}}_0 \xrightarrow{\iota_2} \boldsymbol{\mathcal{V}}_2$, which is the view $\boldsymbol{\mathcal{V}}$, which makes the diagram in Figure 6 commutative. In this span ι_1 and ι_2 are the inclusion applications.

It is worth noting that pushout preserves injective homomorphisms [1], that is, in our case, it transmits the injectivity property to the resulting homomorphism $\mu:\boldsymbol{\mathcal{V}}_0 \rightarrow \boldsymbol{\mathcal{V}}$, where $\mu = \mu_1 o \iota_1 = \mu_2 o \iota_2$. Therefore, the resulting model $\boldsymbol{\mathcal{V}}$ is also a view of M.

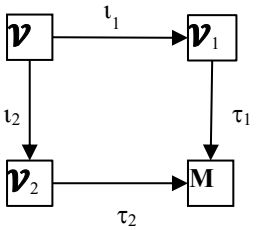


Figure 5. Pullback diagram

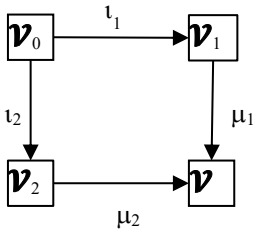


Figure 6. Pushout diagram

Example 3. In Figure 7 we have the pullback of the views in Figures 3 and 4, and in Figure 8 we have the pushout of these models.

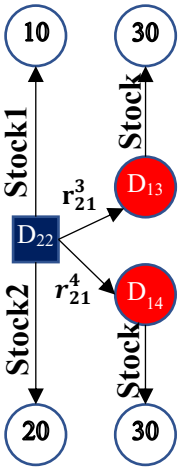


Figure 7. Pullback of the views of Figure 3 and 4

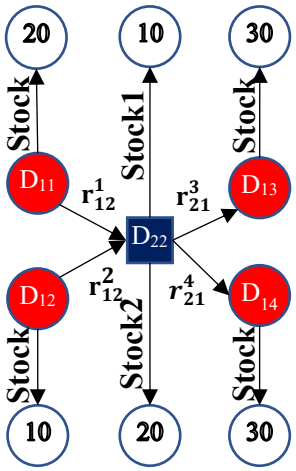


Figure 8. Pushout of the views of Figure 3 and 4

4 Integrating multi-view modeling facilities into diagrammatic languages

In the previous section we saw that a view of a model M is another model $\boldsymbol{\mathcal{V}}$ in the diagrammatic language that maps via an injective homomorphism to M. This is a theoretically satisfactory definition but does not provide us with any solution to actually find these views. In the remainder of this section, we will present a practical way to construct a satisfactory set of views for a given model. Moreover, this mechanism can be implemented generically, at the level of a modeling language, and then used for each model separately. The categorical mechanisms that underlie our approach are representable functors and the Yoneda lemma.

If $\boldsymbol{\mathcal{C}}$ is a category then there is a set of special functors defined on $\boldsymbol{\mathcal{C}}$ with values in Set that are completely determined by the objects in $\boldsymbol{\mathcal{C}}$. More precisely, for every object C∈ob($\boldsymbol{\mathcal{C}}$) there is a set-valued functor represented by C. This functor is Hom$_{\boldsymbol{\mathcal{C}}}$(C,-): $\boldsymbol{\mathcal{C}}$→Set, and is defined as follows: For every object A∈$\boldsymbol{\mathcal{C}}$, we have Hom$_{\boldsymbol{\mathcal{C}}}$(C,-)(A)= Hom$_{\boldsymbol{\mathcal{C}}}$(C,A) and for every arrow f:A→B, we have Hom$_{\boldsymbol{\mathcal{C}}}$(C,-)(f)=Hom$_{\boldsymbol{\mathcal{C}}}$(C,f) where Hom$_{\boldsymbol{\mathcal{C}}}$(C,f)(g)=fog.

We will denote the functor corresponding to the object C∈$\boldsymbol{\mathcal{C}}$ by Y_C. Therefore, we have Y_C(A)=Hom$_{\boldsymbol{\mathcal{C}}}$(C,A) for all objects A∈ob($\boldsymbol{\mathcal{C}}$) and Y_C(f)= Hom$_{\boldsymbol{\mathcal{C}}}$(C,f), for all arrows f in $\boldsymbol{\mathcal{C}}$. It is obvious that if Hom$_{\boldsymbol{\mathcal{C}}}$(C,A)= Y_C(A), is an object in Set and g∈Hom$_{\boldsymbol{\mathcal{C}}}$(C,A) then Y_C(f)(g) = fog∈Hom$_{\boldsymbol{\mathcal{C}}}$(C,B)= Y_C(A) which is an object in Set.

If C is an object from category $\boldsymbol{\mathcal{C}}$, then the functor Y_C is called the functor represented by C and any functor isomorphic to Y_C is called a representable functor.

In order to specify the representable functors, in the case of models, for the sketch $\boldsymbol{\mathcal{S}}$, we will consider the free category generated by the graph $\boldsymbol{\mathcal{G}}$ of the sketch. It is known that a graph can generate a category called free category, which has as objects the nodes of the graph and as arrows the arcs of the graph. The identities are given by the identity arcs attached to each node and the composition of the arrows is given by the concatenation of the paths in the graph.

Example 4. We consider the free category generated by the sketch graph $\boldsymbol{\mathcal{S}}$ from Figure 1. For a more concise presentation we will denote it by cu f^n=f o f o…o f, i.e. f^0=id and f^n=f^{n-1} o f. With this notation the functor represented by T_1, is Y_{T_1}=Hom(T_1,-) defined on components.

1) Each object in $\boldsymbol{\mathcal{G}}$ will be mapped to a set of arcs in the Set category:

$Y_{T_1}(T_1)=\{id_{T_1}\}\cup\{(g_1 \text{ o } f_1)^n, \mid n\geq 0\}\cup\{(g_2 \text{ o } f_1)^n \mid n\geq 0\}$,

$Y_{T_1}(T_2)=\{f_1, \text{Stock}\}\cup\{(f_1 \text{ o } g_1)^n \mid n\geq 0\}\cup\{(f_1 \text{ o } g_2)^n \mid n\geq 0\}$,

$Y_{T_1}(T_3)=\{\text{Stock}, \text{Stock1 o } f_1, \text{Stock2 o } f_1\}$.

2) Each arc in $\boldsymbol{\mathcal{G}}$ will be mapped to a function that maps functions to functions in the Set category as follows:

The function $Y_{T_1}(f_1)$=Hom(T_1,f_1): $Y_{T_1}(T_1)\rightarrow Y_{T_1}(T_2)$ maps each function g∈$Y_{T_1}(T_1)$ to the function $Y_{T_1}(f_1)(g)=f_1$ o g.

The function $Y_{T_1}(g_1)$=Hom(T_1,g_1): $Y_{T_1}(T_2)\rightarrow Y_{T_1}(T_1)$ maps each function g∈$Y_{T_1}(T_2)$ to the function $Y_{T_1}(g_1)(g)=g_1$ o g.

The function $Y_{T_1}(g_2)=\text{Hom}(T_1,g_2)$: $Y_{T_1}(T_2)\rightarrow Y_{T_1}(T_1)$ maps each function $g\in Y_{T_1}(T_2)$ to function $Y_{T_1}(g_2)(g)=g_2 \circ g$.

The function $Y_{T_1}(\text{Stock})=\text{Hom}(T_1,\text{Stock})$: $Y_{T_1}(T_1)\rightarrow Y_{T_1}(T_3)$ maps each function $g\in Y_{T_1}(T_1)$ to function $Y_{T_1}(\text{Stock})(g)=\text{Stock} \circ g$.

The function $Y_{T_1}(\text{Stock1})=\text{Hom}(T_1,\text{Stock1})$: $Y_{T_1}(T_2)\rightarrow Y_{T_1}(T_3)$ maps each function $g\in Y_{T_1}(T_2)$ to function $Y_{T_1}(\text{Stock1})(g)=\text{Stock1} \circ g$.

The function $Y_{T_1}(\text{Stock2})=\text{Hom}(T_1,\text{Stock2})$: $Y_{T_1}(T_2)\rightarrow Y_{T_1}(T_3)$ maps each function $g\in Y_{T_1}(T_2)$ into the function $Y_{T_1}(\text{Stock2})(g)=\text{Stock2} \circ g$.

We note that in the case of a diagrammatic language a functor represented by a certain type represents an abstract model that generically specifies the attributes of the model. We will show in this paper that these functors become views for the concrete models specified by the sketch $\boldsymbol{\mathcal{S}}$. For this we will take advantage of one of the most important results in category theory, namely Yoneda's lemma.

Intuitively, Yoneda's lemma says that any object in a category is completely determined, up to an isomorphism, by its interaction through arrows with the other objects of the category, that is, by the structure of the category.

Specifically, Yoneda's lemma states that if $\boldsymbol{\mathcal{C}}$ is a category, C an object of $\boldsymbol{\mathcal{C}}$, and Ψ:C→Set a set-valued functor, then there is a bijection between the set of homomorphisms from Y_C to Ψ and the set $\Psi(C)$, namely $\text{Hom}(Y_C,\Psi)\xrightarrow{\cong}\Psi(C)$ [5] [1].

We can see that the lemma refers to natural transformations and says that defining a natural transformation from Y_C to Ψ, is equivalent to choosing a concrete element from the set $\Psi(C)$. In the case of models, if $M:\boldsymbol{\mathcal{G}}\rightarrow$Set, and C an object of the free category generated by $\boldsymbol{\mathcal{G}}$, then for each element $c\in M(C)$ there is a unique injective natural transformation $h_c:Y_C\rightarrow M$. Therefore, Y_C is a generic view for the model M.

Example 5. The generic models specified by the sketch $\boldsymbol{\mathcal{S}}$ and the functor Y_{T_1} have the diagrammatic form in Figure 9. We now consider again the concrete model in Example 1, where: $M(T_1)=\{D_{11},D_{12},D_{13},D_{14}\}$; $M(T_2)=\{D_{21},D_{22}\}$; $M(T_3)=\{10,20,30\}$; $M(f_1)=f_1$; $M(g_1)=g_1$; M(Stock)= Stock; M(Stock1)= Stock1; M(Stock2)= Stock2.

We observe that $M(T_1)$ has four elements and therefore it results, according to Yoneda's lemma, that we will have four views of type Y_{T_1}, one for each value given to $x\in M(T_1)$, in the generic model in Figure 9. Thus, for example, for $x=D_{11}$, we have the view in Figure 3, for $x=D_{12}$, we have the view in Figure 4.

Multi-view modeling does not only involve decomposing a model into several sub-models, but also a rigorous mechanism for managing the relationships between these sub-models so that the integrated model remains consistent. As we can see in our approach, the fact that each view is in turn a model of the categorical sketch guarantees the consistency of the integrated model and ensures, at the same time, great flexibility for the sub-models.

The example we used to illustrate our demersal is a very simple one. Obviously, in relation to multi-view modeling, the separation of concerns is also pursued and therefore views can represent not only structural sub-models but also behavioral or functional sub-models. Of course, the approach in these cases does not differ much in the context in which all these aspects will be defined by a categorical sketch and Set-value functors.

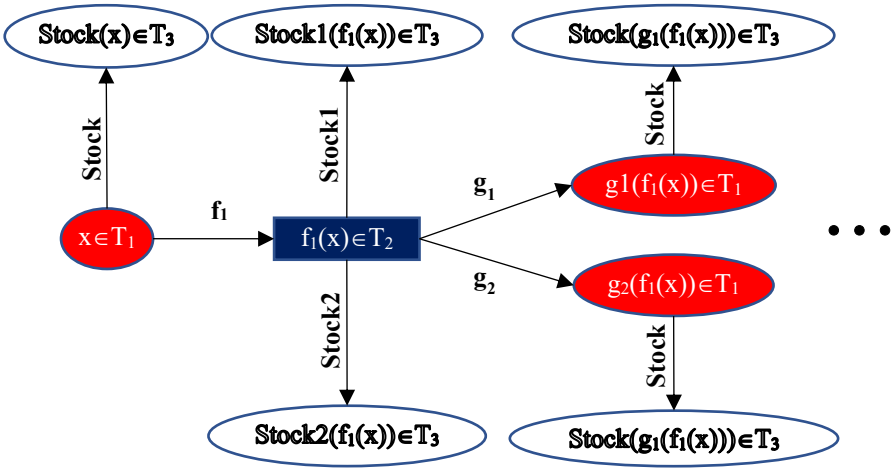


Figure 9. A Y_{T_1} instance of the sketch in Figure 1.

5 Observations and conclusions

Multi-view modeling provides a robust conceptual framework for modeling, specifying, and analyzing complex systems by separating concerns and integrating multiple perspectives coherently. Multi-view modeling is an essential pillar in issues related to the decomposition of systems into subsystems, subsystem interoperability, consistency between sub-models, and integration of sub-models of a complex system.

Modern systems modeling, characterized by structural complexity and heterogeneity, requires the use of powerful formal mechanisms that provide advanced modeling and analysis facilities. In this context, integrated MDE methodologies require more powerful mathematical tools than graph theory. Category theory can provide the context and mathematical mechanisms necessary to solve many modeling problems.

Along with multi-level modeling and co-simulation, multi-view modeling emerges as a complementary pillar of an integrated MDE methodology. Therefore, the use of categorical mechanisms to facilitate multi-view modeling is welcome.

This paper demonstrates that defining diagrammatic languages through metamodels represented by categorical sketches allows the use of important categorical concepts such as representable functors and Yoneda's lemma to endow the diagrammatic language with special facilities for specifying the views of a model.

The use of categorical mechanisms for defining diagrammatic languages through metamodels allows the use of all important theoretical results such as universal constructions, representable functors, Yoneda's lemma and many others obtained in decades of research. The synergy between these mechanisms generates a unified modeling and analysis framework, capable of managing the complexity of cyber-physical systems, flexible manufacturing processes and other domains where interactions between components are critical.

References

[1] Michael Barr, Charles Wells, “Category Theory for Computing Science”, Reprints in Theory and Applications of Categories, No. 22, (2012).
[2] Walters R. F. C., “Categories and Computer Science”, Cambridge Texts in Computer Science, Edited by D. J. Cooke, Loughborough University, (2006).
[3] Hartmut Ehrig, Karsten Ehrig, Claudia Ermel and Ulrike Prange, “Consistent integration of models based on views of meta models in Formal Aspects of Computing”, 22, 327–344, Springer, (2010).
[4] Hartmut Ehrig, Claudia Ermel, Ulrike Golas, Frank Hermann, “Graph and Model Transformation General Framework and Applications”, Springer-Verlag Berlin Heidelberg, (2015).
[5] David I. Spivak, “Category Theory for the Sciences”, The MIT Press Cambridge, Massachusetts London, England, Massachusetts Institute of Technology, (2014).
[6] 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).
[7] Daniel-Cristian Craciunean, “Categorical Mechanisms in Multi-level Modeling Methods”, Publishing House of the "Lucian Blaga" University of Sibiu, (2023).
[8] Daniel-Cristian Crăciunean, D. Karagiannis, “A categorical model of process cosimulation”, Journal of Advanced Computer Science and Applications (IJACSA), 10, (2019).
[9] D. Karagiannis, H.C. Mayr, J. Mylopoulos, “Domain-Specific Conceptual Modeling Concepts, Methods and Tools”, Springer International Publishing Switzerland, (2016).
[10] Kühn, E., & Karagiannis, D., “A multi-view modeling approach for complex systems”, Complex Systems Informatics and Modeling Quarterly, (24), 39–60, (2020).
[11] Dominik Bork , Dimitris Karagiannis, Benedikt Pittl, “A survey of modeling language specification techniques”, Information Systems 87 101425, journal homepage: www.elsevier.com/locate/is. , (2020).
[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).
[14] Markus Voelter, “DSL Engineering Designing, Implementing and Using Domain-Specific Languages”, http://dslbook.org, (2010 – 2013)
[15] Zeigler, B. P., Sarjoughian, H. S., & Hammonds, P., “Modeling and Simulation-Based Systems Engineering Handbook”. Springer, (2020).

DOI: 10.2478/ijasitels-2025-0012

Optimization of Romanian Identity Documents Processing

Delia-Victoria Milaşcon[1], Constantin Constantinescu [2], Raluca Brad [3]

[1,2,3]Lucian Blaga University of Sibiu, Department of Computer Science and Electrical Engineering

Abstract

This paper presents the development of an application based on computer vision and deep learning aimed at the automatic recognition of data from Romanian national identity documents. Its primary goal is to optimize the processing of such documents within public and private institutions. Developing such a system involves several challenges. These include the variability in the visual appearance of documents caused by wear and tear, scanning angles, camera quality, and ambient lighting conditions; the difficulty of accurately recognizing Romanian diacritics; and the correct identification of textual fields based on their positioning within the document. The system must be sufficiently robust to adapt to these variations while maintaining a high level of accuracy.

Keywords: optical character recognition, ID cards, passport, identification

1 Introduction

The processing of identity documents in Romania, encompassing optical character recognition (OCR), data extraction, authenticity verification, and integration with digital services, has undergone significant transformation as of November 2025. Driven by EU regulations like eIDAS 2.0 [1] and national initiatives under the Recovery and Resilience Plan (PNRR) [2], optimizations focus on speed, accuracy, security, and fraud prevention. These advancements address rising fraud (up 7% year-on-year in card transactions [3]) and low digital service adoption (24% for online government services [4]). Key enablers include AI-enhanced OCR, biometric chips, and modular digital ID platforms, reducing processing times from minutes to seconds while achieving over 99% accuracy in data entry [5].

Recent advances in document layout analysis, powered by deep learning, have transformed OCR-based document processing from a theoretical concept into a practical reality. Central to this progress is DocLayNet [7], a dataset comprising over 100,000 annotated pages across invoices, forms, and identity documents. With 11 semantic classes, including text blocks, titles, tables, and figures models, trained on this corpus learn to interpret documents as structured hierarchies rather than flat pixel arrays. Transformer-based architectures like LayoutLMv3 [8] with various fine tunings [9] have achieved mean Average Precision scores exceeding 0.90 on benchmark datasets of 5000 anonymized eID scans collected during a 2024–2025 pilot.

However, standard models encounter challenges when applied to Romanian identity documents, such as the standard ID or even the CEI (eID), which feature bilingual text

(Romanian and English), variable font styles, and densely packed fields. These issues are present for other countries as well. Nguyen-Trong [10] developed a complete system for Vietnamese ID card information extraction with multiple deep learning models, while authors of this paper [11] propose and algorithm for Indonesian ID cards. And the list can go on.

Accurate field detection alone is insufficient. Mobile-captured images often suffer from perspective distortion, motion blur, or poor lighting. To mitigate this, a rectification technique must be incorporated. This system [12] proposes an unsupervised geometric rectification technique, which reconstructs frontal, flat views of documents using detected text lines and border cues. This preprocessing reduces the Character Error Rate in downstream OCR from 4.2% to 0.8%.

Another major issue is the practical deployment of the solutions. Custom-trained architectures face substantial barriers: computational overhead requiring GPU infrastructure, extensive labeled training datasets (often 5,000+ documents per document type), complex fine-tuning workflows, and integration challenges with existing institutional systems. Some real-time applications exist, for example the RT-DETR detector [13], a transformer-based architecture optimized for speed, processes both sides of an ID card in under 50 ms on a mid-range smartphone GPU, enabling deployment in mobile applications like ROeID.

Building on all of the above, we propose a solution that achieves great results in localizing critical fields Romanian IDs, including the CNP (13-digit personal identifier), full names with diacritics, addresses aligned with SIRUTA codes, and validity dates.

This paper presents an intelligent hybrid deep learning architecture specifically engineered for automated processing of Romanian identity documents. Rather than training custom models from scratch, our system optimally configures and orchestrates state-of-the-art pre-trained deep learning engines—PaddleOCR's DBNet [14] for neural text detection and CRNN [15] for character recognition, complemented by Tesseract's LSTM-based [16] enhancement for Romanian diacritics—through sophisticated spatial relationship analysis and domain-specific parameter optimization. The system addresses key challenges in Romanian document processing: visual variability from wear and lighting conditions, accurate diacritic recognition (ă, â, î, ș, ț) essential for legal validity, field localization within complex layouts, and comprehensive validation through Machine Readable Zone (MRZ) integration.

The developed architecture achieves great accuracy on both identity cards and passports through intelligent configuration of complementary deep learning models, advanced geometric preprocessing, and multi-layer validation mechanisms. Processing completes in under 3 seconds on standard institutional hardware without requiring GPU infrastructure, custom model training, or cloud dependencies. Our results demonstrate that strategic optimization of pre-trained deep learning models—when engineered for domain-specific requirements—can achieve accuracy comparable to custom-trained architectures while offering the possibility of immediate deployment and minimal resource requirements, providing a production-ready solution suitable for Romanian institutional environments.

2 The Proposed System Architecture

The developed application is a complete automatic recognition system designed to extract relevant data from Romanian official identity documents. It aims to automate data entry and digitalization by processing document images and generating pre-filled official forms.

The processing pipeline is shown in figure 1 and includes:

1. Image Upload – Users load an image of the document (scanned or photographed).
2. Document Detection – The system isolates the document's contour within the image, correcting distortions.
3. Orientation Correction – It tests four possible rotations (0°, 90°, 180°, 270°) using OCR confidence and Romanian keywords to determine the correct orientation.
4. Image Preprocessing – Noise removal, contrast enhancement, and adaptive binarization prepare the image for text recognition.
5. Text Recognition (OCR) – A hybrid approach uses PaddleOCR for speed and Tesseract for refinement—especially improving Romanian diacritics.
6. Data Association and Validation – Extracted text is mapped to document fields, verified through pattern recognition (CNP, serial number, etc.), and validated for format correctness.
7. User Review and Document Generation – The user can manually correct data before the system automatically generates pre-filled Word documents using the extracted information.

The development of the system was based on a selection of technologies, each chosen for its specific characteristics that contribute to the overall functionality of the application. Python 3.12.6 was selected as the main programming language due to its rich ecosystem of libraries for image processing and machine learning.

CustomTkinter was chosen for developing the graphical user interface. This modern library, built on top of the traditional Tkinter, offers contemporary visual components with a professional look. CustomTkinter allows for the creation of an intuitive interface that adheres to modern design principles while providing the necessary flexibility for displaying and editing complex data extracted from documents.

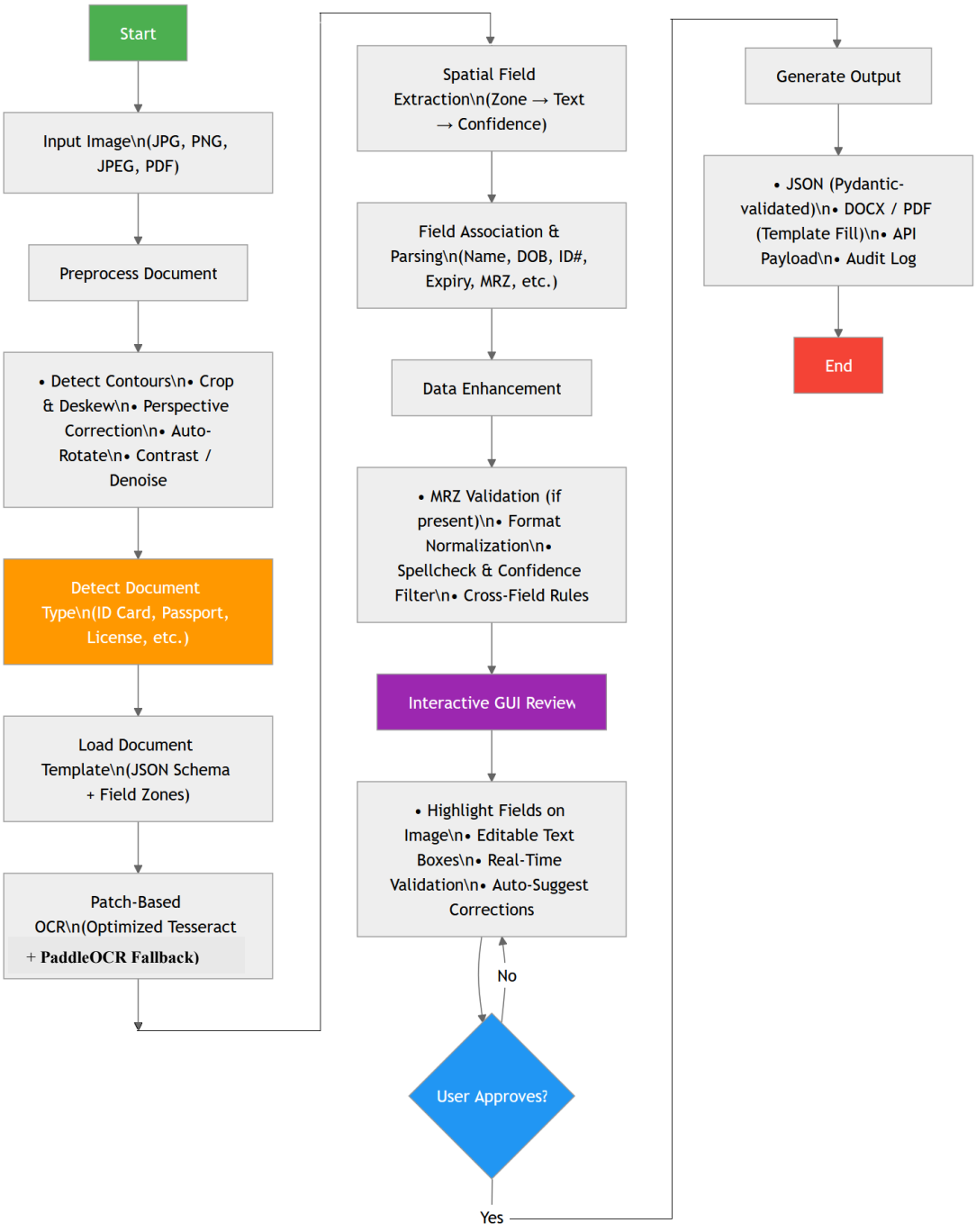


Figure 1. Processing flowchart

For image processing, the application relies on OpenCV, the reference library in the field of computer vision. OpenCV provides optimized implementations for fundamental image processing operations. Complementarily, PIL/Pillow is used for high-level image manipulation and for integration with the Tkinter graphical interface.

The central OCR component is provided by PaddleOCR, a modern framework that delivers excellent performance for text recognition in multiple languages, including Romanian. PaddleOCR uses deep neural network architectures for text detection and recognition, offering an optimal balance between accuracy and speed. To refine the

results, particularly for characters with diacritics, the system also integrates Tesseract OCR, the open-source OCR engine originally developed by HP and currently maintained by Google.
NumPy forms the foundation for efficient numerical operations, being extensively used in pixel matrix processing and in calculations required for various transformations. For document generation, the application uses python-docx, a Python library that allows programmatic manipulation of Microsoft Word documents while preserving the formatting and structure of templates.

3 Image Pre-processing Steps

The process of detecting a document within an image represents the first major technical challenge of the system. In practice, users may provide images in which the document appears under various conditions: it may be photographed on a complex background, occupy only a portion of the image, or be captured at an oblique angle. The function detect_document_contour() implements a robust algorithm that automatically identifies and extracts the document from this variable context.
The algorithm begins by converting the image to grayscale, an essential step that reduces computational complexity and removes color variations that are irrelevant for contour detection. This conversion simplifies the image from three color channels (BGR) to a single intensity channel, while preserving essential information about edges and contrasts.
The next step involves applying a Gaussian filter [17] to reduce noise. This filter is essential because image noise can generate false contours that interfere with the correct detection of the document edges. A 5x5 Gaussian kernel performs smoothing that preserves important edges while eliminating minor intensity variations. Actual edge detection is performed using the Canny algorithm [18], applied with three different threshold sets to cover a wide range of image conditions. This multi-threshold approach ensures robust detection in the presence of varying contrast and lighting.
For each configuration, the algorithm applies morphological operations to connect fragmented edges. Dilation followed by a morphological closing operation helps form complete contours even when the initial detection produces broken edges. The evaluation of detected contours represents the most sophisticated aspect of the algorithm. For each identified contour, the system calculates a score based on multiple criteria reflecting the likelihood that the contour corresponds to an identity document:

- Aspect ratio criterion evaluates how close the detected shape is to the standard proportions of a Romanian ID card (85.6 mm × 54 mm). Contours with aspect ratios near this value receive higher scores.
- Rectangular approximation criterion checks how well the contour can be approximated by a polygon with few vertices. A real document should be approximately rectangular.

After selecting the optimal contour, the algorithm adds generous margins (10% of the document's dimensions) to ensure that no important content is lost during cropping. This conservative approach prioritizes preserving the integrity of the information over achieving a perfectly tight crop. Figure 2 shows a Romanian ID card in different phases of preprocessing.

Figure 2. Document detection and extraction from an image. Original image (1); Document detection (2); Edge detection (3, 4, 5); Final result (6)

4 Optical Character Recognition Optimizations with Tesseract

The development of optimizations for Tesseract has undergone a significant evolution, culminating in the current implementation that prioritizes the preservation of detail over aggressive processing. This section describes the refined approach that maximizes Tesseract's performance for Romanian documents.

Early versions of the system implemented a complex preprocessing pipeline that included intensive bilateral filtering, adaptive histogram equalization (CLAHE), morphological opening and closing operations, and even image sharpening techniques. While theoretically sound, this approach produced disastrous results in practice.

The fundamental problem was that each processing step, although improving the visual appearance of the image for the human eye, actually degraded the information useful for OCR. Excessive sharpening transformed diacritics into unrecognizable artifacts. Aggressive morphological operations incorrectly merged or split character components. Extreme binarization completely eliminated the subtle nuances necessary to distinguish between similar characters.

The final implementation adopts the principle of elegant simplicity. Rather than attempting to "enhance" the image, the system focuses on providing Tesseract with a representation that is as faithful as possible to the original text, with only the minimal cleaning necessary to remove obvious noise. This minimalist approach not only produces superior results but also reduces processing time and code complexity. Removing unnecessary preprocessing steps makes the system more robust and predictable.

The system implements a decision algorithm designed to select the optimal OCR result by intelligently combining outputs from two complementary engines: the primary OCR module and Tesseract. This mechanism goes beyond a simple "best match" approach,

instead using multiple criteria to evaluate the quality of the extracted text and determine which version should be retained for each field.

At the core of this decision process is the diacritics criterion. Romanian documents often contain characters with diacritics (ă, â, î, ș, ț), which are critical for accurate recognition of names, locations, and other key fields. The system counts all variants of diacritics in both the primary OCR and Tesseract outputs, considering uppercase, lowercase, and older cedilla-based forms. This allows the algorithm to prioritize results that preserve these essential linguistic details.

Complementing the diacritics analysis is a similarity criterion. To avoid selecting results that deviate substantially from the original, the algorithm calculates textual similarity between the two outputs. Both texts are normalized to lowercase, ensuring that differences in capitalization do not influence the comparison. This step prevents the system from mistakenly favoring a result that introduces unrelated or erroneous text.

The decision logic itself follows a clear hierarchy. Results from Tesseract that contain more diacritics than the original are given the highest priority, reflecting the importance of preserving linguistic accuracy. Secondary consideration is given to outputs with at least 60% similarity to the original and a diacritic count equal to or higher than that of the primary OCR. For certain fields, such as residence, additional rules apply: if both outputs have the same number of diacritics, the system selects the more complete result, recognizing that Tesseract may capture extra details—such as multi-line addresses—that the primary OCR engine might miss.

To maintain high data quality, the system incorporates several safety validations. Short text segments (fewer than two characters), text resembling MRZ zones, segments containing noise characters, or outputs with similarity below 50% are automatically rejected. This ensures that the algorithm does not introduce errors while attempting to improve results.

The decision-making process is further refined with field-specific handling. Different types of fields have unique requirements: for instance, the place of birth field uses a relaxed similarity threshold and employs fallback logic for detecting counties, while the residence field is processed with multi-line support and additional address-specific validations.

Finally, the algorithm includes an implicit safety mechanism. In cases where none of the improvement criteria are clearly met, the original result from the primary OCR engine is retained. This conservative approach prevents quality degradation in ambiguous scenarios, ensuring that the system avoids introducing new errors while striving to enhance the accuracy of extracted information.

Through this hybrid, criterion-driven process, the system achieves a careful balance between precision and reliability, leveraging the strengths of Tesseract for fine-grained details such as diacritics, while maintaining the robustness and consistency of the primary OCR engine.

This validation function represents a significant innovation in processing Romanian documents. By intelligently combining results from two complementary OCR engines, the system achieves accuracy rates superior to either technology used individually. Practical benefits include correct recognition of Romanian names with diacritics, precise identification of cities and addresses, minimization of transcription errors in critical fields, and optimized processing of multi-line addresses.

The implemented safety mechanisms ensure preservation of the original result in ambiguous cases, validation against MRZ interference, filtering of noise characters, and similarity verification to maintain semantic consistency. This hybrid approach

leverages the strengths of both OCR engines: the speed and robustness of the primary engine combined with Tesseract's precision for special characters, resulting in a system that excels in processing Romanian documents.

5 Experimental Approaches and Challenges with PaddleOCR

Several experimental strategies were explored to improve OCR performance for Romanian documents.

In our first approach we used the EAST text detector [19] to precisely localize text regions, followed by applying PaddleOCR only in these regions. The hypothesis was that EAST, specialized in text detection, could provide more accurate boundaries than PaddleOCR's integrated detector. However, this approach proved problematic:

- Implementation complexity increased due to coordinating two separate systems.
- EAST did not outperform PaddleOCR for official documents.
- Detection errors propagated and amplified through the two-step process.
- Overall performance was inferior to direct PaddleOCR usage.

Applying the intensive preprocessing pipeline developed for Tesseract to images for PaddleOCR also failed. Modern neural network-based PaddleOCR performs best on natural images, and excessive preprocessing removed subtle textures and gradients crucial for convolutional recognition. In many cases, text detection failed entirely after preprocessing.

Attempts to train PaddleOCR models to improve recognition of Romanian diacritics were abandoned due to high computational and time requirements and the complexity in creating a suitable training dataset for official documents, while the hybrid approach using Tesseract proved more practical and effective for the application.

These experiments highlighted that direct PaddleOCR usage on minimally processed natural images, combined with hybrid integration of Tesseract for diacritics, provides the most practical and reliable solution.

6 Machine Readable Zone Processing

The Machine Readable Zone (MRZ) is a standardized component located at the lower portion of modern identity documents. It encodes essential information in a format optimized for automated reading. In Romanian identity documents, the MRZ typically consists of two or three lines containing uppercase letters, digits, and the filler character '<', which separates fields and fills unused space. The rigid standardization of the MRZ makes it an invaluable resource for validating and supplementing data extracted through conventional OCR methods.

The first step in MRZ processing is to identify its spatial location within the document. The system analyzes the vertical distribution of all detected text elements, determining that the MRZ consistently occupies the bottom 25% of the document. A vertical threshold is computed so that all text elements below this line are considered potential MRZ components. This approach is robust across document types and image qualities, relying on the fixed relative position of the MRZ within the overall document layout.

Once the region of interest is established, several criteria are applied to identify actual MRZ lines:

1. Presence of '<': This character serves as a field separator, fills empty space to maintain line length, and marks the end of variable fields.
2. Character composition: MRZ lines consist exclusively of uppercase letters and include at least a few digits, distinguishing them from supplementary text in the lower document region.
3. Minimum line length: MRZ lines contain at least 30 characters, adhering to international formatting standards.

For Romanian ID cards, the first MRZ line always begins with the prefix "IDROU" (ID Romania), followed by the surname and given name of the holder, separated by '<<' sequences. Subsequent lines encode numeric identifiers, document type, country codes, dates, and check digits in fixed positions.

Before parsing, the extracted MRZ text undergoes cleaning to remove OCR-induced spaces, as the MRZ contains no whitespace. Common OCR errors are also corrected:

- '0' vs. 'O' and '1' vs. 'I' are resolved contextually.
- Missing or extra characters are identified by verifying line length against the standard.

For each document type, patterns based on the MRZ's fixed structure are used to extract fields:

- Names: Sequences between '<<' markers are converted from uppercase MRZ format to properly capitalized names.
- Numeric fields (e.g., CNP): Extracted from known positions, validated for length and format, and converted from YYMMDD to the standard Romanian date format (DD.MM.YYYY), with the century inferred from context.

MRZ data is cross-checked with fields extracted via standard OCR:

- When values match, correctness is confirmed.
- When discrepancies arise, the system applies rules: OCR versions with diacritics are preferred for names, while MRZ values are considered more reliable for numeric codes and standardized fields.

MRZ extraction also enables automatic completion of fields that are missing or poorly captured in visual OCR, such as sex, document expiration date, and country/type codes. This significantly increases the overall success rate of complete data extraction.

The system incorporates strategies for damaged or partially unreadable MRZs:

- If only a portion of the MRZ is legible, data is extracted from intact sections.
- Ambiguous characters are inferred using the known MRZ structure.
- For example, the algorithm can automatically correct the expected "ROU" country code even if OCR introduces errors.

Although MRZs are standardized, minor variations exist across document series and passport types. The system adapts by:

- Using flexible searches for known markers such as "IDROU" or '<<'.
- Automatically detecting document type and applying corresponding parsing rules.

MRZ processing is integral to the OCR system, providing:

- Redundancy: Critical information appears in two formats, enhancing extraction reliability.
- Independent validation: Standardized fonts and format offer a trustworthy verification source.
- Completeness: Fields missing from the visual OCR zone are present in MRZ.
- International compliance: Accurate MRZ parsing ensures compatibility with global document processing systems.

By intelligently integrating MRZ extraction with standard OCR, the system achieves a level of accuracy and completeness that would not be possible using a single method, ensuring reliable recognition of Romanian identity documents across various conditions.

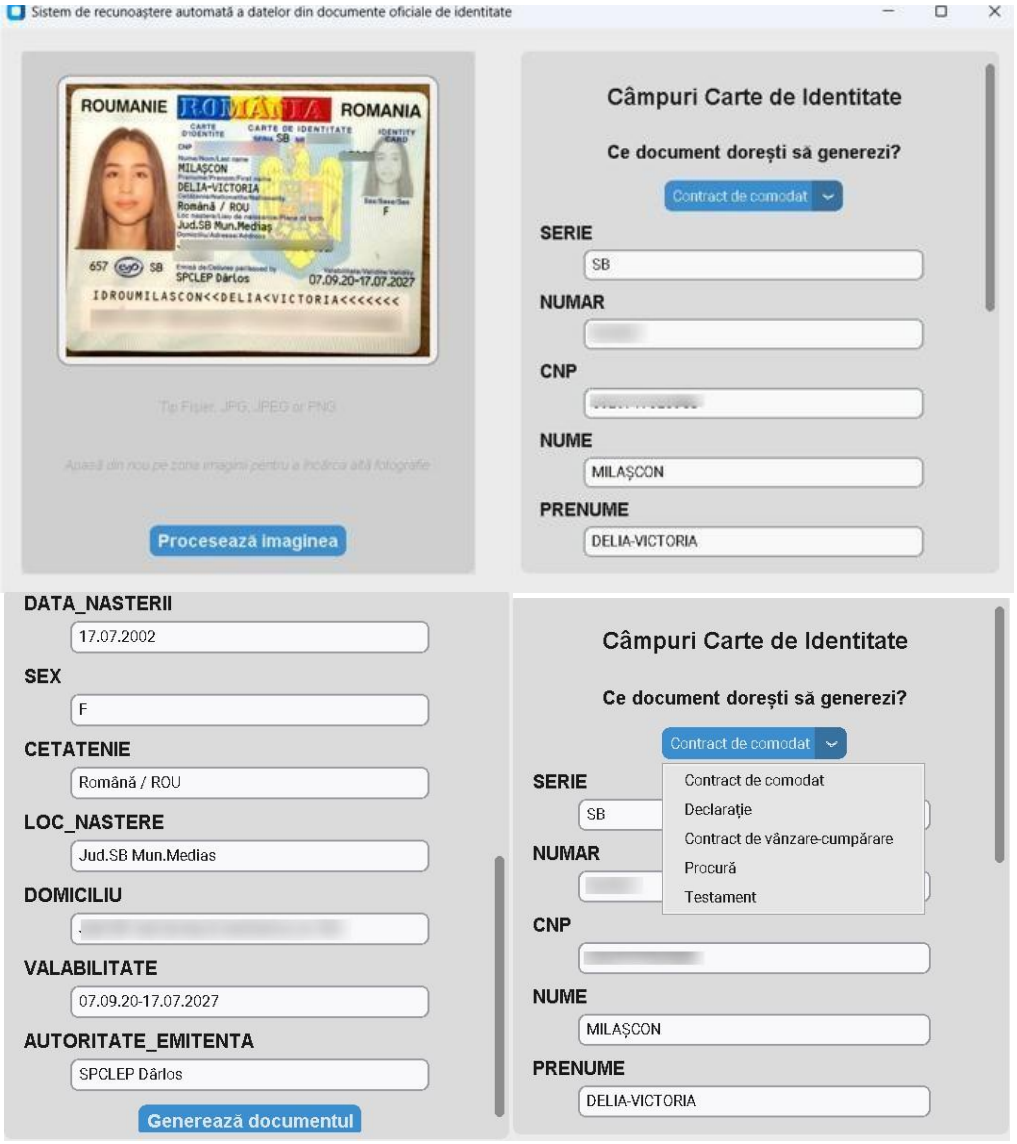


Figure 3. Graphical User Interface

7 Graphical User Interface

The graphical user interface (GUI) of the application was designed following modern UX/UI design principles, focusing on simplicity, clarity, and efficiency as shown in Figure 3. The use of the CustomTkinter library enables the creation of a contemporary-looking interface that integrates naturally with modern operating systems.

The main structure of the interface is divided into two distinct panels, each with a clearly defined role:

• Left Panel (ImageFrame): serves as the image viewing and control area.

It includes:

- a 400x300 pixel canvas area for image display
- informative text with user instructions
- the main “Process Image” button
- visual indicators showing the processing status

• Right Panel (FieldsFrame): implemented as a scrollable frame to accommodate a variable number of fields, depending on the type of document being processed.

The visual design uses a neutral color palette with blue accents for interactive elements. Fonts are chosen for maximum clarity.

8 Results and Performance Evaluation

8.1 Recognition metrics on ID cards and passports

The evaluation of the automatic recognition system for data extracted from Romanian identity cards was conducted on a comprehensive test set, aiming to determine the extraction accuracy for each individual field, as well as the overall performance of the application. The testing included both a quantitative evaluation of the correct recognition rate and a qualitative analysis of the types of errors encountered.

The system was tested on a set of identity card images, each image being processed through the complete pipeline — preprocessing, OCR detection, and validation. The extracted results were automatically compared with a reference dataset (ground truth) to calculate the performance metrics.

The system was evaluated on two separate test sets: 36 identity cards and 17 passports. The results, shown in table 1, demonstrate excellent performance for both document types:

• Identity cards: 98.5% overall accuracy

• Passports: 89.2% overall accuracy

Table 1. Results

Metric	ID Cards	Passports
Overall Accuracy	98.5%	89.2%
Documents Processed	36	17
Perfect Matches	80.3%	78.3%
Minor Errors	17.1%	8.6%
Major Errors	2.5%	13.1%
Best Field	sex (100.0%)	Passport number (100.0%)

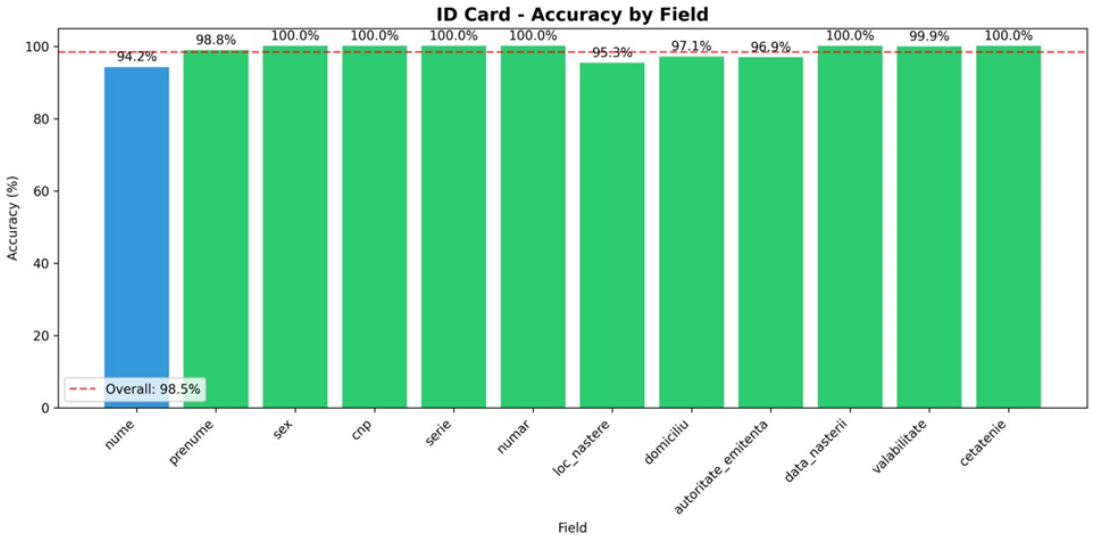


Figure 4. Data extraction accuracy for each field of the identity card

The system achieves perfect performance (100%) for most fields (Figure 4):

- Sex, CNP, series, number: Algorithmic validation ensures maximum accuracy.
- Date of birth, validity, citizenship: Perfect extraction due to standardized formats.
- Last name and first name: 98.7% and 100%, respectively, demonstrating the efficiency of the hybrid approach for handling diacritics.

The only fields with performance below 100% are:

- Last name (98.8%): Minor errors in complex names with multiple diacritics.
- Place of birth (95.3%): Localities with complex spelling.
- Address (97.1%): Long addresses with variable formatting.
- Issuing authority (96.9%): Acronyms and specific formatting styles.

The system's performance varies significantly across the different fields of the identity card, reflecting the varying recognition complexity for each type of information.

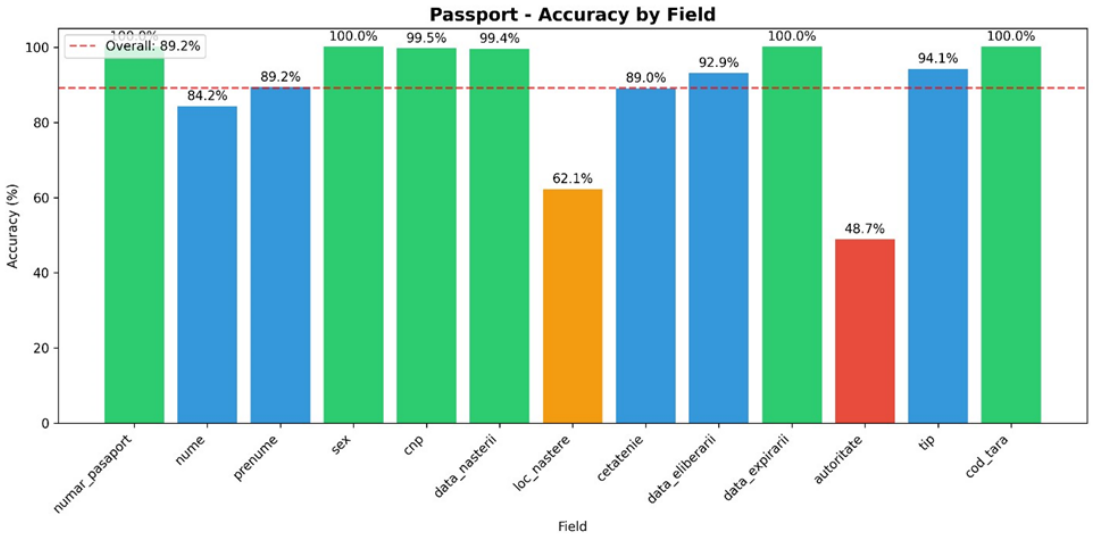


Figure 5. Data extraction accuracy for each field of the passport

The results for passports (Figure 5) show a more varied distribution:

•Excellent performance (>99%): passport_number, sex, date_of_birth, expiration_date, country_code
• Good performance (84–94%): last_name, first_name, citizenship, type
• Moderate performance: place_of_birth (62.1%)
• Low performance: issuing_authority (48.7%)

The *issuing_authority* field represents the main challenge due to its variable position and complex formatting within the passport layout

8.2 Error Distribution Analysis

The system evaluation revealed significant differences in the types and distribution of errors between the two categories of processed documents, graphically represented in figure 6. This detailed analysis provides valuable insights for further system improvements and for understanding the current technological limitations.

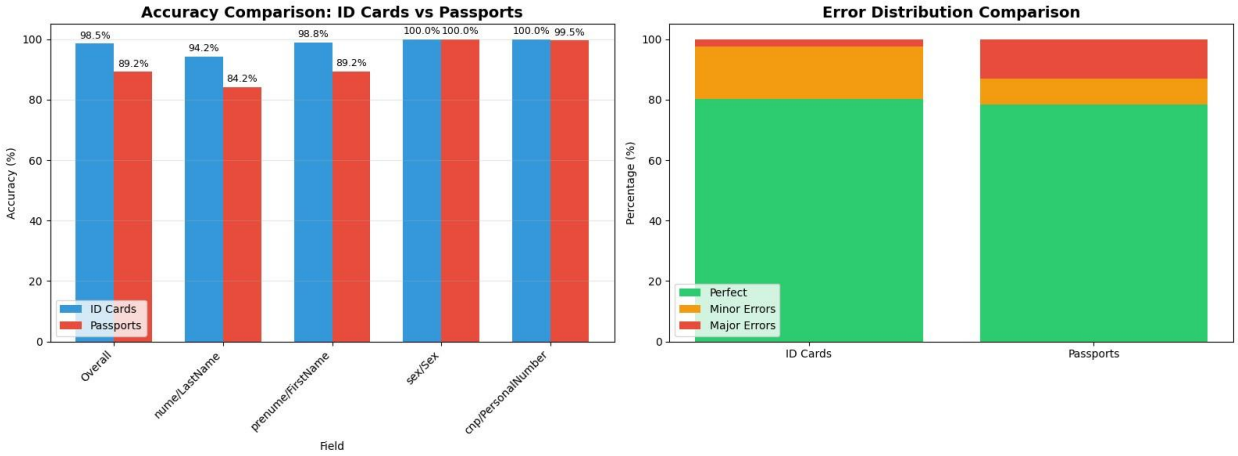


Figure 6. Error distribution comparison

8.2.1 Error Analysis for Identity Cards

Figure 7 shows the overall error distribution reveals a system that performs reliably, with 80.3% of extractions classified as perfect. However, 17.1% contain minor errors, most of which stem from diacritic inconsistencies or spacing issues, while 2.5% fall into the category of major errors, typically caused by complete OCR failures.

Minor errors arise predominantly from spacing and formatting issues, which account for 60% of all minor inaccuracies. These problems are most common in composite fields, where words are frequently concatenated. For instance, entries such as issuing_authority or place_of_birth sometimes appear compressed, producing artifacts like “Jud.IFSat.Vidra(Com.Vidra)” instead of the correctly spaced “Jud. IF Sat. Vidra (Com. Vidra)”. In other cases, simple spaces are misinterpreted by the OCR engine and replaced with special characters such as slashes, generating outputs like “SPCLEP/Vinga”.

The remaining 40% of minor errors are linked to Romanian diacritic challenges, particularly notable in the place_of_birth field. Complex city names—Bârlad, Drobeta-Turnu Severin, Râmnicu Vâlcea—often appear in smaller text sizes on identity documents, making them more susceptible to misinterpretation. Characters such as ț

and ș, with their visually intricate shapes, are at times interpreted as noise by PaddleOCR. Although a hybrid enhancement strategy using Tesseract increases recognition accuracy by 15–20%, some regions of interest are not accurately isolated, limiting the benefits of reprocessing.

Major errors, which make up 2.5% of all extractions, result from more severe breakdowns in the OCR pipeline. These include outputs containing nonsensical, unintelligible character sequences or instances where fields are not detected at all due to unusual text placement or degraded image quality. Together, these issues highlight the key limitations of the current OCR setup and indicate directions for further refinement.

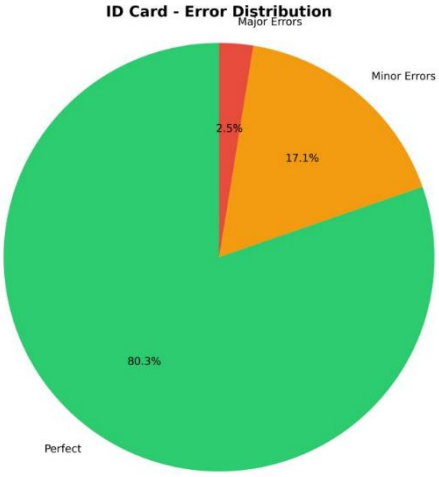


Figure 7. Error distribution for identity cards

8.2.2 Error Analysis for Passports

The analysis of OCR performance on passports (Figure 8) reveals both encouraging results and specific challenges.

Out of all passport extractions, the system achieved the following error distributions:

- 78.3% perfect extractions: All fields were correctly recognized without any errors.
- 8.6% minor errors: Extractions with small, non-critical mistakes that do not substantially affect data usability.
- 13.1% major errors: Extractions with significant issues, potentially rendering the data unreliable.

These minor mistakes rarely compromise the usability of the extracted data but highlight areas where OCR could be refined for improved diacritic handling. While passports generally yield a high rate of perfect extractions due to their standardized, concise layout, OCR performance is occasionally hindered by technical challenges such as fragmented labels, undetected fields, and interference from security features. These findings emphasize the need for adaptive recognition strategies and robust error-handling mechanisms, particularly for critical fields like *place_of_birth* and *issuing_authority*.

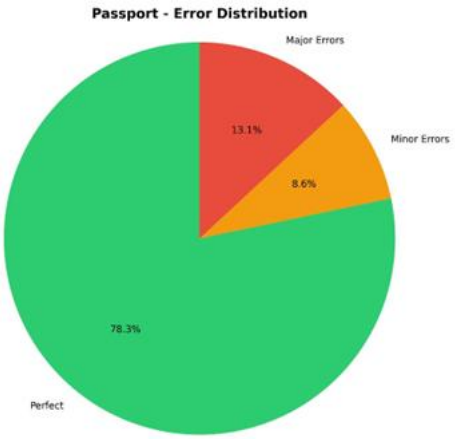


Figure 8. Error distribution for passports

9 Conclusions

With an accuracy of 98.5% for identity cards and 89.2% for passports, the developed system demonstrates the practical viability of an automatic digitization solution specialized for Romanian documents. The excellent performance across critical fields validates the main architectural decisions:

- The hybrid PaddleOCR + Tesseract approach provides the best compromise between speed and accuracy.
- MRZ (Machine Readable Zone) processing adds valuable redundancy and enhances the system's robustness.
- The modular structure allows easy adaptation and expansion.

The system enhanced in recognition of structured fields (CNP, series, number) with very good accuracy, the preservation of Romanian diacritics, achieving higher success rates than generic commercial solutions and automatic generation of pre-filled documents, significantly reducing processing time.

References

[1] European Commission, *eIDAS 2.0 Regulation*, EU 2024/1183, Official Journal of the EU, April 2024. ↩
[2] Ministerul Investițiilor și Proiectelor Europene, *Planul Național de Redresare și Reziliență Componenta C7: Transformare Digitală*, 2025 Update. ↩
[3] *, *European Payments Council*, 2025 Fraud Report, Brussels, October 2025. ↩
[4] *, *Eurostat, Digital Economy and Society Index (DESI)*, Romania 2025, Publications Office of the EU. ↩
[5] *, *BCR Internal Audit Report, AI-OCR Deployment Results Q3 2025*, Bucharest, September 2025. ↩
[6] Zhao Y., Lv W., Xu S., Wei J., Wang G., Dang Q., Liu Y., *DETRs Beat YOLOs on Real-time Object Detection*, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16965–16974, ISBN 978-1-6654-6547-0, Vancouver, Canada, 2024
[7] Pfitzmann B., Auer C., Dolfi M., Nassar A.S., Staar P., DocLayNet: *A Large Human-Annotated Dataset for Document-Layout Analysis*, Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD), 2860–2870, ISBN 978-1-4503-9232-0, Washington DC, USA, 2022.
[8] https://ubiai.tools/the-role-of-layoutlmv3-in-document-layout-understanding/, accessed 4.11.2025

[9] Huang Y., Lv T., Cui L., Lu Y., Wei F, *LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking*, Proceedings of the 30th ACM International Conference on Multimedia (MM), 4083–4092, ISBN 978-1-4503-9203-0, Lisboa, Portugal, 2022.
[10] Nguyen-Trong, Khanh. "An End-to-End Method to Extract Information from Vietnamese ID Card Images." *International Journal of Advanced Computer Science and Applications* 13, no. 3 (2022).
[11] Satyawan, W., Pratama, M. O., Jannati, R., Muhammad, G., Fajar, B., Hamzah, H., ... & Kristian, K. (2019, June). Citizen ID card detection using image processing and optical character recognition. In *Journal of physics: Conference series* (Vol. 1235, No. 1, p. 012049). IOP Publishing
[12] Li H., Wang Y., Chen Y., Liu Z. *Foreground and Text-lines Aware Document Image Rectification*, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 19574–19583, ISSN 2380-7504, Paris, France, 2023.
[13] Li, M., Lv, T., Chen, J., Cui, L., Lu, Y., Florencio, D., Zhang, C., Li, Z., Wei, F, *TrOCR: Transformer-Based Optical Character Recognition with Pre-trained Models,* Proceedings of the AAAI Conference on Artificial Intelligence, 37(11), 13094-13102, ISSN 2374-3468, Washington DC, USA, 2023.
[14] Low, L. M., Salleh, F. H. M., Law, Y. F., & Zakaria, N. Z. (2024). Detecting and recognizing seven segment digits using a deep learning approach. In *ITM Web of Conferences* (Vol. 63, p. 01007). EDP Sciences.
[15] Tan, Z., Zhou, J., & Liu, Y. (2023, August). CRNN-based abstract artistic text recognition. In *2023 IEEE Smart World Congress (SWC)* (pp. 652-659). IEEE.
[16] Smith, R. (2007, September). An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)* (Vol. 2, pp. 629-633). IEEE.
[17] Deng, G., & Cahill, L. W. (1993, October). An adaptive Gaussian filter for noise reduction and edge detection. In *1993 IEEE conference record nuclear science symposium and medical imaging conference* (pp. 1615-1619). IEEE.
[18] Rong, W., Li, Z., Zhang, W., & Sun, L. (2014, August). An improved CANNY edge detection algorithm. In *2014 IEEE international conference on mechatronics and automation* (pp. 577-582). IEEE.
[19] Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., & Liang, J. (2017). East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 5551-5560).

DOI: 10.2478/ijasitels-2025-0013

Open-loop BLDC motor simulation using the Motor Control blockset from Matlab library

Alina Cristina Viorel[1], *Gabriela Crăciunaș*[1]

[1] *Computer Science and Electrical Engineering Department, "Lucian Blaga" University of Sibiu, Romania, E-Mail : alina.viorel@ulbsibiu.ro, gabriela.craciunas@ulbsibiu.ro*

Abstract

Special electrical machines play an increasingly important role in modern applications. They are designed to address situations that were previously handled by conventional machines, offering improved efficiency and sustainability. The growing need to save resources and protect the environment has encouraged the development of new strategies in motor design and control. This paper presents a BLDC behaviour using datasheets obtained from a PMSM motor. Both BLDC (Brushless DC) and PMSM (Permanent Magnet Synchronous Motor) are permanent magnet machines and their mathematical models are closely related, differing mainly in how the back-EMF is shaped (trapezoidal for BLDC, sinusoidal for PMSM). Key parameters like stator resistance, inductance, and permanent magnet flux linkage are common to both machines. This means PMSM data can be repurposed for BLDC simulations without needing a completely separate dataset. MATLAB/Simulink, provide PMSM blocks that can be adapted for BLDC by changing the EMF profile. This paper presents an open-loop control scenario for simulating the BLDC motor performance using the Motor Control Blockset from the MATLAB/Simulink library starting from a PMSM open loop control strategy.

Keywords: Brushless DC Motor (BLDC), Permanent Magnet Synchronous Motor (PMSM), Open-loop control, MATLAB/Simulink, Motor Control Blockset, Back-EMF, Torque ripple.

1 Introduction

Electric drives based on permanent magnet (PM) machines have become increasingly popular in modern applications such as electric vehicles, robotics, and industrial automation. Among these machines, the Brushless DC Motor (BLDC) and the Permanent Magnet Synchronous Motor (PMSM) are two of the most widely used types due to their high efficiency, compact size, and excellent dynamic performance. Although they share a similar construction—both having a permanent magnet rotor and a three-phase stator—their electrical characteristics and control strategies differ significantly.

However, their behaviour differs significantly due to the type of supply current and the shape of the Back-EMF waveform. Consequently, the control methods also differ: the BLDC motor by using electronic commutation based on Hall sensors [1], while the

PMSM employs vector control (Field-Oriented Control) [2] or sinusoidal PWM. These differences notably affect the torque ripple, noise and overall efficiency [3].
An open loop speed control of BLDC represents an important step in obtaining a model which contains less switching action and a sensorless control technique applied in electric vehicle drives [4].
The BLDC motor is more efficient, requires less money, has a large ratio of torque to weight and it is easy to at all speeds. Hall sensors are used to determine the rotor's position. In an open loop, BLDC motor the control technique involves voltage control, pulse-width modulation, and frequency control. The open-loop control is more accurate than the closed-loop speed control as external factors like temperature and load variations are included [5].
To reduce the ripple torque commutation, a strategy of direct torque control (DTC) is used for a dual inverter-fed three-phase BLDC motor drive with open-end stator winding [6]. Another possibility of reducing the ripple torque is to connect each phase winding of the winding BLDC to an H-bridge inverter, so its phase voltage and current could be controlled independently [7].
MATLAB and its Simulink environment provide a powerful platform for modelling, simulation, and analysis of electrical machines and control systems. The Motor Control Blockset within MATLAB offers prebuilt models, parameterized motor components, and control algorithms that significantly simplify the development and testing of motor control strategies. These tools enable the accurate simulation of both BLDC and PMSM motors under various operating conditions, allowing designers to evaluate torque, speed, current, and efficiency characteristics without the need for physical prototypes. Furthermore, MATLAB's flexibility in integrating mathematical models with control algorithms facilitates a deeper understanding of system dynamics and the impact of different control techniques such as electronic commutation (BLDC) and field-oriented control (PMSM). As a result, MATLAB serves as an essential tool for validating motor performance and optimizing design parameters in a cost-effective and time-efficient manner.
Motor parameter estimation is the key element for implementing motor control algorithms with good results. Accurate motor parameters enable the algorithm to compute the control parameters with precision. Therefore, accurate representation of motor parameters is necessary for a finer speed and torque control when you run PMSMs using control techniques such as field-oriented control (FOC). Motor parameter estimation also enables you to verify the parameter values provided by the motor datasheet. In addition, it enables you to accurately replicate the plant model in Simulink® using which you can simulate real-world scenarios and tests that are difficult to execute using a physical hardware setup.[8]
The paper presents the behavior of a BLDC motor using a PMSM parameters offered by MATLAB library "Motor Control Blockset" and the tutorial applications. From a PMSM datasheet, the most useful parameters for BLDC motor simulation are electrical constants (resistance, inductance), mechanical constants (inertia, friction), and motor ratings (voltage, current, torque, speed, poles). These values allow to build an accurate open-loop strategy of the motor for control and performance analysis.
The objective of this paper is to analyze the behavior of a BLDC motor using 'Motor Control Blockset toolbox and highlight the impact of control method (open-loop speed by VbF control) on Back_ EMF waveform and dynamic response if the procedure starts with data sheet provided by a PSMS motor. It's should be a starting point in developing

a complex strategy in BLDC drive simulation using Motor Control Blocks toolbox developed by MATLAB platform.

2 System modelling

2.1 The structure of BLDC diagram

Several types of electric motors are commonly used in EV applications, including permanent magnet synchronous motors (PMSMs), brushless DC (BLDC) motors, induction motors, and synchronous reluctance motors. Due to its characteristics, BLDC motors are preferred because could achieve precise speed control and reducing torque ripple, particularly under varying load conditions, remains a significant challenge that requires advanced control strategies. BLDC motor control techniques are primarily categorized into sensor-based and sensorless methods.[9]

The sensorless method estimates position by analyzing the back-electromotive force (back-EMF) waveform and identifying zero-crossing points. While sensorless control offers advantages such as reduced hardware complexity and cost, it struggles with rapid variations in load and speed, which can lead to misinterpretation of rotor position and potential system failure. Due to these limitations, sensor-based control techniques are commonly preferred in EV applications to ensure reliable and precise motor operation.[10]

A typical structure for supplying and controlling a BLDC is presented in fig. 1 [9],

Where is illustrates a simplified configuration of a BLDC motor drive system, wherein a battery serves as the power source and delivers energy through a three-phase inverter composed of six MOSFET switches. The rotor, embedded with permanent magnets, directly contributes to the flux linkage, making magnetic flux saturation a prominent aspect of the motor's dynamic behavior. The analysis is carried out under the following assumptions:

1. To prevent magnetic core saturation, the motor is regulated to operate within its nominal current rating.
2. The 3-phase windings are symmetrically arranged.[9]

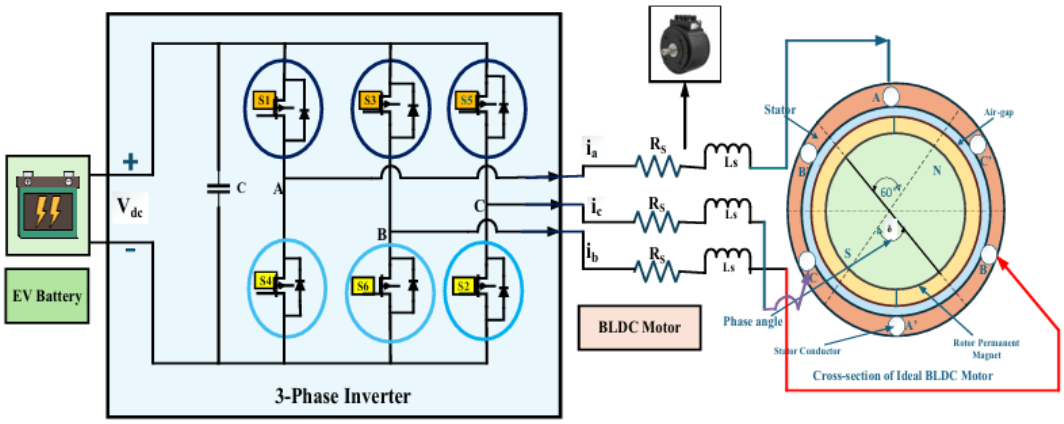


Figure 1. Simplified layout of a BLDC motor drive

The stator windings are a star connection and the voltage relations are the next expressions:

$$V_a = Ri_a + e_a + L_a \frac{di_a}{dt}$$

$$V_b = Ri_b + e_b + L_b \frac{di_b}{dt}$$

$$V_c = Ri_c + e_c + L_c \frac{di_c}{dt}$$

Where, $L_a = L_b = L_c = L_S - L_M$ [Henry]
And $R_a = R_b = R_c = R$

R, L_M, and L_S, are armature resistance in ohms, the stator mutual inductance, and self-inductance.

An example of control of BLDC motor without position sensor created in MATLAB/Simulink blocks are presented in fig. 2. [8]

The system describes a simulation for a BLDC motor, to control the speed in open loop electric drive. The system is feeds by a DC voltage source through a controlled three-phase inverter. The Control subsystem implements the sensorless speed control strategy. A hysteresis controller controls the phase currents.

The model is done by Simscape blocks from electrical and mechanical area. The model proposed uses dedicated blocks for each element like: inverter (with ideal semiconductor switch), ideal torque sensor, inertia, dc sources etc. The load torque is done through a step time source. The BLDC has a star stator winding connection with 4 number of pairs.

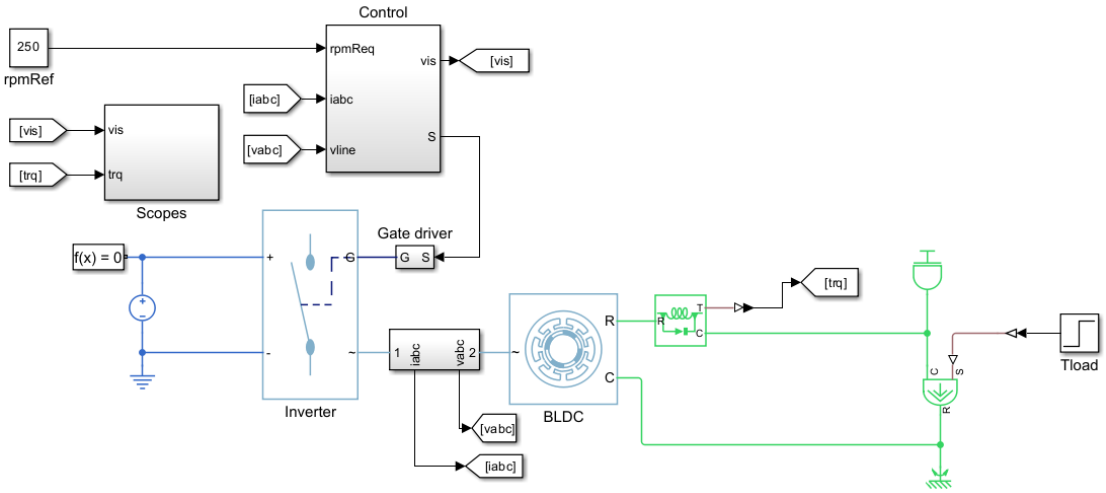


Figure 2. Physical system model of BLDC motor

2.2 The structure of PMSM diagram

In a small commercial application where not much precision is required, open-loop or voltage/frequency (Volt/Hz) control is used, due to their easy calculation and implementation and becomes a low-cost product for PMSM. [11]

Open-loop speed control for a PMSM (Permanent Magnet Synchronous Motor) involves controlling motor speed by adjusting the applied voltage and frequency

without using a speed sensor for feedback. The most common method is V/f constant, where the voltage-to-frequency ratio is kept constant to maintain magnetic flux. This is often used for starting the motor before transitioning to a closed-loop system.
Scalar Control controls only magnitudes. The simplest method to control a PMSM is scalar control, where the relationship between voltage or current and frequency are kept constant through the motors speed range. The frequency is set according to the wanted synchronous speed, and the magnitude of the voltage/current is adjusted to keep the ratio between them constant. No control over angles is utilized, hence the name scalar control. The method uses an open-loop control approach without any feedback of motor parameters or its position.[12]
Scheme presented in Fig. 3 is an open-loop PMSM electric drive. [14]
In this situation, the system does not measure or adjust based on rotor position or speed and have a fixed input: the controller sends a predefined signal to the inverter.
Sensorless control strategies for PMSM eliminate the need for physical position sensors by estimating rotor position and speed using mathematical models and observers. These methods improve reliability, reduce cost, and are widely used in automotive, industrial, and consumer applications. The simulation is realized with Simscape Electrical Specialized Power Systems Blocks types of blocks which will be not sustained by the versions of MATLAB anymore.

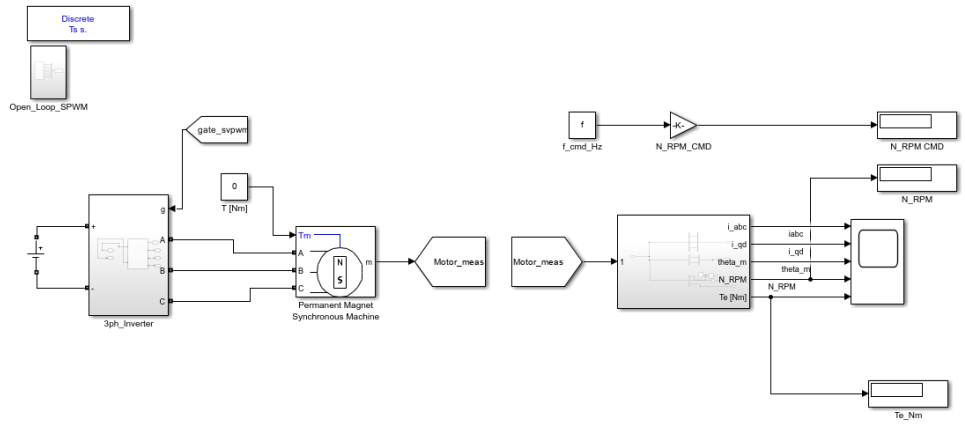


Figure 3 An open-loop PMSM control.

2.3 MATLAB simulations of BLDC motor

2.3.1 Motor Control Blockset toolbox

The Simscape library contains dedicated blocks in electrical and mechanical area for simulating the behavior of PMSM/BLDC motors in different situations with appreciate results. In many applications simulations part and practical determination are separately created and the aim is to validate simulation results with practical one.
Control Motor blocks are the building blocks of modern motor control systems. They abstract complex math and hardware interactions into manageable modules, making it easier to design safe, efficient, and scalable robotics or automation solutions.
Motor Control Blockset™ is a MATLAB/Simulink toolbox designed to help engineers design, simulate, and deploy advanced motor control algorithms on real hardware. It

provides ready-to-use blocks, reference examples, and automatic code generation for microcontrollers, FPGAs, and SoCs.
Motor Control Blockset bridges the gap between motor control theory and real-world deployment, making it easier to move from simulation to hardware implementation.
This toolbox includes prebuilt Simulink blocks for common motor control techniques like Clarke and Park transforms, six-step commutation, lookup table or sensor support. It's able to use encoders, Hall sensors, and resolvers. More offers to users sensorless estimators for rotor position and speed and generates different types of codes.
This toolbox supports popular hardware platforms such as Texas Instruments C2000, STM32G4 boards and enables both open-loop (scalar control) and closed-loop (vector control) methods, including Field-Oriented Control (FOC). The toolbox is focused on PMSM structure but it could be extended to other motor types also.
Thus, using the block named "Interior PMSM" from "Motor Control" toolbox, is possible to simulate the behavior of a PMSM only introducing the motor parameters for start an application. These parameters describe the physical properties of the motor and help the built-in blocks build accurate models and design effective controllers.
The toolbox permits to observe how motors behave—speed, torque, position—by combining sensor inputs, mathematical transforms, and control strategies into modular blocks that can be reused and deployed on hardware.
MATLAB datasheet is a very useful tool in observing the motor behavior in several situations, starting for simple to complex, without using very complicated formula and calculus.
Unfortunately, in practice are just few manufacturers who will supply a datasheet that provides information on the motor's properties and performance. For this particular reason in this paper the parameters described in MATLAB tutorial are used for implementing an open-loop speed strategy for BLDC using comparable parameters from PMSM. This should be the first step in designing a model for BLDC drive and then to extend the research to complex strategies, like: closed-loop control speed, using Hall sensors to determine exactly the rotor position.
The motor parameters offered by MATLAB simulation are done in Table 1.

Table 1 PMSM motor parameter

Description	Variables	Values
Number of Pole Pairs	`pmsm.p`	4
Rotational Inertia	`pmsm.J`	0.0227 kg*m^2
Stator Resistance	`pmsm.Rs`	0.02 Ω
D-Axis Inductance	`pmsm.Ld`	61.31 µH
Q-Axis Inductance	`pmsm.Lq`	122.63 µH
Permanent Magnet Flux	`pmsm.FluxPM`	0.0725 Wb
Rated Voltage	`pmsm.V_rated`	325 V
Rated RPM	`pmsm.N_rated`	500 RPM
Rated Current	`pmsm.I_rated`	200 A

These tools will output a data structure containing the estimated parameters, which can then be loaded into a MATLAB® workspace.

2.3.2 BLDC block from Motor Control Blockset

This toolbox contains a BLDC block which implements a three-phase brushless DC (BLDC) motor with a trapezoidal back electromotive force that remains constant for a position range of 120 electrical degrees. The block uses the three-phase input voltages to regulate the individual phase currents, allowing control of the motor torque or speed. The motor parameters are taken from Table 1.
In many situations the values of inductance after both axes are equal for a BLDC, but when you want to assure a good balance between torque and speed control different values are recommended and exploited in advanced control methods like field-oriented control. Thus, the motor datasheet plays a very important role in motor's behavior and in choosing a control strategy.
This strategy is used into a testing procedure, when the motor doesn't have any load and it will be connected into a larger structure after that.
The MATLAB block used in applications is presented in fig.4. It has 2 inputs, LdTrq (load torque on motor), PhaseVolt (stator terminal voltages) and 3 outputs Info (contains the block calculations), PhaseCurr (phase a, b, c current, i_a, i_b, and i_c) and MtrSpd(angular speed of the motor). Being on the testing phase the value of load torque on motor is setting to zero.

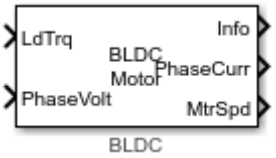


Figure. 4 Three-phase brushless DC motor with trapezoidal flux distribution

The paper presents an open-loop speed control strategy applied on the BLDC motor which have different values for inductances on both axes in several situations:

- 1. The motor is feed with three-phase voltage input.
- 2. The motor is feed by a DC voltage through an inverter.
- 3. To create an open-loop speed control (Vbf controller)

2.3.3 BLDC – speed control strategy using a three-phase AC voltage signal.

The first step in developing a control strategy is to implement the motor data sheet into the model. The three-phase AC voltage signal is the primary driver of the motor response.
The structure builds up with blocks from MATLAB/Simulink are presented in fig. 5.

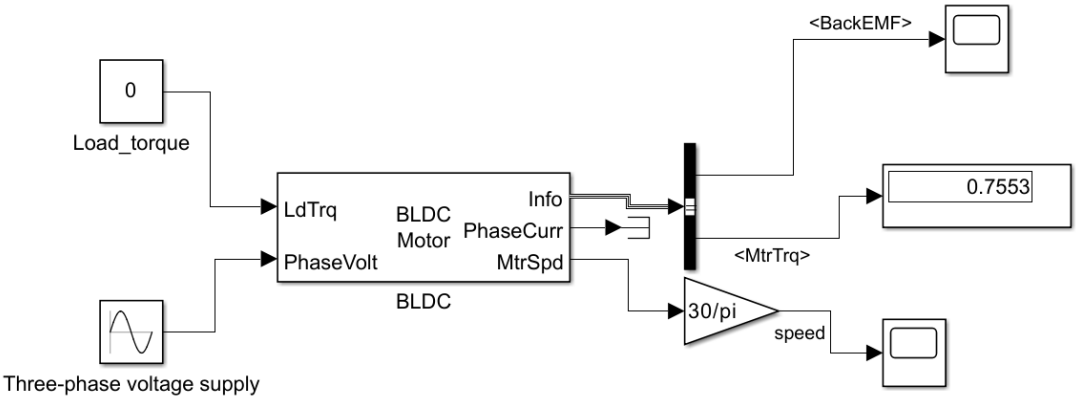


Figure 5. The primary driver of the motor response

The three-phase AC voltage signal is the primary driver of the motor response with the parameters considered: amplitude 50, frequency 60 rad/sec and the phase (rad) [0, -2*pi/3, -4*pi/3].
In the *Display block* connected at the BLDC motor output, the motor torque output doesn't settle quite at zero but rather at a very small number, which represents a good start in developing a testing procedure.
The variation of motor speed and the shape of back EMF in certain conditions is presented in figure 6,7.

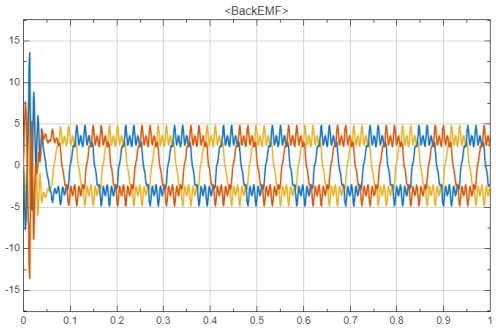


Figure 6. Output of BLDC motor Back_EMF of phases (A, B, C)

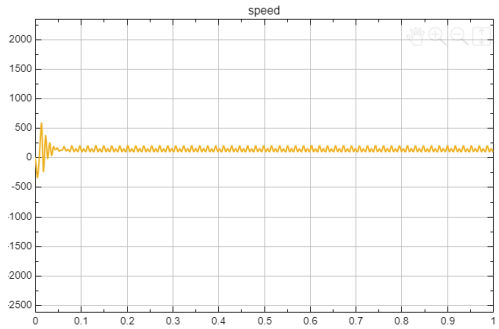


Figure 7. The BLDC motor speed

The motor BLDC has some ripple effects on Back_EMF and speed shapes that could affect torque stability, efficiency, noise, and control accuracy. Higher ripple speed (rapid fluctuations in torque or current) can cause vibration, acoustic noise, and reduced smoothness of rotation, while lower ripple speed improves precision and efficiency.

2.3.4 BLDC inverter solutions.

In electric drive the nature of source plays a very important role due the applications which involved BLDC motors. Most of the applications supposed to have DC sources like electric vehicle for example. In all these situations an inverter is necessary to add. The Average-Value Inverter block provides an easy-to-use inverter model for solving this situation. The diagram control of motors became like in fig. 8.

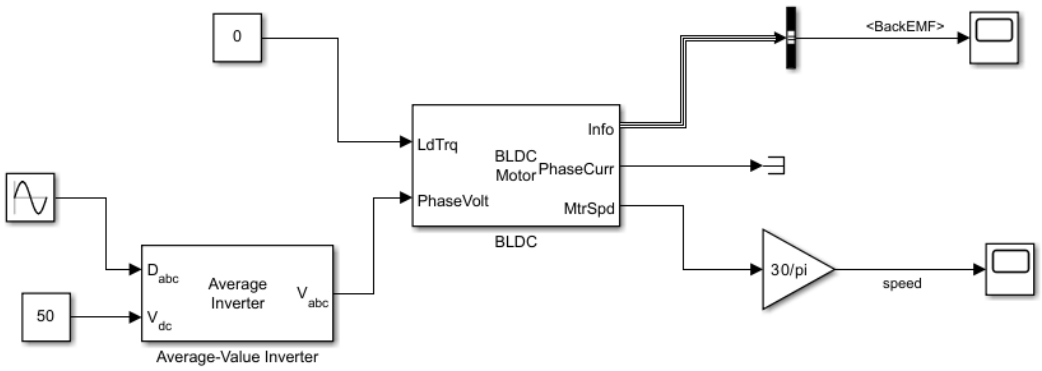


Figure 8 The BLDC motor fed by battery thorough average-value inverter

The Average-Value Inverter models an average-value and full-wave inverter. It computes the three-phase AC voltage output from inverter DC voltage by using the duty cycle information. The input D_{abc} of inverter needed a three-phase signal for its duty cycle, so a *Sine Wave block* is used for this reason. The duty signal needs to be between 0 and 1, similar to a Pulse Width Modulation (PWM) signal and the data for Sine Wave block are: Amplitude = 1, Bias = 0.5, frequency= 120 rad/sec and phase [0, -2*pi/3, -4*pi/3]. The value of voltage DC input Vdc is the same like in the previous case. The BLDC motor shape of Back-EMF and speed response is presented in fig. 9, 10.

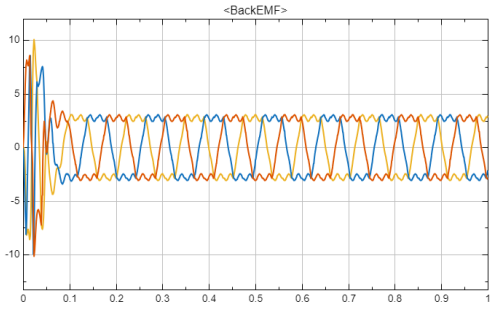


Figure 9. Output of BLDC motor Back_EMF of phases (A, B, C) with inverter

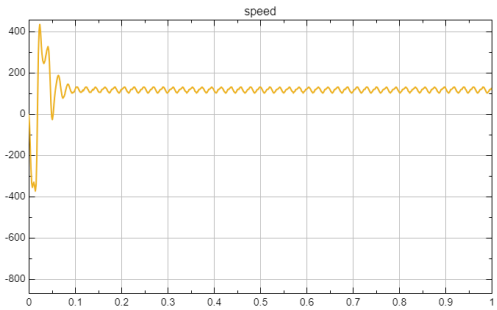


Figure 10. The BLDC motor speed with inverter

The values of the start-up procedure and the ripple of Back_EMF and motor speed are reduced compare with the previous case. The inverter plays a positive role in order to have a system more stable.

2.4 Open loop control for BLDC motor

2.4.1 Voltage by Frequency Control

When the motors are stopped suddenly having the entire supply voltage could create negative behaviours. The motors would be difficult to control, and it would not have smooth movement. A slow, gradual acceleration would be much safer and more comfortable. Control voltage by frequency (V/F) is an open loop control strategy for steadily increasing the applied voltage and signal frequency to an AC motor [15].

In this type of control, the phase voltage and frequency of the signal sent to the motor are increased proportionally. Both parameters are increased till the desired voltage is reached even the frequency continue to increase. The procedure helps the motor to start smooth.

2.4.2 VbyF Controller block in BLDC simulation.

In Motion Control Blokset is used a *VbyF Controller block* to simplify the implementation process and to get quickly a working controller desired motor. The input block contains information about desired rotational speed (Ref_RPM), the rated rotational speed (Rated_RPM), the number of pole pairs in the rotor (No_poles), the rated peak voltage of the motor (Rated_Voltage), the minimum operating voltage of the motor (Minimum_Voltage), considered to be 15% of the motor's rated voltage is often a good place to start, the number of simulation time steps(N-ramp) the controller takes to get the motor up to the reference speed. The input Enable acts as a switch. A value of 0 will turn the controller off, while a value of 1 will turn it on.
The paper presents an open-loop control for a BLDC type using a VbyF controller, a block who permit to implement the value of the speed desired for the considered model. The input values used are presented in the Table 1. In fig. 11 the simulation diagram is presented.

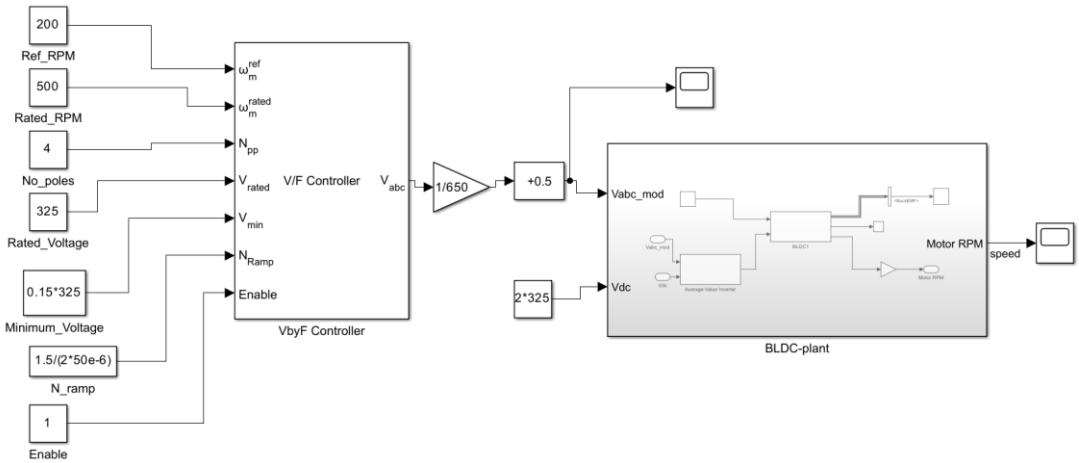


Figure 11 Open-loop BLDC speed control

The subsystem *BLDC-plant* contains the structure created before which contains an inverter. A voltage by frequency (V/F or VbyF) controller was added and the existing system for controlling the speed motor safely from a stop up to the desired speed. This block has an output V_{abc} which is a three-phase voltage signal. The meaning of the block is to simulate the action of accelerating pedal in a safe mode (slowly and gradually). Most of the previous parameters used to set up the motor-inverter plant model still work here for notice the impact of control by VbF controller has on the BLDC's behavior.
The value to the motor's peak-to-peak rated voltage is using like value for the dc input of the inverter.
The output of BLDC motor (Back_EMF and speed motor) when a VbyF controller is used are presented in figures 12 and 13.

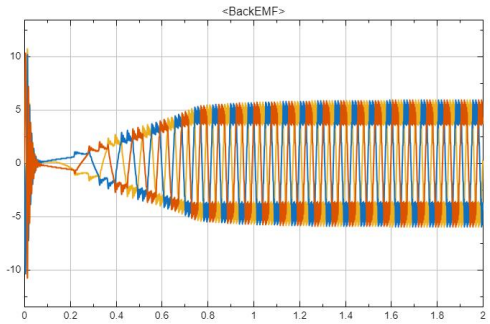


Figure 12. Output of BLDC motor Back_EMF of phases (A, B, C) with Vbf controller

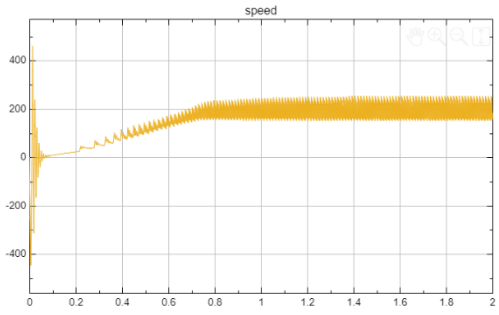


Figure 13. The BLDC motor speed with Vbf controller

The simulation emphasis some ripple in Back_EMF shape and motor speed. Ripple in the speed of a BLDC motor is mainly caused by torque ripple, which arises from the motor's non-ideal commutation, magnetic design, and control methods.
The results obtained emphasis the utility of the procedure in testing motor wiring and inverter operation with reduce cost and complexity, provides insights into performance limitations, and supports applications where simplicity and efficiency outweigh precision. More, motor parameter estimation represents a very important element for implementing motor control algorithms with good results.
Open-loop control eliminates sensors and feedback mechanisms, making systems cheaper and easier to implement. This is especially valuable in low-cost consumer devices or applications where precise speed/torque control is not critical. It also lays the groundwork for improving closed-loop systems by understanding baseline behaviors.

3 Conclusions

This simple control method is suggested to replace conventional induction motors in applications like fans, pumps and ventilation systems where high dynamic performance is not a requirement.
The paper presents a modality to apply the new toolbox implemented by MATLAB, *Motor Control Blockset* who offers strategies to control systems.
Open-loop systems are more vulnerable to disturbances such as sudden load changes, supply voltage fluctuations, or mechanical stress. This can reduce reliability in industrial or automotive contexts.
The results obtained in simulations in motor speed especially are good even the speed ripple is significative. This simulation could a pointing start for developing a closed loop control strategy where precise speed and torque regulation using feedback (Hall sensors, encoders, or sensorless estimation are necessary especially when varying loads existing. Closed-loop control offers precision and robustness but at higher cost and complexity. Future research should focus on sensorless and hybrid strategies to bridge the gap, making BLDC drives both affordable and efficient across diverse applications.

References

[1] Pallavi K Gamit, *Simulation of BLDC Motor Control Using Conventional PI Controller in MATLAB Simulink'*, *jes*, vol. 20, no. 3, pp. 8051–8061, Apr. 2024, doi: 10.52783/jes.7807.

[2] Xiaotong Li, Daoliang Xi, Yullin Zhang, Permanent *magnet synchronous motor vector control based on MATLAB/Simulink*, First International Conference on Electronics Instrumentation& Information Systems (EIIS), Febr. 2018, doi:10.1109/EIIS.2017.8298737.
[3] A. Sonkar, S. Sashidhar, and B. Subudhi, 'Maximum Torque per Ampere Control of Interior PMSM and BLDC Motor Using Co-Simulation', in *2024 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES)*, Mangalore, India: IEEE, Dec. 2024, pp. 1–6. doi: 10.1109/PEDES61459.2024.10961197.
[4] P. V. Kulkarni, P. P. Asnotkar, P. M. Katageri, N. Parameshwar, and K. Chachadi, '*Simulation of open loop speed control of BLDC motor for electric vehicle application'*, in 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India: IEEE, Aug. 2020, pp. 577–585. doi: 10.1109/ICSSIT48917.2020.9214085.
[5] B. Ramesh, K. Chenchireddy, B. N. Reddy, B. Siddharth, C. V. Kumar, and P. Manojkumar, '*Closed-loop control of BLDC motor using Hall effect sensors*', IJAPE, vol. 12, no. 3, p. 247, July 2023, doi: 10.11591/ ijape.v12. i3.pp247-254.
[6]] H. K. U and R. P. P, '*Commutation Torque Ripple Reduction in Open-end Winding BLDC Motor Drives*', in 2024 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES), Mangalore, India: IEEE, Dec. 2024, pp. 1–6. doi: 10.1109/PEDES61459.2024.10961717.
[7] Q. Huang, L. Luo, X. Zhang, and L. Diao, '*Fault Tolerant Control for Open Winding Brushless DC Motor with Power Device Failure*', J. Electr. Eng. Technol., vol. 19, no. 5, pp. 3175–3189, July 2024, doi: 10.1007/s42835-024-01784-0.
[8] www.mathworks.com.
[9] Abhishek Kumar, M.Venkatesh Naik, Rahul Kumar , *Comparative Analysis of Speed Control for BLDC Motors in Electric Two-Wheeler Applications* , 2025 IEEE International Conference on Computer, Eleclectronics, Electrical Engineering & their Applications(IC2E3), Srinagar Garhwal, India, 2025, pp.1-6, doi:10.1109/IC2E365635.2025.11167351.
[10] G. Gupta and M. Sreejeth, "*Comparative analysis of Speed control of BLDC motor using PWM and Current Control Techniques*," 2022 IEEE IAS Global Conference on Emerging Technologies (GlobConET), Arad, Romania, 2022, pp. 610-614, doi: 10.1109/GlobConET53749.2022.9872173.
[11] S. Gokul Krishnan, Sahishnu Raju Kakarlapudi, R. Vezhavendhan. *Modeling of BLDC Motor with Enhanced Controller Tuning: Employing Golden Eagle Optimization for Efficient Automated Manufacturing Systems*, Springer Nature Singapore Pte Ltd. 2024 S. G. Ponnambalam et al. (eds.), Recent Advances in Industrial and Systems Engineering, Lecture Notes on Multidisciplinary Industrial Engineering, https://doi.org/10.1007/978-981-97-4700-9_31323.
[12] Jung, T.-U., Nam, N.N.: *A High-Efficiency Driving Method of BLDC motor based on modified Trapezoidal Method*. J. Electric. Eng. Technol. Springer Science and Business Media LLC 2022.
[13] P.D. Chandana Perera, F. Blaabjerg, J.K. Pedersen, and P. Thogersen, *A sensorless, stable V/fcontrol method for permanent-magnet synchronous motor drives*, IEEE Trans. Ind. Applicat., vol. 39, pp. 783-791, May/June 2003.
[14] Neeraj Y. *Open loop V/F control of PMSM using Sine PWM* https://ch.mathworks.com/matlabcentral/fileexchange/177949-open-loop-v-f-control-of-pmsm-using-sine-pwm, MATLAB Central File Exchange. Retrieved November 25, 2025.
[15] Yashvi Parmar, Priyanka Patel, Nitiksha Pancholi, Chand Thakor, Unnati Mali *Scalar control of Permanent Magnet Synchronous motor*, International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395 -0056, Volume: 03 Issue: 12 | Dec -2016.
[16] Munaf S. N. Al-Din, Majid A. Al-Taee - *Efficiency optimization of an open loop controlled permanent magnet synchronous motor drive using adaptive neural networks*, European Scientific Journal January 2014 edition vol.10, No.3 ISSN: 1857 – 7881 (Print) e - ISSN 1857- 7431, 2014.

DOI: 10.2478/ijasitels-2025-0014

Deep Learning Segmentation for stem cells images

Constantin-Cristian DRĂGHICI[1], *Cătălina NEGHINĂ*[1]

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

Cristian DRĂGHICI, e-mail: constantincristian.draghici@ulbsibiu.ro
Cătălina NEGHINĂ, e-mail: catalina.neghina@ulbsibiu.ro

Abstract

In this study, we investigate the application of deep learning techniques for automatic segmentation of the Fluo-N2DH-GOWT1 dataset, which consists of time-lapse grayscale images of rat neural stem cells. We employ the DeepLabV3+ semantic segmentation framework with a ResNet-18 backbone, chosen for its balance between accuracy and computational efficiency on relatively small biomedical datasets. To improve generalization and robustness, we apply data augmentation strategies including rotation, scaling, shear, and reflection. The performance of the proposed model is evaluated using standard metrics such as F1-score, Intersection of Union, Precision and Recall. Experimental results demonstrate that the ResNet-18–based DeepLabV3+ achieves reliable segmentation of stem cells, effectively distinguishing cells from the background.

Keywords: Stem Cells, Semantic Segmentation, Instance Segmentation, Deep Learning, ResNet

1 Introduction

Accurate detection and segmentation of stem cells represent a critical step in the quantitative analysis of cellular dynamics within biomedical research. Advances in computer science have profoundly transformed biological research, particularly through computational image analysis. In microscopy, automatic cell segmentation enables quantitative and reproducible study of morphology and dynamics, overcoming the limitations of manual annotation like time inefficiency, subjectivity, and variability [1], [2]. Recent progress in deep learning, especially convolutional neural networks, has revolutionized image understanding. CNNs learn hierarchical visual representations directly from data, achieving high accuracy in biomedical image segmentation and enabling the precise delineation of complex cellular structures [3], [4].

The focus of this study was the segmentation of fluorescence microscopy images from dataset Fluo-N2DH-GOWT1 [5]. Previous studies using this dataset employed architectures such as U-Net [6], [7], Motion UMiTNet [8], and hybrid or recurrent variants like C-LSTM [9], combining spatial and temporal modeling for improved segmentation accuracy. This work highlights the potential of deep learning techniques in automating the analysis of microscopy images, providing accurate and efficient tools for stem cell research and other biomedical applications [10], [11].

Convolutional Neural Networks (CNNs) extract hierarchical spatial features using convolution and pooling operations, achieving high performance in tasks such as classification and segmentation. Residual Networks (ResNet) introduce skip connections to ease the training of deeper models by learning residual mappings. For semantic segmentation, encoder-decoder architectures predict pixel-level class labels while preserving spatial context. DeepLabV3+ extends this design by integrating Atrous Spatial Pyramid Pooling (ASPP) for multi-scale context extraction and a refinement decoder for boundary accuracy [3], [12], [13].

The primary goal of this study is to set up and evaluate a context-aware deep learning framework for automatic segmentation of stem cells in fluorescence microscopy images. Specifically, the objectives are:

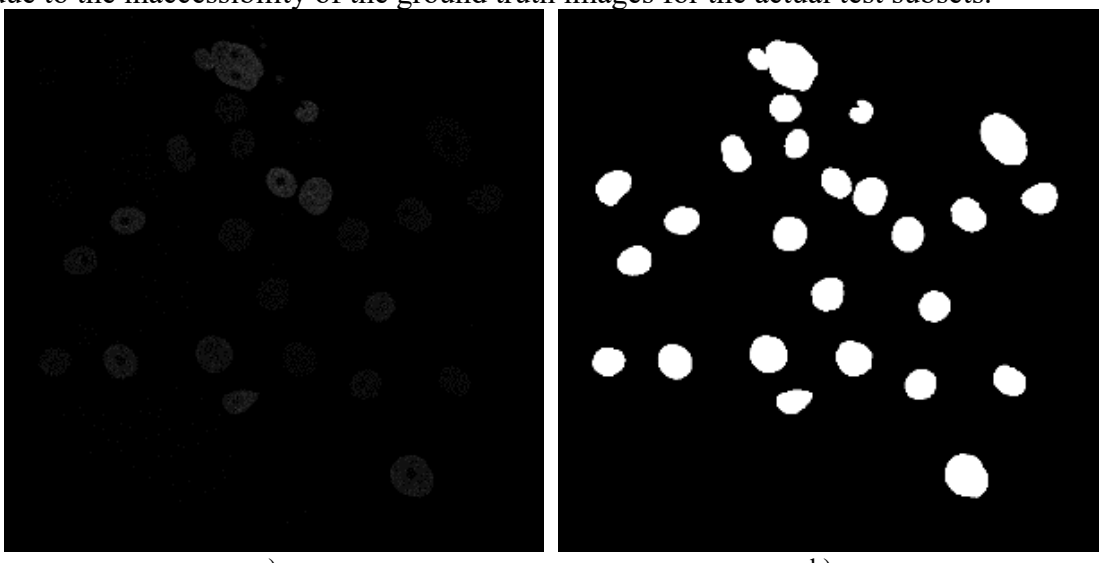
- To implement an image segmentation framework based on the DeepLabV3+ architecture with a ResNet-18 backbone [12], [13], fit for small biomedical datasets
- To train the network on the Fluo-N2DH-GOWT1 dataset with appropriate data augmentation techniques to improve generalization.
- To evaluate the semantic segmentation performance using standard quantitative metrics such as *Intersection over Union* (IoU) and *F1*-score
- To evaluate the instance segmentation using metrics such as *Precision* and *Recall*

By addressing these objectives, this work highlights the potential of contextual deep learning in automating the analysis of microscopy images, providing accurate and efficient tools for stem cell research and other biomedical applications.

2 Dataset

The Fluo-N2DH-GOWT1 dataset, part of the *Cell Tracking Challenge* [14], contains 2 subsets with 2D fluorescence microscopy sequences (*dataset1* with 92 frames and *dataset2* also with 92 frames) of GOWT1 stem cells. Each image is accompanied by instance-level annotations, such as in *Figure 1*. The image resolution is 1024x1024 px. Each dataset consists of successive frames of the same cell culture.

The dataset was split into training (*dataset2*) and validation subsets (*dataset1*), due to the inaccessibility of the ground truth images for the actual test subsets.



a) b)

Figure 1. a) Original image from *dataset2,* b) ground truth instance

3 Methods

A flowchart of the proposed methods can be seen in *Figure 2*.

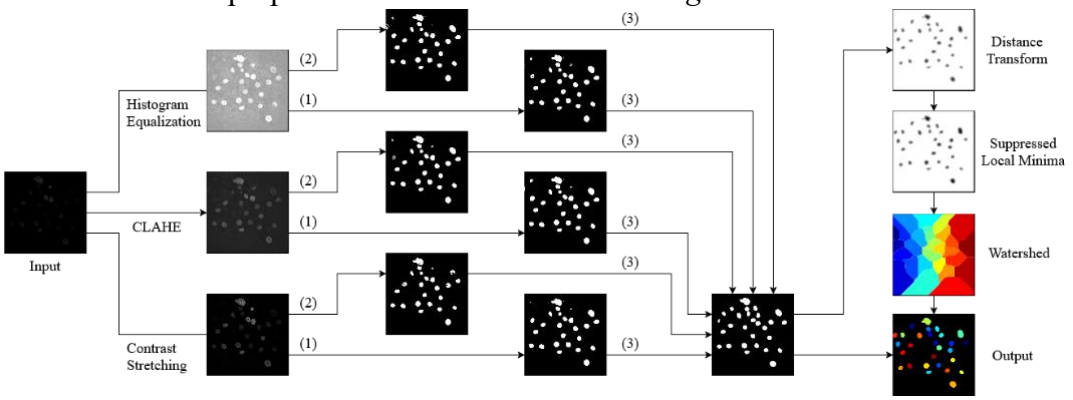


Figure 2. Scenario scheme: (1) Segmentation using a model trained on images without brightness, contrast modifications, or added noise; (2) Segmentation using a model trained on images with brightness and contrast adjustments and with added noise; (3) Series of morphological operations: removal of objects smaller than a given threshold, dilation, hole filling, and erosion.

3.1 Preprocessing

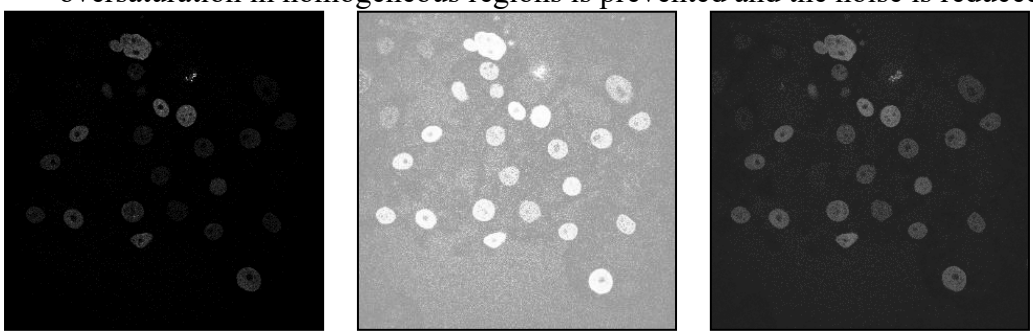
To enhance feature extraction, three grayscale preprocessing techniques were applied, which can be seen in *Figure 3*:

a) *Contrast Stretching*. The method consists of linearly adjusting the image contrast so that the range of pixel values is stretched across the entire available domain [0, 255] using eq.1, where I_{min} and I_{max} represents the minimum respectively the maximum intensity of the image $Imag_{orig}$.

$$Imag_{New} = \frac{Imag_{orig} - I_{min}}{I_{max} - I_{min}} \cdot 255 \quad (1)$$

b) *Histogram Equalization*. This method improves contrast by redistributing pixel intensities so that the histogram becomes more uniformly distributed [15]. However, this technique may lead to amplification of the noise present in the image.

c) *Contrast Limited Adaptive Histogram Equalization (CLAHE)* [16]. This method divides the image into blocks (in this case of dimension 8x8 px) and applies local histogram equalization to each of them. In this way the contrast oversaturation in homogeneous regions is prevented and the noise is reduced.



a) b) c)

Figure 3. a) Maximized contrast, b) Equalized Histogram, c) CLAHE

Each method produced a distinct dataset for comparative analysis of segmentation performance.

3.2 Training and Segmentation

The model consists of a DeepLabV3+ network with a ResNet-18 backbone [12], [13] pretrained on the ImageNet dataset [17]. The network, optimized for cross-entropy loss with a final classification layer, contains approximately 20.6 million trainable parameters and is configured for binary semantic segmentation (*cell* vs. *background*).

The binary masks were mapped from the 255 and 0 intensities to categorical labels for "*cell*" and "*background*" respectively.

To enhance generalization, the training set underwent extensive augmentation. Geometric transformations included random rotations (0°–360°), horizontal/vertical flips, scaling (0.7–1.4), shear (±10°), and translations (−50 to 50 pixels). A second variant added photometric changes like brightness/contrast adjustments (±5%) and Gaussian noise (σ=0.05) to simulate diverse acquisition conditions.

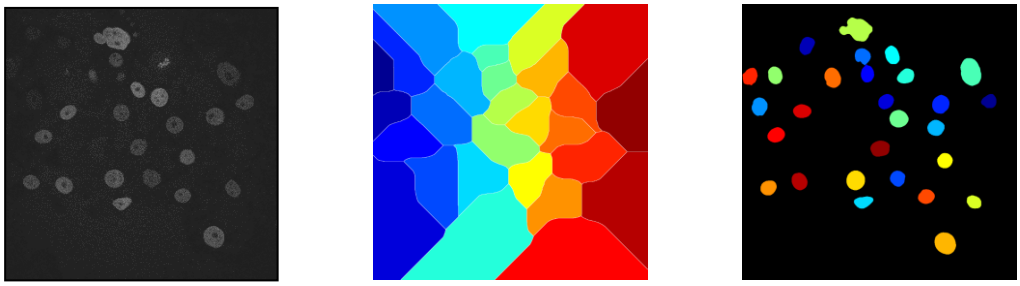
The network was trained with a batch size of 8, using the Adam optimizer, learning rate of 0.001, and 10 epochs. These parameters balanced GPU memory constraints with convergence speed. Data were shuffled each epoch, and validation metrics were monitored every five iterations. Checkpointing allowed recovery in case of interruptions. After training, the network produced binary segmentation images with white (255) representing estimated cells and black (0) representing the background.

3.3 Postprocessing

Postprocessing was applied to all segmentation results to correct local artefacts. Small regions were removed to eliminate possible noise or artefacts. Then, the binary image obtained is dilated, expanding the edges of the objects to close any discontinuities or gaps that may have appeared during segmentation. Subsequently, the internal holes within the binary objects are filled, ensuring a coherent representation of the cells as compact and continuous regions. Finally, a morphological erosion using the same structuring element as the morphological dilation is used, in order to restore object sizes while maintaining boundary improvements.

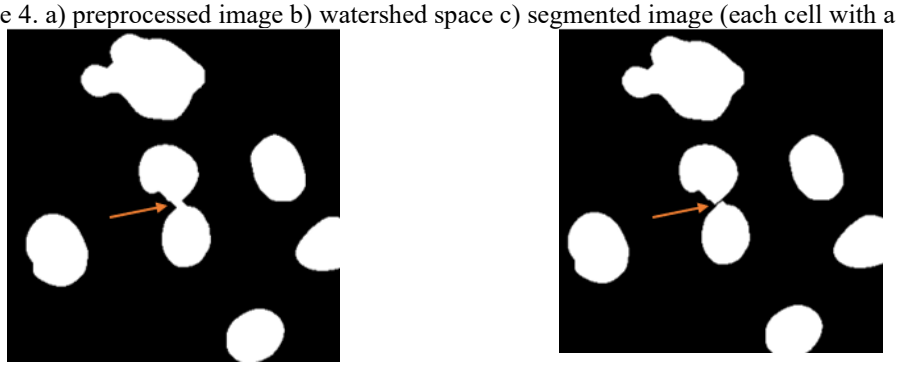
In some cases, the segmentation may produce merged regions corresponding to multiple adjacent cells (as in *Figure 5*). To separate merged neighboring cells, the *watershed transform* [11] was applied. The watershed algorithm starts from the *distance transform* [15] of the complement of the segmented image after morphological operators. The *distance transform* calculates, for each pixel, the minimum distance to the edge of the object, resulting in a three-dimensional map where the central areas of the objects appear as valleys and the edges as peaks. To avoid over-segmentation caused by minor intensity variations, less significant local minima are suppressed. Watershed lines then separate individual cells, effectively resolving overlaps.

This postprocessing pipeline produces cleaner, more consistent segmented images (as in *Figure 4*), providing a reliable foundation for the quantitative evaluation.



a) b) c)

Figure 4. a) preprocessed image b) watershed space c) segmented image (each cell with a color)



a) b)

Figure 5. a) Merged cells b) Result after applying watershed transform

4 Results and Conclusions

The segmentation performance of the DeepLabV3+ network with a ResNet-18 backbone (including pre-processing and post-processing) was evaluated by comparing the binary images obtained after segmentation with ground-truth annotations. Two aspects were analyzed:

a) **Semantic segmentation:** how well the pixels were correctly identified
b) Instance segmentation: how well the cells were identified

4.1 Semantic segmentation

In order to determine how well the pixels were correctly classified as *background* or *cell* pixels, the confusion matrix was computed from:

- True Positives (TP) = *cell* pixels classified correctly as *cell*
- True Negatives (TN) = *background* pixels classified correctly as *background*
- False Positives (FP) = *background* pixels classified as *cell*
- False Negatives (FN) = *cell* pixels classified as *background*

The F1-score (eq. 2) and Intersection of Union (IoU - eq. 3) were also derived.

$$F1\ Score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \tag{2}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{3}$$

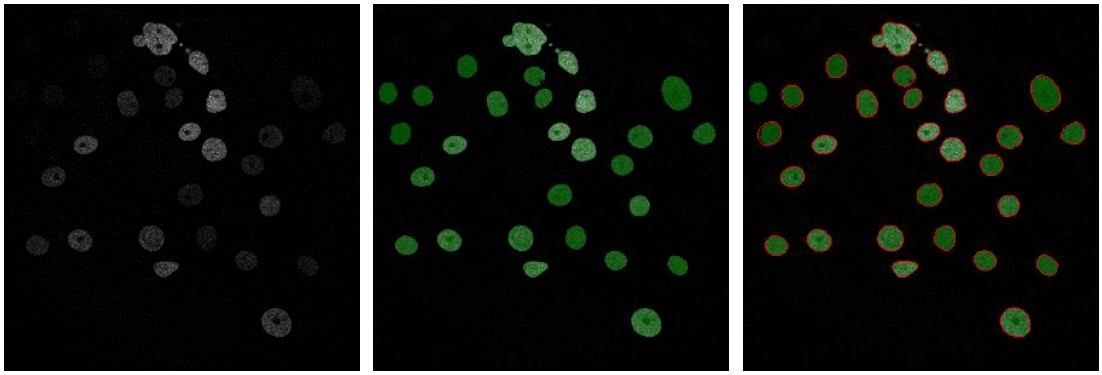
Six preprocessing strategies were tested: contrast maximization, histogram equalization, and adaptive histogram equalization (CLAHE), each of them with and without photometric changes. Among them, CLAHE without photometric changes produced the best results, as shown in *Figure 6*.

For postprocessing, several structuring element shapes and sizes were evaluated, and the diamond-shaped element of size 3 provided the most accurate results, effectively removing small artifacts while preserving cell contours.

With this configuration, the model consistently achieved an F1-score of 95.40%, indicating high segmentation accuracy and confirming the effectiveness of the DeepLabV3+ architecture for fluorescence microscopy cell images. For a complete overview of the evaluation metrics, the full results are provided in *Table 1*.

Table 1. Segmentation evaluation metrics for the validation dataset

	Structural element and size	**Diamond**					
		3 pixels		5 pixels		7 pixels	
	Metric	F1-Score	IoU	F1-Score	IoU	F1-Score	IoU
No photometric changes	Contrast maximization	94.66%	89.86%	94.66%	89.86%	94.65%	89.86%
	Histogram Equalization	92.72%	86.48%	92.75%	86.53%	92.77%	86.56%
	CLAHE	**95.40%**	**91.23%**	95.40%	91.22%	95.38%	91.20%
Photometric changes	Contrast maximization	83.31%	71.57%	83.34%	71.62%	83.38%	71.67%
	Histogram Equalization	94.38%	89.37%	94.38%	89.36%	94.37%	89.36%
	CLAHE	95.09%	90.66%	95.11%	90.69%	95.14%	90.75%
	Structural element and size	**Disk**					
		3 pixels		5 pixels		7 pixels	
	Metric	F1-Score	IoU	F1-Score	IoU	F1-Score	IoU
No photometric changes	Contrast maximization	94.65%	89.84%	94.66%	89.86%	94.66%	89.86%
	Histogram Equalization	92.72%	86.47%	92.74%	86.51%	92.77%	86.57%
	CLAHE	**95.40%**	**91.22%**	95.40%	91.22%	95.39%	91.20%
Photometric changes	Contrast maximization	83.29%	71.54%	83.34%	71.62%	83.39%	71.69%
	Histogram Equalization	94.38%	89.36%	94.38%	89.37%	94.37%	89.36%
	CLAHE	95.09%	90.65%	95.11%	90.69%	95.15%	90.77%



a) b) c)

Figure 6. Segmentation results for the best parameters
a) Preprocessing using CLAHE, b) green overlay of the detection, c) the mask contour (red) overlaid on image 6.b.

4.2 Instance segmentation

To determine how well the cells have been identified, for each frame, each cell from the segmented image must be matched to the corresponding cell from the reference image. For the reference cells, the size of each cell is additionally extracted by estimating the effective radius as half of the diagonal of the bounding box rectangle. This approximation is justified by considering the cell as an elliptical shape, and the bounding box diagonal provides a reasonable estimate of the maximum extent.

In order to establish the association between predicted and actual cells, a matrix of Euclidean distances is constructed between all the centroids of the cells in the segmented image and those in the reference image (*Figure 7*). The matrix is then filtered to retain only the values below the effective radius associated with each reference object; larger values are marked as ∞ (infinity), indicating that the prediction is considered too far to be a valid match. This restriction serves to prevent false matching between spatially distant objects.

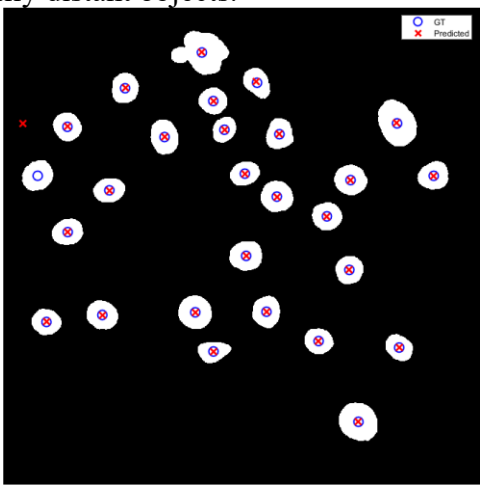


Figure 7. Image with the centroids of the segmented cells and the centroids of the cells from the reference images

For the best configuration for semantic segmentation (diamond 3px, CLAHE, no photometric changes), the segmentation performance was further evaluated by counting the detected and missed cells, resulting in the values from *Table 2*.

Table 2. Results for instance segmentation

Metric	Subset 1 (validation)	Subset 2 (training)
TP (*cells* as *cells*)	No. of detected cells: 1982	No. of detected cells: 2282
FP (non-*cells* as *cells*)	No. of detected cells: 138	No. of detected cells: 167
FN (*cells* as non-*cells*)	No. of detected cells: 26	No. of detected cells: 5
$Precision: \frac{TP}{TP + FP} \cdot 100$	93.49%	93.18%
$Recall = \frac{TP}{TP + FN} \cdot 100$	98.70%	99.78%
The average distance between the centers of the reference cells and the centers of the associated predicted cells	1.03px	0.54px

5 Conclusions

The proposed DeepLabV3+ model with a ResNet-18 backbone achieved reliable segmentation of GOWT1 stem cells, reaching on the validation dataset a semantic segmentation F1-score of 95.40% under optimal preprocessing and postprocessing conditions and an instance segmentation precision score of 93.49% and a recall of 98.70%. These results confirm the effectiveness of context-aware CNNs for fluorescence microscopy, offering a strong balance between precision and computational efficiency. The approach proved robust to variations in illumination and noise, providing consistent, biologically meaningful segmentations suitable for downstream quantitative analyses and future extensions toward tracking or division detection.

References

[1] E. Meijering, *Cell Segmentation: 50 Years Down the Road [Life Sciences]*, IEEE Signal Process. Mag., vol. 29, no. 5, pp. 140–145, doi: 10.1109/MSP.2012.2204190, Sept. 2012.

[2] C. F. Koyuncu, R. Cetin-Atalay, and C. Gunduz-Demir, *Object-Oriented Segmentation of Cell Nuclei in Fluorescence Microscopy Images,* Cytometry Part A, vol. 93, no. 10, pp. 1019–1028, doi: 10.1002/cyto.a.23594, 2018.

[3] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, *DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs*, *arXiv*: arXiv:1606.00915. doi: 10.48550/arXiv.1606.00915, May 12, 2017.

[4] S. Zhang, Z. Ma, G. Zhang, T. Lei, R. Zhang, and Y. Cui, *Semantic Image Segmentation with Deep Convolutional Neural Networks and Quick Shift*, Symmetry, vol. 12, no. 3, p. 427, doi: 10.3390/sym12030427, Mar. 2020.

[5] M. Maška *et al.*, *The Cell Tracking Challenge: 10 years of objective benchmarking*, Nat Methods, vol. 20, no. 7, pp. 1010–1020, doi: 10.1038/s41592-023-01879-y, 2023.

[6] D. K. Gupta, N. de Bruijn, A. Panteli, and E. Gavves, *Tracking-Assisted Segmentation of Biological Cells*, *arXiv*: arXiv:1910.08735. doi: 10.48550/arXiv.1910.08735, Oct. 19, 2019.

[7] A. R. Menezes and B. Ramsundar, *Open Source Infrastructure for Automatic Cell Segmentation*, *arXiv:* arXiv:2409.08163. doi: 10.48550/arXiv.2409.08163, Sept. 12, 2024.

[8] G. Rahmon, I. E. Toubal, D. Cornelison, and K. Palaniappan, *Marker and Motion Guided Deep Networks for Cell Segmentation and Detection Using Weakly Supervised Microscopy Data*, in 2023 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), pp. 1–12. doi: 10.1109/AIPR60534.2023.10440692, Sept. 2023.
[9] A. Arbelle and T. R. Raviv, *Microscopy Cell Segmentation Via Convolutional LSTM Networks*, in 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019), Venice, Italy: IEEE, pp. 1008–1012. doi: 10.1109/ISBI.2019.8759447, Apr. 2019.
[10] M. Arvidsson, S. K. Rashed, and S. Aits, *An annotated high-content fluorescence microscopy dataset with Hoechst 33342-stained nuclei and manually labelled outlines*, Data Brief, vol. 46, p. 108769, doi: 10.1016/j.dib.2022.108769, Nov. 2022.
[11] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer Nature, 2022.
[12] Y. Wang, L. Yang, X. Liu, and P. Yan, *An improved semantic segmentation algorithm for high-resolution remote sensing images based on DeepLabv3+*, Sci Rep, vol. 14, no. 1, p. 9716, doi: 10.1038/s41598-024-60375-1, Apr. 2024.
[13] K. He, X. Zhang, S. Ren, and J. Sun, *Deep Residual Learning for Image Recognition*, *arXiv*: arXiv:1512.03385. doi: 10.48550/arXiv.1512.03385, Dec. 10, 2015.
[14] *Cell Tracking Challenge* – "Where your software moves cells…."
[Online]. Available: https://celltrackingchallenge.net/ Accessed: Nov. 24, 2025.
[15] R. C. Gonzalez and R. E. Woods, *Digital image processing*, Fourth edition, Global edition. New York: Pearson, 2017.
[16] K. Zuiderveld, *Contrast Limited Adaptive Histogram Equalization*, Elsevier, pp. 474–485. doi: 10.1016/B978-0-12-336156-1.50061-6, 1994.
[17] *ImageNet*. Available [Online]: https://image-net.org/, Accessed: Nov. 24, 2025.

DOI: 10.2478/ijasitels-2025-0015

Emergent Strategies in Multi-Agent Systems

Vlad OLEKSIK[1], Cantemir MIHU[1], Antoniu PITIC[1]
[1]Computer Science and Electrical and Electronics Engineering Department, Faculty of Engineering, "Lucian Blaga" University of Sibiu, Romania
{vladandrei.oleksik, cantemir.mihu, antoniu.pitic} @ulbsibiu.ro

Abstract

This article aims to present an overview of the emerging protocols and strategies used for achieving scalability of performance- and efficiency-centric multi-agent systems as this field is rapidly developing and focus is shifting towards architectures employing multiple, tool-enabled models that need to both split tasks, working concurrently, and follow step-based problem-solving strategies, while maintaining effective communication throughout the system.

Keywords: Multi-agent system, memory, agentic workflow, evaluation, communication, dispatching.

1. Introduction

A significant part of today's development in terms of machine learning applications comes from the recent waves of breakthroughs in the field of large language model (LLM)-based systems. The emergence of LLMs has started with the perfection of the transformer with the attention mechanism [1], a component that represents the functional unit of a system that enables 'talk' between different features of data streams situated at a significant distance away.

However, as information recollection and filtering became a problem in the context of large time costs for cold-start finetuning, the idea of delegating long-term memory to a different service resulted in the introduction of the concept of retrieval-augmented generation, employing embedding models to point the 'attention' of the model towards a specific subset of given data and introducing this data straight into the context window of the model to obtain much improved results. Despite this practice being rapidly adoptable, multiple different paradigms for data storage exist today, from centralized vector stores to distributed constructs.

Moreover, the finite number of rounds of transformations, as well as some intrinsic properties of language involved in some problem-solving steps led to a problem as most of the computation essential to solving a problem would be focused into just several predictions [2], which led to the problem of task planning, with different 'reasoning' schemes being introduced to encourage more generation in order to enable the model to use as many steps as necessary to reach the desired output.

In this context, the lingering limitations of single models marked, in part, by the scarcity of new and high-quality training data have opened the door to the use of agentic paradigms, in which a model is given access to tools for information processing and retrieval and imposed planning paradigms to solve a problem. The emergence of multi-agent paradigms, characterized as a multi-level, multi-cluster network of agents exchanging and processing data, has thus furthered the horizons of high-performance, as well as high-efficiency systems for both reasoning and simpler processing tasks.

To date, there is little standardization in the protocols used in task delegation, evaluation, planning and memory management, [3] while specific use-cases are benefiting from particular architectural traits and more paradigms are still being introduced. Thus, the aim of this paper is to provide a representative overview of the different approaches and paradigms present in modern LLM systems and their applications, as well as a comparison between the performance of different approaches for benchmarks representing various high-applicability tasks.

2. Theoretical concepts

This section outlines the main sources of variation in multi-agent system design, both from a technical perspective and in terms of system integration.

As Yan et. al. presented in their survey [4] of such systems, with a highlight, among others, on communication, specific approaches emerge as solutions for the previously described problems. Yan et. al. challenges the previous classification of design "patterns" into *agent-specific* and *architectural* strategies, highlighting the need for a paradigm to address the highly varying structural and functional traits of architectures.

As described in this paper, along with studies such as the one published by Han et. al. [3], some of the most relevant aspects to consider in a multi-agent LLM system architecture and implementation are task delegation, communication, planning and tooling, and memory management. Each of these will be briefly discussed in the following subsections.

2.1. Task delegation

With the introduction of multiple LLMs in an agentic workflow, one of the first issues that arises is the topology to be used for creating and handling these agents. Some of the approaches represent a variation on the hierarchical model, where each agent can generally spawn and manage other agents, tasked with solving a subproblem of the current goal, while others resemble a horizontal, clustered structure, where models associate based on certain criteria to solve problems and shift these themes over time.

2.1.1. Hierarchical structures

This first approach resumes to the protocol of an agent being able to add several others at each point, up to a certain depth, each one being tasked with a subdivision of the task of the parent agent.

Thus, higher-level agents will perform management and supervision tasks, while lower-level ones will be relied upon for detailed, stepwise problem-solving actions [4]. As a characteristic of this strategy, when working with specific tasks requiring various, well-defined subtasks and/or roles, multiple different agent types can be instantiated to fit the problem with high flexibility and very little friction (Fig. 1).

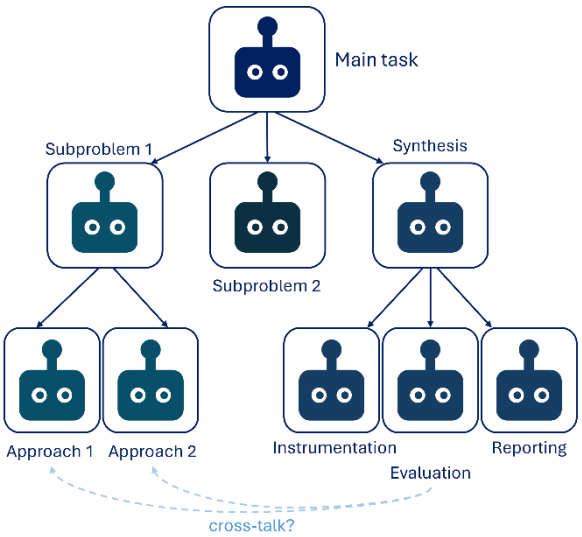


Fig. 1. The hierarchical (compartmentalized) model for multi-agent systems.

This is used in many versatile frameworks, however, the main limitation of this scheme is the lack of flexible and scalable opportunities for communication among agents, which affects the unitary structure of results through compartmentalization.

The Mixture-of-Experts topology [5] is an example of such a hierarchy, encountered with a model designated to decide what models (called “experts”) to dispatch for each input.

2.1.2. Homogeneous and clustered structures

As an alternative to using the hierarchy as a state and structure handling tool, many less hierarchical task delegation topologies exist, the main common element of which is the existence of a horizontal component to communication – that is, the ability of an agent to refer and report to agents not directly overseeing its activity (Peer-to-Peer structures).

These approaches vary from a fully-horizontal (homogeneous) model, to more structured approaches, where, despite it being possible to communicate with any other model, subgroups (teams/clusters) can be distinguished among the agents for certain tasks (Fig. 2).

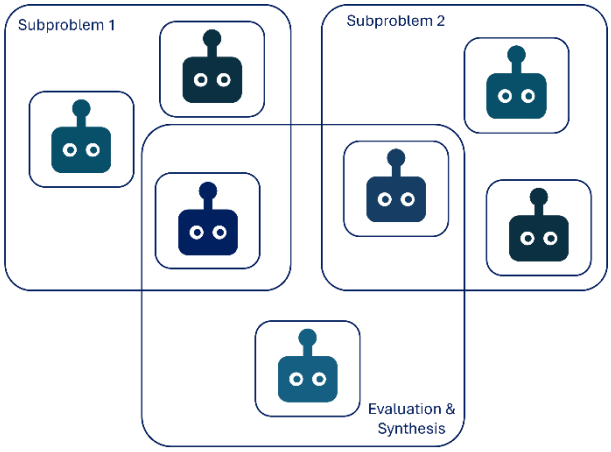


Fig. 2. Homogeneous models in multi-agent LLM systems

The downside of such systems is that it is often more difficult to exercise a given planning or structure for solving a given problem. Thus, elements of planning from the hierarchical model might be observed in some of the systems studied.

2.2. Memory management

Another mandatory feature of a performant multi-agent system is a form of information storage and retrieval mechanism. This needs to be implemented in order to enhance the relatively short working memory (context window) of a single agent, while being able to work in an environment of highly specific or quickly updating information. While most applications use a centralized data store – typically a vector store – some decentralized alternatives are better suited for scenarios involving rapid, localized updates that need to propagate only to a small subset of relevant agents.

2.2.1. Centralized stores

The idea of a vector database emerged in industry implementations for various similarity search algorithms, relying on previously well-known multi-dimensional data storage and retrieval structures [6]. In the context of multi-agent systems, it represents a database designed to retrieve data based on multi-dimensional vector key-value similarity, where all agents are free to read from and writes performed to this medium have global effect on the system (Fig. 3).

Integration of such a store into the agentic workflow is done as a tool/oracle that can be accessed (called) by an agent in a specific context. For the purpose of this study, a memory system is considered centralized not specifically from the point of view of the database implementation but based on the visibility of documents for specific agents.

Vector databases are a crucial component of most retrieval-augmented generation implementations and are, by far, the most widespread approach [7] for multi-agent systems, due to their high reliability and ease of integration into the vast majority of systems, regardless of their planning workflow and communication topology.

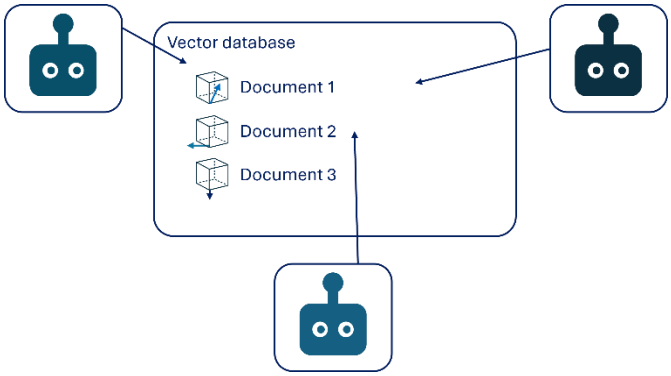


Fig. 3. Centralized memory approach

2.2.2. Decentralized topologies for memory management

Considering many of the multi-agent LLM systems have to handle growing complexity in order to produce quality results, overcoming limitations of particular agents, alternative memory management systems can not only help control the complexity when introducing many agents, but also limit the propagation of bad reasoning and low-quality documentation in tasks concerning only a subset of agents.

In this regard, a variety of approaches can be found across implementations in practice, with some topologies splitting the information stored among nodes/agent memories and submitting information to a propagation process for retrieval (Fig. 4).

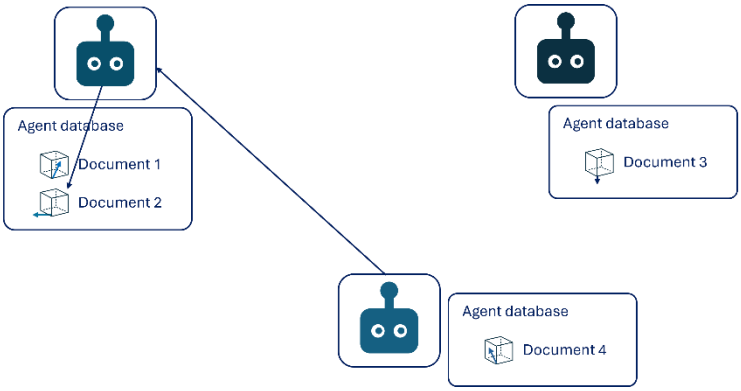


Fig. 4. The decentralized storage architecture

One emerging topology belonging to the class of decentralized memory management in multi-agent systems is an implementation relying on the concept of transactive memory. Stemming from a model in psychology proposed by Wegner [8], this structure translates in the current context to a specific workflow for updating the memory within the scope of each agent, as well as a mechanism for heuristically querying the system, relying on insight into which node is the most likely to hold particular pieces of information [9]. Still, while this aspect can enable scalability, few models, like the one proposed by Shang et. al. [10] have implemented this specific memory model to date.

2.3. Planning and tooling

Although not unique to multi-agent workflows, the concept of planning – that is, encouraging or enforcing a defined sequence of steps for each agent when generating a response – can enhance the output quality of individual models. Moreover, this approach can serve as a bridge from single-model systems to multi-agent settings by enforcing structured interaction through traditional interfaces and inter-agent communication alike.

A thorough overview of planning frameworks is presented by Aratchige and Ilmini [11] in a review encompassing an array of schemes for planning ranging from static (same flow of steps repeated for every stage of a problem) to dynamic (updating based on context, in designs such as AdaPlanner [12]) and from sequential (such as the ReAct framework [13], where steps are invariably executed one after another) to concurrent (such as the Tree of Thoughts framework [14], which introduces the possibility to handle multiple reasoning paths).

2.4. Communication and evaluation

While the problem of choosing a communication topology is partly influenced by and treated in the choice of task delegation framework, there are aspects specific to structuring of information passed among agents that significantly impact the functionality of a multi-agent system. As the complexity of the system increases with that of the tasks and the overlap between multiple agents begins to decrease in size, the methodology used to perform evaluation and to achieve consensus becomes of growing importance.

Among the frameworks already implemented, the treatment of the evaluation question is split between it being performed by the supervisor agent, shall one exist, and a more horizontal evaluation, performed by a subset of the other agents, according to a specific topology. It is of note that, while similar to the hierarchy problem described in 2.1., the approaches used for task delegation and evaluation are not necessarily correlated in the studied implementations.

In multi-agent systems – given their diverse and often complex topologies – it's important to distinguish between two notions: the local processes of voting or evaluation used to assess the quality and relevance of an individual agent's output, and the global consensus that must be maintained across the network to produce a coherent, unified result. In order to achieve the latter, techniques can be employed relying on expander graphs [15] to study and scalably enforce such convergence in the generation process.

Another issue regarding the communication in multi-agent LLM systems is that of the scope: while most implementations employ cooperative agents, there are notable examples of adversarial systems, such as the one proposed by Zhang and Eger [16]. The latter are often engaged in architectures where output generated by competitor

models is judged by a third agent, emulating feedback extrapolated from select human input.

A factor not extensively studied is that of the format of communication among agents. Many paradigms are in place for multi-agent systems, with Model Context Protocol (MCP) being the most used one in practice in the context of a single agent interacting with its available functional tools, while Agent-to-Agent (A2A) is among the most prominent protocols for handling inter-agent communication. While most frameworks employ plain-text communication or text-based tags for issuing signals or events, little is studied about the passing of embeddings from agent to agent as "paralinguistic" information. As a multi-agent system scales upwards, such information can serve as a mechanism to facilitate consensus by regulating the flow and scope of inter-agent communication.

3. Representative Multi-Agent Architectures

Recent advances in multi-agent LLM systems have led to a growing interest in frameworks to guide and integrate multiple LLM instances' cooperation or competition for solving complex tasks [3]. These systems seek to exploit specific emergent behaviors of distributed reasoning, division of labor, and interactive problem solving. In this context, recent research has focused mainly on defining scalable architectures, novel, efficient communication protocols, and structured memory mechanisms to improve the performance and interpretability of multi-agent LLM systems.

A detailed examination of the scalability problem forms the foundation of the MegaAgent architecture, proposed by Wang et al. [17]. This framework is designed to achieve relatively high efficiency on tasks of arbitrary complexity through a hierarchical team-based model. In MegaAgent, GPT-4o-based agents operate within a unified shared memory and employ a divide-and-conquer approach to task delegation. An important design choice is the prescriptive communication strategy: interactions occur exclusively along the vertical hierarchy or within localized groups. While this design enhances control and scalability, it can reduce the semantic cohesion of the group's collective reasoning. Despite these trade-offs, MegaAgent achieves competitive results with up to approximately 600 active agents, demonstrating one of the most scalable implementations reported to date.

A similar direction is explored in AgentVerse, introduced by Chen et al. [18], which represents a significant and methodically constructed contribution to the field. The authors distinguish between vertical and horizontal organizational structures, showing that a vertically layered arrangement benefits analytical tasks such as mathematical reasoning and software development, while a Peer-to-Peer structure performs better for decision-making and consultation scenarios. AgentVerse also contributes dedicated benchmarks for evaluating multi-agent collaboration patterns, establishing a valuable foundation for comparative research in this rapidly evolving domain.

In contrast to Q. Wang's approach, TalkHier, proposed by Zh. Wang and Moriyama [19], prescribes a rigorously defined hierarchical task organization coupled with a structured communication topology. The system integrates hierarchical evaluation mechanisms to assess intermediate outputs at multiple levels of the hierarchy. Also implemented using GPT-4o as the underlying model, TalkHier is claimed to outperform the selected baselines, highlighting the effectiveness of combining structured dialogue with systematic oversight. This reinforces the notion that well-balanced topologies and supervisory hierarchies can enhance the quality and coherence of distributed reasoning processes.

Based on current research, multi-agent systems are increasingly characterized through a diverse set of attributes drawn from multiple domains, including distributed computing, complex adaptive systems, organizational theory, and collective intelligence, traits largely dependent on the type of structure it uses. Table 1 captures the impact system topologies have on such systems' ability to coordinate, scale, adapt, and maintain robustness. Different system topologies — such as centralized, decentralized, or peer-to-peer structures — exhibit these qualities to differing degrees.

Table 1. Qualitative assessment of traits for the main system topologies

Attribute / Quality	Centralized	Hierarchical	Peer-to-Peer
Scalability	Medium	Good	Good
Coordination Overhead	Low	Medium	High
Fault Tolerance	Poor	Medium	Good
Communication Efficiency	Excellent	Good	Medium
Decision Latency	Low	Medium	High
Control Transparency	Excellent	Good	Poor
Specialization Support	Medium	Excellent	Good
Adaptability / Flexibility	Poor	Good	Excellent
Conflict Resolution	Excellent (via hub)	Good (via manager)	Poor
Robustness to Noise / Error	Poor	Medium	Good
Implementation Complexity	Low	Medium	High
Learning Convergence	Good	Good	Variable
Explainability / Auditability	Excellent	Good	Poor
Emergent Behavior Potential	Poor	Medium	Excellent

Regarding the use of decentralized memory in multi-agent architectures, Shang et al. [10] present a detailed implementation that integrates a shared, well-defined transactive memory topology governed by moderation-based communication. The system, as proposed, employs Gemini models of varying sizes and includes a thorough evaluation across multiple benchmarks. However, this framework lacks a structured evaluation component and does not support dynamic restructuring of the agent network during execution. Moreover, the authors do not employ heterogeneous agent scales (i.e., models of different sizes within a single system) nor maintain interpretability logs that trace the system's internal reasoning steps. These omissions point toward several promising directions for future research, including adaptive team composition, explainability mechanisms, and dynamic reconfiguration of agent hierarchies.

From a communication standpoint, among the models studied and presented, the overwhelming majority use cooperation as the main tactic for generation, as opposed to competitiveness-based frameworks. One common factor of these approaches is represented by the low computational overhead needed to implement such a framework, over the design challenges of rigorous formats needed to achieve quality consensus with a limited number of agents (among which the most commonly used formats are inspired by voting or debating) or the scalability issues of evolutionary systems that need to employ a multitude of agents and a referee system, such as the model proposed by [20].

A developing line of inquiry involves interpreting multi-agent LLM systems from the point of view of criticality theory**.** In the context of multi-agent LLM architectures, this concept has been used to characterize the phase transitions between under-coordinated (fragmented) and over-constrained (rigidly hierarchical) interaction regimes. Hierarchical and modular systems such as MegaAgent and TalkHier tend to operate below the critical point, favouring stability and control but limiting emergent creativity. Conversely, horizontally-structured or loosely-moderated environments, as explored in AgentVerse, move closer to the critical regime, where agent interactions can generate novel behaviours at the cost of reduced predictability.

Overall, current multi-agent LLM research, while proving rapid progress, creates significant open areas of interest. While hierarchical and modular structures—such as those implemented in MegaAgent, AgentVerse, and TalkHier—enable controlled scaling and improved efficiency, they also risk constraining creativity and reducing emergent coordination. Meanwhile, systems employing transactive or shared memory, such as the work by Shang et al. [10], show promise for enhancing collective recall and continuity, yet still face limitations in adaptability and transparency.

4. Synthesis and conclusions

The present paper has presented an overview of the most relevant framework architectures for multi-agent LLM systems, classifying them from various points of view, such as architecture, tooling, and communication.

While many frameworks exhibit features not fully inscribed in the proposed classification, a discussion section has aimed to address the most significant talking points of their architectures.

The study identified the most widely used models to be hierarchical ones, featuring centralized memory and sequential planning for reasons of flexibility. Nevertheless, limitations of these models open the possibility for use-case-specific architectures. This paper has also identified as possible future study areas the use of transactive memory systems, of diversifying the communication formats, as well as of graph-based evaluation schemes for highly scalable, decentralized frameworks.

Future work is also expected to focus on interpretable coordination strategies, dynamic scalability, and the development of new metrics to support design principles for next-generation multi-agent LLM ecosystems.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, *Attention is all you need,* Advances in Neural Information Processing Systems, **30**, Curran Associates, Inc., 2017.

[2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, *Chain-of-thought prompting elicits reasoning in large language models*, Proceedings of the 36th International Conference on Neural Information Processing Systems, pp. 24824–24837, Red Hook, New York, 2022. https://dl.acm.org/doi/10.5555/3600270.3602070

[3] S. Han, Q. Zhang, Y. Yao, W. Jin, Z. Xu, C. He. *LLM Multi-Agent Systems: Challenges and Open Problems*, preprint, arXiv:2402.03578, 2024. https://doi.org/10.48550/arXiv.2402.03578

[4] B. Yan, X. Zhang, L. Zhang, L. Zhang, Z. Zhou, D. Miao, C. Li, *Beyond Self-Talk: A Communication-Centric Survey of LLM-Based Multi-Agent Systems*, preprint, arXiv:2502.14321, 2025. https://doi.org/10.48550/arXiv.2502.14321

[5] S. Mu, S. Lin, *A Comprehensive Survey of Mixture-of-Experts: Algorithms, Theory, and Applications*, preprint, arXiv:2503.07137, 2025. https://doi.org/10.48550/arXiv.2503.0713

[6] L. Ma, R. Zhang, Y. Han, S. Yu, Z. Wang, Z. Ning, J. Zhang, P. Xu, P. Li, W. Ju, C. Chen, D. Wang, K. Liu, P. Wang, P. Wang, Y. Fu, C. Liu, Y. Zhou, C.-T. Lu. *A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge*, preprint, arXiv:2310.11703, 2021.

[7] J. Pan, J. Wang, G. Li, *Survey of Vector Database Management Systems*, VLDB J., **33**, pp. 1591-1615, 2023. https://doi.org/10.48550/arXiv.2310.14021

[8] D. M. Wegner, *Transactive memory: A contemporary analysis of the group mind*, Theories of group behavior, pp. 185–205, Springer-Verlag, New York, 1986.

[9] M. Catasta, A. Tonon, D. E. Difallah, G. Demartini, K. Aberer, P. Cudré-Mauroux, *TransactiveDB: Tapping into Collective Human Memories*, Proc. VLDB Endow., **7** (14), pp. 1977-1980, 2014. https://www.vldb.org/pvldb/vol7/p1977-catasta.pdf

[10] H. Y. Shang, X. Liu, Z. Liang, J. Zhang, H. Hu, S. Guo, *United Minds or Isolated Agents? Exploring Coordination of LLMs under Cognitive Load Theory*, preprint, arXiv:2506.06843v1, 2025. https://doi.org/10.48550/arXiv.2506.06843

[11] R.M. Aratchige and W. M. K. S. Ilmini, *LLMs Working in Harmony: A Survey on the Technological Aspects of Building Effective LLM-Based Multi Agent Systems*, preprint, arXiv:2504.01963, 2025. https://doi.org/10.48550/arXiv.2504.01963

[12] H. Sun, Y. Zhuang, L. Kong, B. Dai, C. Zhang, *AdaPlanner: Adaptive Planning from Feedback with Language Models*, preprint, arXiv:2305.16653, 2023. https://doi.org/10.48550/arXiv.2305.16653

[13] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, *ReAct: Synergizing Reasoning and Acting in Language Models*, preprint, arXiv:2210.03629, 2022. https://doi.org/10.48550/arXiv.2210.03629

[14] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, K. Narasimhan, *Tree of Thoughts: Deliberate Problem Solving with Large Language Models*, preprint, arXiv:2305.10601, 2023. https://doi.org/10.48550/arXiv.2305.10601

[15] L. Becchetti, A. Clementi, E. Natale, *Consensus Dynamics: An Overview*, ACM SIGACT News, **51** (1), pp.57, 2020. https://doi.org/10.1145/3388392.3388402

[16] R. Zhang, S. Eger, *Llm-based multi-agent poetry generation in non-cooperative environments*, preprint, arXiv:2409.03659, 2024. https://doi.org/10.48550/arXiv.2409.03659

[17] Q. Wang, T. Wang, Z. Tang, Q. Li, N. Chen, J. Liang, B. He, *MegaAgent: A Large-Scale Autonomous LLM-Based Multi-Agent System Without Predefined SOPs*, preprint, arXiv:2408.09955, 2024. https://doi.org/10.48550/arXiv.2408.09955

[18] W. Chen, Y. Su, J. Zuo, C. Yang, *AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors in Agents*, preprint, arXiv:2308.10848, 2023. https://doi.org/10.48550/arXiv.2308.10848

[19] Z. Wang, S. Moriyama, W.-Y. Wang, B. Gangopadhyay, S. Takamatsu, *Talk Structurally, Act Hierarchically: A Collaborative Framework for LLM Multi-Agent Systems*, preprint, arXiv:2502.11098v1, 2025. https://doi.org/10.48550/arXiv.2502.11098

[20] J. Cai, J. Li, M. Zhang, M. Li, C.-S. Wang, K. Tei, *Language Evolution for Evading Social Media Regulation via LLM-based Multi-agent Simulation*, IEEE Congress on Evolutionary Computation (CEC), 1–10, 2024..

DOI: 10.2478/ijasitels-2025-0017

Building a power meter for cycling

Raimund TOMA[1], Macarie BREAZU[1]

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

Abstract

The main goal of this research was to design and implement a simple power meter for cycling. The designed power meter had to be affordable, so we chose to use the crank arm solution. In order to achieve the objectives, we had to measure the tangential force applied on the crank arm using tensometric stamps. Also, we had to measure the cadence based on zero-crossing of the applied force, and to transmit data to external applications using Bluetooth Low Energy (BLE). The hardware implementation is based on BF350-3AA tensometric stamps, HX711 amplifier and Seeed XIAO nRF52840 Sense microcontroller, while the software implementation is based on the HX711_MP.h and bluefruit.h libraries. Experiments prove that the proposed solution (cost approx. 29€) meets the requirements. After calibration, our system achieved a deviation of ~10% compared to Tacx virtual power data, confirming its feasibility for amateur training applications.

Keywords: power meter, cycling, tensometric stamps, crank arm, force, cadence

1 Introduction. The importance of measuring the power

Modern cycling, whether recreational or competitive, has benefited in recent years from advanced technologies, such as power meters, which measure the power generated by the cyclist, providing essential data for optimizing performance. These tools provide objective data on effort, helping cyclists optimize their training and compete more effectively, but their high costs often make them inaccessible to amateurs. This paper explores the development of an affordable power meter, designed to bring the benefits of this technology to a wider audience, without claiming the accuracy of top-of-the-line commercial devices.

The objectives of this research include:

- Designing a simple power meter prototype, using affordable components to measure power and cadence.
- Implementing a hardware and software system to process and transmit data via Bluetooth Low Energy (BLE) to external applications.
- Testing the prototype (on a trainer), comparing the data with the virtual power values provided by the manufacturer (TACX), to evaluate the functionality.

2 State-of-the-art in power meters for cycling

Since the 2000s, the power meter market has diversified with the emergence of new technological solutions that have reduced costs and simplified installation [1]. In 2006, PowerTap introduced a power meter integrated into the rear wheel hub. In 2012, Stages Cycling launched a single-sided power meter, mounted on the crank arm. The 2010s brought significant advances in portability and accuracy, with the launch of pedal-based power meters, such as the Garmin Vector (2013) and Favero Assioma (2017) [2]. Additionally, the integration of wireless protocols, such as ANT+ and Bluetooth Low Energy, has facilitated connectivity with bike computers and virtual cycling applications [3], [4].

Commercial power meters are classified by the location of the sensors used to measure power, with each type having specific advantages and limitations in terms of accuracy, portability, and bicycle compatibility [1]. The main categories include systems mounted on the bottom bracket, in the pedals, in the rear wheel hub, and other solutions such as in the gear or optical sensors, each tailored to different needs and budgets [5].

Crank-arm-mounted power meters, such as Stages and SRM, measure power by detecting the force applied to the crank using tensometric stamp sensors [6]. Advantages of these systems include compatibility with most bicycles and durability, but disadvantages include high costs for bilateral models and possible errors with unilateral systems. [5]. These power meters are popular both in professional cycling, due to their accuracy, and in the amateur segment, due to the affordability of models like Stages [6].

Pedal-mounted power meters, such as the Favero Assioma, Garmin Vector, and PowerTap P1, are popular for their portability and ease of installation, allowing for quick transfer between bikes. Advantages include portability and the ability to analyze pedaling asymmetries, but disadvantages include relatively high cost ($500–1000) and vulnerability of the pedals to wear or impact. [5].

Hub-mounted power meters, such as the PowerTap, measure power after chain and drivetrain losses, resulting in readings that are 5–10 W lower than those obtained at the pedals or gears [7]. Advantages include moderate cost (around $500) and reliability in a variety of conditions, but disadvantages include lack of portability (the hub is wheel-specific) and underestimation of power due to drivetrain losses. [5].

In addition to the main categories, there are gear-mounted (spider-based) power meters, such as Quarq and Rotor INpower, which measure power at the chainring level, offering accuracy comparable to SRM (±1÷2%) and compatibility with various bicycles. Other solutions include optical sensors, such as Ergomo, which use monoblock deformation to estimate power, but were abandoned due to high errors (over 5%) [8]. Systems based on "opposing forces", such as PowerPod, measure aerodynamic drag, slope and speed to infer power, without requiring sensors on the bike (error 5÷10%) [9]. Other experimental technologies, such as tire pressure sensors (TPS), are in development, but have not yet reached a commercial version [10].

The accuracy, precision, and efficiency of commercial power meters vary depending on the technology and usage [1]. Accuracy reflects the closeness to the true value of

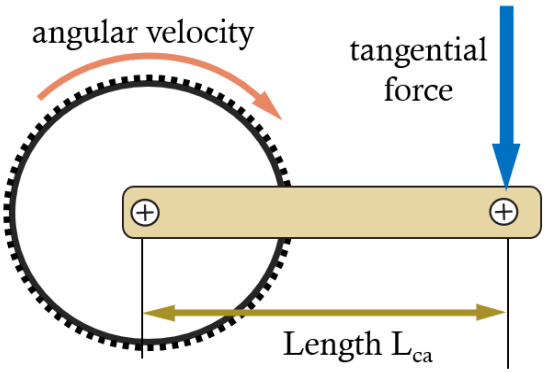


Figure 1. Force applied on the crank arm

power, precision indicates the consistency of measurements, and efficiency refers to the reliability of data with minimal energy consumption and reduced maintenance [5].

Comparative studies evaluate power meters against the SRM standard (±1-2% accuracy), and indicate significant variability among commercial power meters, from accuracy of ±1-2% to ±5-10% [11], [7], [9]. Maier et al. tested 54 units, reporting a mean deviation of -0.9% (standard deviation ±3.2%), but with some models deviating by more than ±5% [11]. Kirkland et al. determined that the contribution of work from each leg was 48.89% from the left leg and 51.11% from the right leg [8].

Efficiency varies depending on portability, battery and maintenance, with pedal-powered systems being the most versatile. Cyclists should perform regular calibrations and choose a model that suits their needs (e.g. bilateral for precision, unilateral for budget).

3 Theoretical basis for power meter on crank arm

The instantaneous power generated by a cyclist pressing on a crank arm (Fig. 1) is given by

$$P_{inst} = \tau * \omega \tag{1}$$

where P_{inst} represents the instantaneous power, τ the torque and ω the angular velocity. Further we have

$$\tau = F_T * L_{ca} \tag{2}$$

where F_T represents the tangential force (the only force component that develops power) and L_{ca} the length of the crank arm.

Instead of using ω it is more common to rely on the cadence RPM (Rotations Per Minute), so we have

$$\omega = RPM * \frac{2*\pi}{60} \tag{3}$$

The instantaneous power becomes

$$P_{inst} = F_T * L_{ca} * RPM * \frac{2*\pi}{60} \tag{4}$$

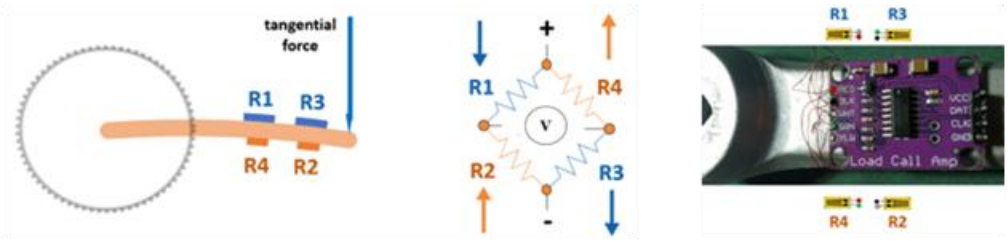


Figure 2. Arrangement of tensometric stamps in a complete Wheatstone bridge

When (as usual) we are interested in the average power we must consider the average tangential force. To maintain the hardware implementation as simple and cheap as possible, we will measure the force only on a single crank arm (as described later), then double the result (to factor in both pedals). In our case L_{ca} = 0.170 m.

4 Design of the power meter

The project aimed to create an accessible solution for real-time measurement of power and cadence in cycling, intended mainly for amateur cyclists. The system measures the force applied to the left crank arm and calculates the generated power, estimating the total value by doubling the contribution of the left leg. Data is transmitted wirelessly via Bluetooth Low Energy [4], complying with the Cycling Power Service standard [12] for direct compatibility with virtual cycling applications and existing cycling computers. The main goal was to experiment and test the feasibility of a functional and significantly cheaper alternative to commercial power meters, without claiming professional-level accuracy.

4.1 Crank arm with Wheatstone bridge

The approach of installing sensors on the left crank arm aims to directly measure the elastic deformation associated with the tangential component of the pedaling force, the only one responsible for generating useful torque. The mechanical configuration uses four resistive tensometric stamps arranged in a complete Wheatstone bridge, oriented "in an X" on the upper and lower faces of the arm. This arrangement, shown in Fig. 2, maximizes bending sensitivity, reduces temperature effects by exposing all stamps similarly, and increases the signal-to-noise ratio required for robust detection of useful force.

In this architecture R1 and R3 are positioned on the upper surface, and R2 and R4 on the lower one, so that the tension and compression generated by bending produce opposite resistance variations and, therefore, double the output of the bridge, simultaneously compensating for thermal influences and lateral forces. When the pedal is pressed, the upper surface stretches (R1 and R3 resistance values decrease) and the lower surface compresses (R2 and R4 resistance values increase), all four tensometric stamps contributing to the differential signal.

The integration of tensometric stamps on the crank arm requires careful surface preparation, precise alignments with the main stress directions, as various factors (such as orientation errors or uneven adhesive thickness) can induce parasitic sensitivity and

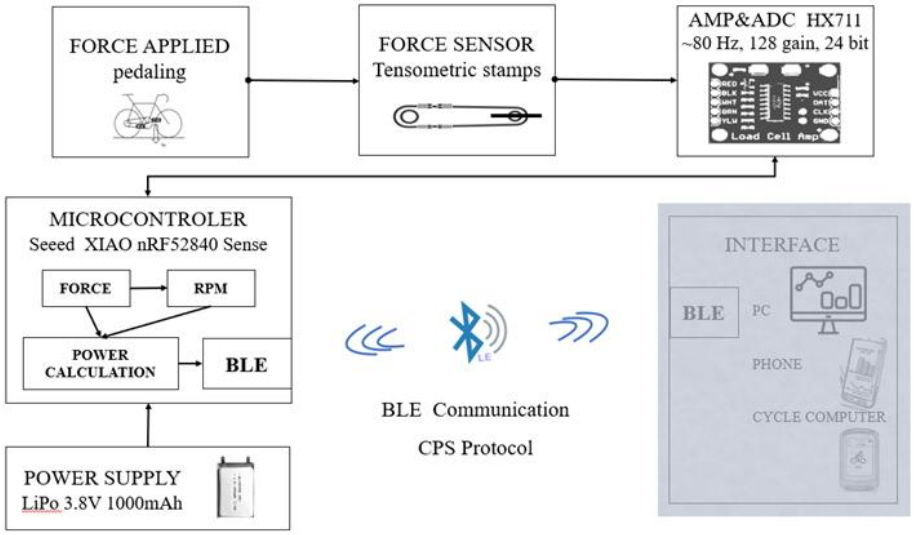


Figure 3. Block diagram of the designed system

drift. Choosing unilateral mounting (left crank arm) simplifies integration and costs, accepting that total power can be estimated by doubling the measured value, assuming approximately symmetrical pedaling. In practice, deviations from symmetry can introduce systematic errors, an aspect to be considered in calibration and data interpretation [5].

4.2 Hardware design

The measurement process is a classic for low-level force applications: a Wheatstone bridge with four tensometric stamps feeds an HX711 analog-to-digital amplifier/converter, whose data is processed by the Seeed Studio XIAO BLE nRF52840 Sense microcontroller and transmitted via BLE using the Cycling Power Service protocol. The block diagram of the designed system (Fig. 3) clearly illustrates the flow: force application → arm deformation → resistance variation in the bridge → amplification and sampling → processing and filtering → cadence/power calculation → CPS packaging and BLE transmission.

The HX711 provides 24-bit resolution, 128 programmable gain, and a sample rate configured at 80 SPS (SamplePerSecond). This rate is adequate for profiling force per rotation even at 180 RPM, when ~26 acquisitions per rotation are made (one reading at ~13.5°). Communication between the HX711 and the nRF52840 is performed through a two-wire serial interface (DAT and CLK), using a proprietary serial protocol implemented through bit-banging: the microcontroller generates the clock pulses, respecting the minimum setup and hold times (0.1 μs), and reading is triggered when the DAT pin transitions to LOW, signaling the availability of data. For robustness, the HX711 module is mounted on the inside of the crank arm, close to the axle, reducing the length of wires to the tensometric stamps and susceptibility to interference.

The XIAO BLE nRF52840 Sense central unit provides stable 3.3 V power supply (for HX711), has appropriate pinout for interfacing with HX711 (D2–CLK, D3–DAT) and includes an RGB LED for signaling statuses (initialization, BLE connection, errors). The platform was also chosen for its balanced combination of size, performance and superior energy efficiency compared to ESP32-based alternatives, with sleep

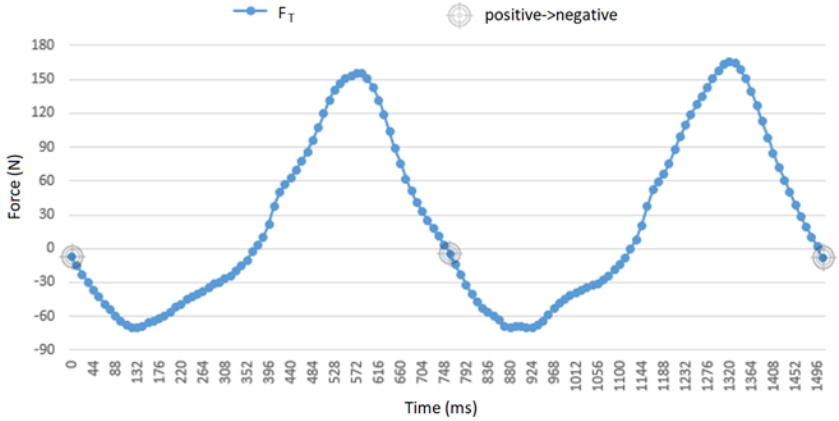


Figure 4. The force applied within an interval of two complete revolutions

consumption dropping below 2 µA, compared to ~43 µA (ESP32-C3) and ~15 µA (ESP32-C6), thus supporting portable applications with extended autonomy.

The system is powered by a 523450 LiPo battery with a capacity of 1000 mAh and a nominal voltage of 3.7 V, enough for approximately 10 hours of continuous use.

4.3 Measuring force and cadence

The acquisition of raw values from the HX711 is synchronized on the interrupt, when the DAT pin goes LOW. The conversion into physical force is achieved through a two-step calibration: offset compensation (zeroing) and application of an experimentally determined scaling factor. Following the tests, using the calibration method with the application of predefined forces, a scaling factor of 0.000165 N/unit was established. This conversion is paramount for subsequent torque and power calculations.

Cadence is determined by detecting zero-crossings of the calibrated force during rotation, using a hysteresis to avoid false triggering induced by noise and vibration. Threshold filters for RPM (realistic range 10 to 200 RPM) and power (10 to 400 W) eliminate noise and false rotations caused by oscillation at low speeds. Fig. 4 provides visual context for the signal dynamics and marking of rotation events used in the cadence calculation, illustrating the effect of filtering on consistent detection of zero crossings.

4.4 Software design

The software architecture is modular and mirrors a clear flow: initialization of subsystems (HX711, BLE, GPIO, LED), configuration of interrupts for data acquisition, main loop for signal processing (offset, scaling, filtering), detection of rotation events and calculation of cadence/power, followed by packaging and publishing of data via BLE. The libraries used include HX711_MP.h [13] for interfacing with the HX711 load cell amplifier and bluefruit.h [14] for the BLE stack. BLE callbacks handle connection/disconnection events, initialize services/features, and trigger notifications when new valid results occur. Thanks to the applied "publish-on-event" strategy, data is transmitted only when a complete rotation is detected, optimizing energy consumption and avoiding unnecessary traffic.

4.5 BLE protocol frame

Data transmission is done over Bluetooth Low Energy, using Cycling Power Service (CPS), a standardized GATT service for devices that measure cycling power. The BLE architecture involves the application-level GAP/GATT layers, the Attribute Protocol, and the L2CAP/HCI interfaces, which interact with the radio Link Layer; this structure ensures interoperability with cycle computers and software platforms [15]. Data is delivered via the Cycling Power Measurement feature (UUID: 0x2A63), configured with the notification property. The data frame has a fixed length of 8 bytes and is transmitted in little-endian format. The first two bytes represent the flags field; in the current implementation, the value 0x0020 is used, which sets bit 5 (Crank Revolution Data Present) and indicates the inclusion of crank revolution data. Bytes 2–3 contain the 16-bit signed instantaneous power (in watts), bytes 4–5 represent the total number of rotations, and bytes 6–7 encode the time of the last rotation event (in 1/1024s quanta). A complete example is shown in Table 1 for the following values; power = 251 W (0x00FB), total Crank Revs = 7 (0x0007), last Crank Event Time = 5120 (0x1400) and flags 0x0020.

Table 1 Creating a frame for BLE transfer

Octet	Value (Hex)	Description	
0	0x20	Flags (LSB)	0x0020
1	0x00	Flags (MSB)	
2	0xFB	Power (LSB)	251W
3	0x00	Power (MSB)	
4	0x07	Number of rotations (LSB)	7 rotations
5	0x00	Number of rotations (MSB)	
6	0x00	Event time (LSB)	5120 (=5 sec)
7	0x14	Event time (MSB)	

Notifications are only triggered when a full rotation has been detected and there is at least one active connection, and the blue LED is activated with each transmission for visual feedback. This implementation minimizes power consumption and ensures compliance with the CPS standard, thus facilitating interoperability with specific applications and devices.

4.6 Results

The prototype reached a sampling rate of 80 SPS, enough for capturing the force profile, thus at cadences of 180 RPM capturing ~26.7 samples per rotation. However, accuracy is influenced by the limited precision of the scaling factor and sensor drift, influenced by the adhesive used to attach the tensometric stamps, and by fluctuations in battery voltage. Initial dynamic tests revealed a discrepancy between the theoretically estimated power (100 W) and the values measured on the crank arm (60–70 W), a discrepancy attributed to the imprecision of the scaling factor. The tests were performed using the Tacx Blue Motion trainer [16] and the Tacx Training app, which estimates power based on speed and the known power curve for a given resistance level (100 W corresponding to 15.7 km/h in level 4). Subsequent more rigorous recalibration improved the accuracy, ultimately reaching a deviation of ~10%.

The usefulness of the system, even if it does not reach the precision of professional systems and does not provide exact information about the power exerted by the cyclist,

lies in its accessibility and the ability to provide real-time feedback in a format compatible with cycling devices and applications. Also, its consistent use in training provides amateur cyclists with a reproducible power metric over time, which allows for consistent monitoring of progress and comparison of training sessions, offering an excellent cost-benefit ratio.

5 Conclusions and future work

The development of a power meter involved the integration of hardware and software components in a dynamic environment, with additional requirements for cost, precision and robustness. The lack of specialized equipment, the variable quality of components, and the complex nature of cycling power measurement have generated significant challenges, both hardware and software.

The BF350-3AA stamps were attached to the crank arm with professional general-purpose adhesive, selected for its adhesion to metal surfaces. However, the adhesive exhibited slow elastic recovery at high loads, which influenced the stability of the readings.

Integrating the HX711 library with BLE communication was problematic. The standard HX711 library (Bogde) worked correctly in individual tests, but when combined with the BLE stack (bluefruit.h), the code stopped running, with no obvious compilation errors.

The HX711_MP.h library was adopted as a solution, but it exhibited sporadic resets at 10 SPS and frequent resets at 80 SPS, caused by overlapping HX711 interrupts with BLE radio activity. Changes to the library included reducing the interrupt blocking period (from a block for the entire loop to a block for a single instruction in the loop body), thus preventing the HX711 from entering "power-down" mode and resetting the microcontroller.

Table 2 Cost of the designed system

Component	Approx. Price (€)	Total cost (€)
Seeed XIAO	18	29
Battery 523450 LiPo	8	
HX711	2	
BF350-3AA stamps (all)	1	

Real-time testing and debugging were hampered by the inability to use the serial monitor while pedaling. The use of the BleUART protocol and BLE applications (e.g. nRF Connect, Bluefruit Connect) for debugging introduced a latency of ~200 ms per transmission, affecting real-time data acquisition.

The prototype, based on the Seeed XIAO nRF52840 Sense microcontroller, HX711 amplifier and BF350-3AA tensometric stamps, proved the feasibility of a DIY alternative to commercial power meters, offering promising results despite the technical challenges. A major advantage of the prototype is its significantly lower cost compared to commercial power meters, such as SRM or PowerTap, which cost hundreds or thousands of euros. The use of affordable components, with a total cost of ~29€ as detailed in Table 2, and open-source software (Arduino IDE, HX711_MP.h and bluefruit.h libraries) makes the prototype attractive for amateur cyclists. The prototype,

being much cheaper, can be easily multiplied, without the need for transfer when one owns multiple bikes.

Certainly, to improve the performance and reliability of the power meter, other developments should be considered in future, such as:

- Using a specialized adhesive and better quality tensometric stamps to minimize the observed springback, adding a second set of stamps on the second crank arm.
- Design of a robust, weatherproof housing (e.g. IP65 standard), to protect the components (HX711, microcontroller, battery) against dust, moisture and vibrations during pedaling. Possibly using a modular design, with a fixed part (the tensometric stamps and HX711, mounted on the crank arm) and a portable part (the microcontroller and battery), connected by a quick coupling system (e.g. magnetic connectors or pins).
- Implementing a microcontroller "deep sleep" mode during periods of inactivity and adding battery status monitoring and notification via BLE.
- Comparison of the prototype with a commercial reference power meter (e.g. SRM) under controlled conditions, to adjust the scaling factor and increase and validate the accuracy.

References

[1] H. Allen and A. Coggan, *Training and Racing with a Power Meter*, Boulder, CO, USA: VeloPress, 2010.

[2] Favero Electronics, *Assioma Technical Specifications and IAV Power Technology*, 2022. [Online] https://cycling.favero.com/en/blog/assioma/all/data

[3] ANT+ Alliance, *ANT+ Device Profile: Bicycle Power*, 2023. [Online]. https://www.thisisant.com/developer/ant-plus/device-profiles/#521_tab

[4] Bluetooth SIG, *Bluetooth Low Energy Overview*, 2023. [Online]. https://www.bluetooth.com/learn-about-bluetooth/tech-overview/

[5] A. Bouillod, J. Pinot, G. Soto-Romero, W. Bertucci, and F. Grappe, *Validity, Sensitivity, Reproducibility and Robustness of the Powertap, Stages and Garmin Vector Power Meters in Comparison With the SRM Device*, International Journal of Sports Physiology and Performance, vol. 11, no. 8, pp. 1077–1083, Nov. 2016.

[6] Stages Cycling, *Stages Power Meter User Guide*, 2023. [Online]. https://manuals.stagescycling.com/en/power-meter/stages-power-meter-user-guide/

[7] A. S. Gardner, D. T. Martin, D. G. Jenkins, and C. J. Gore, *Accuracy of SRM and PowerTap power monitoring systems for bicycling*, Medicine & Science in Sports & Exercise, vol. 41, no. 7, pp. 1259–1265, Jul. 2009.

[8] A. Kirkland, D. Coleman, J. Wiles, J. Hopker Validity and Reliability of the Ergomo®pro Powermeter, Int J Sports Med 2008; 29: 913–916

[9] P.F.J. Merkes, P. Menaspà, C.R. Abbiss, *Validity of the Velocomp PowerPod power meter in comparison with the Verve Cycling InfoCrank power meter*, International Journal of Sports Physiology and Performance, vol. 14, no. 10, pp. 1382–1387, Nov. 2019.

[10] N. J. Fiolo, H. Y. Lu, C. H. Chen, P. X. Fuchs, W. H. Chen, and T. Y. Shiang, *The Validity and Reliability of a Tire Pressure-Based Power Meter for Indoor Cycling*, Sensors, vol. 21, no. 18, p. 6117, Sep. 2021, doi: 10.3390/s21186117.

[11] T. Maier, B. Schmid, T. Steiner, and C. A. Wehrlin, *Accuracy of cycling power meters against a mathematical model of treadmill cycling*, International Journal of Sports Medicine, vol. 38, no. 6, pp. 456–461, Jun. 2017. [Online] https://www.thieme-connect.com/products/ejournals/pdf/10.1055/s-0043-102945.pdf

[12] Bluetooth SIG, *Cycling Power Service 1.1*, 2016. [Online] https://www.bluetooth.com/Specifications/specs/html/?src=CPS_v1.1/out/en/index-en.html

[13] R. Tillaart, *Arduino library for HX711 load cell amplifier, 24 bit ADC*. [Online]. https://github.

com/ RobTillaart/HX711_MP
[14] *Adafruit code for the Nordic nRF52 BLE SoC on Arduino.* [Online]. https://github.com/adafruit/Adafruit_nRF52_Arduino/blob/master/libraries/Bluefruit52Lib/src/bluefruit.h
[15] Analog, *Understanding the Architecture of the Bluetooth Low Energy Stack*, 2024. [Online] https://www.analog.com/en/resources/technical-articles/understanding-architecture-bluetooth-low-energy-stack.html
[16] *Features of the Tacx Blue Motion Basic Trainer.* [Online]. https://support.garmin.com/en-US/?faq=SJoC24KZo31tV6nKhBmH5A

DOI: 10.2478/ijasitels-2025-0019

A Deep Dive into Wavelet-Based Neural Architectures for Inpainting Forgery Detection

Adrian Alin Barglazan[1][0009-0009-4420-3639],
Remus Brad[1] [0000-0001-8100-1379],
Pitic Elena Alina[1],
Berghia Ștefania Eliza [1]

[1] *Lucian Blaga University of Sibiu*

Abstract

This paper provides a comprehensive analysis of several neural network architectures designed for the detection of generative image inpainting. The central challenge addressed is the increasing sophistication of inpainting technologies, driven by generative AI, which renders traditional forensic methods based on simple statistical or visual artifacts obsolete. This paper investigates a feature-centric hypothesis: that visually seamless inpainting manipulations introduce consistent, subtle, and detectable artifacts in specific high-frequency and statistical domains. The primary domain of interest is the feature space defined by the Dual-Tree Complex Wavelet Transform (DTCWT).

The core of this report details the methodology and evaluation of six novel deep learning architectures, each custom-designed to operate on these wavelet-domain features. These architectures explore diverse design philosophies, including dense feature aggregation (Custom UNet++), explicit global context integration (UNet++ with Global Average), recurrent spatial modeling (ConvLSTM), unsupervised generative anomaly detection (Variational Autoencoder), and channel-specific processing (StackDeepAll).

A rigorous experimental evaluation was conducted on a custom inpainting detection dataset. The quantitative results demonstrate the clear superiority of the StackDeepAll architecture. This model, which processes each of the 12 DTCWT channels through independent, parallel UNet-like encoders before fusion, achieved a state-of-the-art Intersection over Union (IoU) of 0.8126 in validation. This result significantly surpassed all other proposed models, including a stable Variational Autoencoder (0.3711 IoU) and overfitting-prone ConvLSTM models (0.4514 IoU). The findings establish that for high-dimensional, engineered features like complex wavelets, a channel-specific processing paradigm that avoids premature feature-blending is a powerful and highly effective design strategy. This feature-centric approach is validated as a potent alternative to contemporary end-to-end spatial-domain models.

Keywords: image inpainting; forgery detection; digital image forensics; wavelet-domain features; deep neural architectures; Dual-Tree Complex Wavelet Transform

1 Introduction

The field of digital forensics is in a constant state of escalation against the rapid advancement of digital media manipulation. In recent years, the "accelerated progress of inpainting technologies" [1] has created an "urgent need for robust forgery detection methods" [2, 3]. Driven by powerful generative AI tools, including Generative Adversarial Networks (GANs) and modern diffusion models, image inpainting, the process of reconstructing missing or removing undesired regions in an image, has evolved from a simple restoration tool into a sophisticated method for creating "highly realistic" or "photorealistic" alterations [2, 3].

These advanced techniques, capable of synthesizing semantically coherent and contextually appropriate content, now "challenge traditional forensic tools" [1]. Conventional detection methods, which traditionally rely on identifying discernible artifacts such as blurring, color inconsistencies, or unnatural edges, are progressively failing. As the generative models producing these manipulations become more "unprecedented[l] realistic", the "visible signs" of tampering are effectively erased, forcing the forensic community to seek out new, more subtle forensic signals [4, 5].

The failure of traditional methods necessitates a deeper investigation into the fundamental artifacts introduced by inpainting, even by the most advanced generative models. An extensive evaluation revealed that while classical inpainting techniques, such as diffusion-based or patch-based algorithms, are "relatively easy to detect," images altered using modern machine learning-based methods present a "greater challenge" [1]. On these challenging datasets, even the most effective conventional detection methods were found to achieve an average Intersection over Union (IoU) and F1 score below 35%, highlighting a significant capability gap [1].

However, this analysis also revealed that inpainting artifacts, while visually imperceptible, are not non-existent. Instead, they manifest subtle but "statistically significant deviations in the texture's structure" [1]. This finding motivated an exploration of feature domains specifically sensitive to such structural and textural irregularities. This exploration identified three key forensic signals: complex wavelets, semantic segmentation, and the analysis of noise level inconsistencies [1].

Based on this hypothesis, a novel classical (non-machine learning) detection method was developed, which integrated these three signals. By performing semantic segmentation, extracting complex wavelet features, and analyzing noise inconsistencies at a segmental level, this classical algorithm was able to achieve an overall IoU of approximately 53% [1]. This result, while a significant improvement over the 35% baseline, represents a performance "ceiling" for rigid, statistical-based algorithms. Such algorithms, while effective, lack the capacity to learn the complex, non-linear relationships and high-dimensional correlations that define these subtle generative artifacts.

This leads to the central premise of the work detailed in this report: to break this 53% performance ceiling, a deep neural network is required. The hypothesis is that a deep learning model, with its inherent ability to learn complex hierarchical patterns, can more effectively model and detect these subtle wavelet-domain and noise-domain artifacts than any manually engineered statistical algorithm.

This paper presents the culmination of this feature-centric hypothesis: the design, implementation, and rigorous evaluation of six novel, custom-built neural network architectures. These architectures are not generic, end-to-end spatial-domain models.

They are specialized; purpose-built systems specifically designed to ingest and process high-dimensional, frequency-domain features—primarily the 12-channel coefficients of the Dual-Tree Complex Wavelet Transform (DTCWT). The objective is to determine which architectural philosophy is best suited for exploiting this complex feature space. The contributions of this paper are as follows:

- A comprehensive review of the current (2023-2025) state-of-the-art in inpainting forgery detection, identifying key trends in dataset creation and detection model paradigms.
- A detailed methodological breakdown of six novel, custom neural network architectures, each representing a different approach to processing wavelet-domain features (e.g., dense aggregation, recurrent modeling, unsupervised generative detection, and channel-specific processing).
- A quantitative experimental evaluation of these six architectures on a custom inpainting dataset, using Intersection over Union (IoU) as the primary performance metric.
- A deep analysis of the results, leading to the identification of the optimal architecture (StackDeepAll) and providing a data-driven explanation for its success, while also diagnosing the failures of alternative designs.
- A concluding discussion that contextualizes these findings within the broader SOTA landscape, validating the power of a feature-centric design philosophy as a potent alternative to purely end-to-end approaches.

2 State-of-the-Art in Inpainting Forgery Detection

Before detailing the proposed architectures, it is essential to contextualize this work within the current (2023-2025) research landscape. The field is evolving rapidly, with two dominant trends: a “dataset arms race” to address the realism gap, and a “paradigm shift” in detection models to fight generative AI with its own tools.

2.1 Baseline Methods

The application of deep learning to image forgery detection was popularized by several foundational models. Architectures such as ManTra-Net [6] and IID-Net [7] were seminal in demonstrating that deep convolutional neural networks could learn to identify a wide range of manipulation artifacts, moving beyond handcrafted features. These models serve as important baselines and established the core task as a pixel-level segmentation problem.

2.2 The Dataset Realism Gap

A significant theme dominating recent literature is the explicit recognition that existing forgery detection datasets are no longer sufficient. It is now widely accepted that “current datasets... are limited in scale and diversity” [2, 3] and that “none of the existing datasets have sufficient size, realism and pixel-level labeling at the same time”

[5]. This deficit is a critical bottleneck for research, as models trained on older, less realistic datasets fail to generalize to modern, generative inpainting.
This problem is exacerbated by the infeasibility of manual dataset creation; as one research group noted, it "can take hours for an image editing expert to manipulate just one image" [5].
In response, the SOTA has seen the introduction of several large-scale, automated, and highly realistic benchmarks:

- DiQuID: A massive dataset comprising over 95,000 inpainted images. Its methodology is particularly novel, featuring:
  - Semantically Aligned Object Replacement (SAOR): Uses instance segmentation to identify objects and generates contextually appropriate text-to-image prompts for their replacement.
  - Multiple Model Image Inpainting (MMII): Employs various state-of-the-art inpainting pipelines, primarily based on diffusion models, to create diverse manipulations [2, 3].
- COCOInpaint: Another large-scale benchmark containing 258,266 inpainted images. It is explicitly designed to test generalization by using six different SOTA inpainting models and four distinct mask generation strategies, including optional text guidance [7].
- CIMD: A benchmark created specifically for "Challenging" Image Manipulation Detection. Its subsets focus on two notoriously difficult scenarios: the detection of "small tampered regions" and forgeries subjected to double compression with identical quality factors [8].
- DIS25k: A dataset generated using SOTA image composition models, rather than just inpainting. The goal is to produce spliced images that are "close to the quality of real-life manipulations" and are demonstrably "more difficult to detect" [5].

The emergence of these datasets signifies a major shift in the field. Performance is no longer measured against simple, academic examples. Any new detection method proposed today must demonstrate its efficacy against these large-scale, diverse, and diffusion-based benchmarks to be considered a viable SOTA contender. The most critical SOTA development (2023-2025) is the pivot in detection methodology. As inpainting models have achieved "unprecedented realism", traditional detection models that search for "visible signs" or simple statistical artifacts are failing [4].
The new philosophy is that only a model with a deep, generative understanding of how an image is formed can reliably detect a generative forgery. This has led to the development of detection methods that are themselves based on diffusion models.

- InpDiffusion: This method "propose[s] a new paradigm that treats IIL (Image Inpainting Localization) as a conditional mask generation task". Instead of a discriminative segmentation, it uses a conditional denoising process, enhanced by image semantic and edge conditions, to "progressively refine predictions". This approach is explicitly designed to tackle two of the hardest challenges: model "overconfidence" and the detection of "subtle tampering boundaries" [4].

- End4 / UpDM: This work is cited as the first to "update the diffusion reconstruction model for the forgery detection task". It trains a model with a "noise-prediction objective" and leverages "one-step denoising." This technique is reported to "better align the latent spaces of reconstruction and detection" than previous reconstruction-based methods (like VAEs or autoencoders). The authors claim this method "significantly outperform[s]" other approaches, especially on diffusion-based inpainting forgeries [9].

Alongside the diffusion-based paradigm, other advanced deep learning techniques are being explored:

- FOCAL: Employs "contrastive learning and unsupervised clustering" to help the model learn a better feature space for differentiating forged and authentic regions [10].
- TruFor & MMFusion: MMFusion is an extension of the TruFor model. It acts as a "fusion architecture" that explicitly combines semantic (high-level) artifacts with low-level artifacts, using "more filter convolutions" to enhance the signal from both streams [11].
- SegFormer: A Transformer-based model noted for its simple and-efficient design for semantic segmentation, which has been identified as a strong and relevant architecture for the pixellevel localization task required by forgery detection [12].

3 Proposed Neural Network Architectures for Wavelet-Domain Detection

This section presents the core technical contribution of this work: an exhaustive analysis of the six custom neural network architectures. These architectures were not developed in a vacuum; they are the direct result of the feature-centric hypothesis established in Section 1.2. Each architecture represents a different strategy for processing high-dimensional, wavelet-domain features to detect inpainting.

3.1 Foundational Input: Dual-Tree Complex Wavelet Coefficients (DTCWT)

With the exception of the Variational Autoencoder (which also uses a related feature set), the proposed architectures are all designed to accept a specific, engineered input. This input is the 12-channel output of a first-level Dual-Tree Complex Wavelet Transform (DTCWT) decomposition [1].

This 12-channel tensor is composed of:

- 6 channels of real coefficients
- 6 channels of imaginary coefficients

Together, these coefficients represent the image's high-frequency information, separated into six distinct directional orientations ($\pm 15^\circ$, $\pm 45^\circ$, and $\pm 75^\circ$). This input is hypothesized to be an ideal forensic signal, as it captures the "localized directional

information" and "fine-grained anomalies" in texture and structure that are the hallmarks of inpainting artifacts [1]. The challenge, therefore, is not finding the signal, but processing this high-dimensional (12-channel) and sparse feature representation effectively.

3.2 Architecture 1: Custom UNet++ (MyUnetPlusPlus)

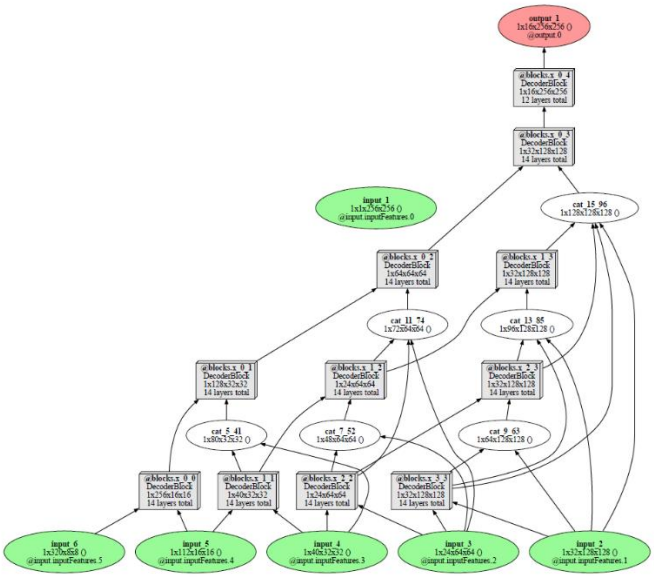


Figure 1MyUnetPlusPlusDecoder

This architecture serves as the baseline, adapting the well-established UNet++ framework to the novel 12-channel wavelet input.

- Methodology:
  - Encoder: The encoder is flexible, designed to use a pre-trained backbone from the timm library (e.g., ResNet, EfficientNet, DenseNet). This backbone is adapted to accept the 12-channel DTCWT coefficient tensor as its input [1].
  - Decoder: Two decoder variants are proposed to process the features from the encoder:
    1. MyUnetPlusPlusDecoder (Figure 1): This decoder implements the "classical dense skip connections" characteristic of the Unet++ framework. This design philosophy is built on feature re-use and aggregation, allowing the network to "preserve fine details while incorporating high-level context" from different scales of the encoder [1].
    2. MyMAnetDecoder (Figure 2): This is an enhanced decoder variant that integrates "additional attention mechanisms". Specifically, it

incorporates a "Multi-Scale Fusion Attention Block (MFAB)" designed to learn to "emphasize important features while suppressing irrelevant ones" [1].

- Strengths: The primary strength of this design is its explicit intent to "exploit the hierarchical structure of DTCWT coefficients". The dense skip connections are, in theory, perfectly suited for a forensic task, as they ensure that fine-grained, low-level artifact information is not lost during the deep propagation through the network and is re-combined with high-level semantic context [1].
- Limitations: This architecture suffers from significant, known drawbacks. The "dense skip connections require storing intermediate feature maps" at all nested levels, leading to a high computational and memory footprint, which increases training time. More fundamentally, its performance is entirely contingent on the assumption "that the wavelet decomposition captures meaningful features" [1]. If the artifacts are not well-represented in the DTCWT domain, the network will fail.

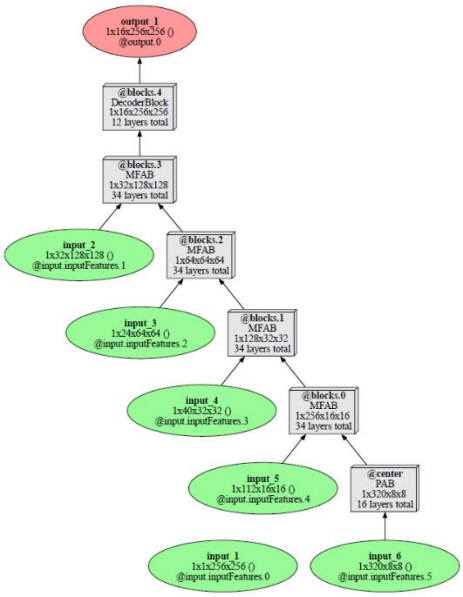


Figure 2 MyMAnetDecoder

3.3 Architecture 2: UNet++ with Global Context (MyUnetPlusPlusWithGlobalAverage)

This architecture is a direct modification of the first, designed to test a specific hypothesis: will adding explicit global context improve detection accuracy? [1].

- Methodology:
  - This model builds directly upon the MyUnetPlusPlus architecture but introduces a Global Average Pooling (GAP) module at the end of the decoder stream [1].

  - As detailed in Figure 3, this GAP module "computes the average response across spatial dimensions" of the final feature map, collapsing it into a single global feature vector that represents the "bird's-eye view" of the image's features [1].
  - This global vector is then "expanded to match the spatial resolution" of the decoder's output and is concatenated with the high-resolution feature map just before the final segmentation head [1].
- Strengths: This design is theoretically superior for specific types of forgeries. It is intended for scenarios "where anomalies are distributed across large portions of the image" or when "contextual information is critical" for classifying an artifact. The GAP mechanism is designed to provide this global context and "avoid overfitting to localized noise or small-scale anomalies" [1].
- Limitations: This architecture suffers from a critical, and potentially fatal, design flaw. While it adds global context, the very act of averaging all spatial features together risks "diminish[ing] effectiveness". Inpainting artifacts are, by nature, often "small or isolated." The GAP operation may simply "average out" these subtle forensic signals, effectively blinding the network to the very evidence it is designed to find [1].

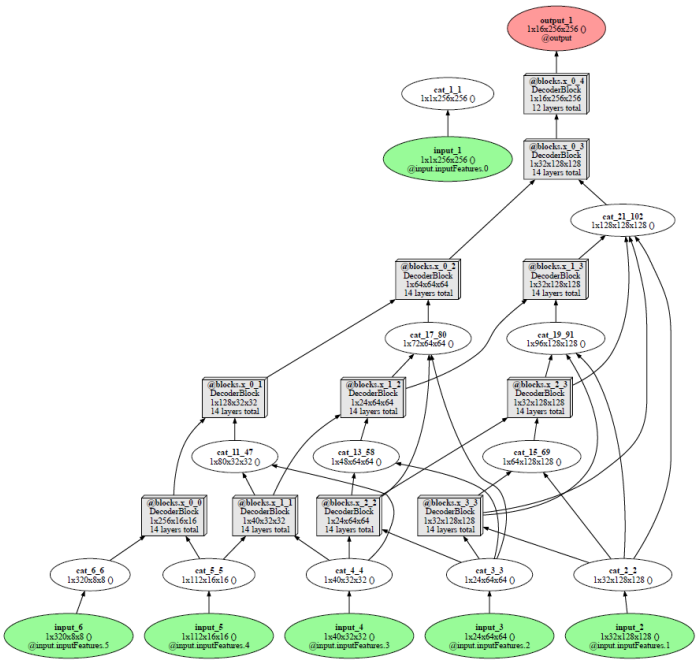


Figure 3 MyUnetPlusPlusDecoder with GlobalAverage

3.4 Architecture 3 & 4: Recurrent Architectures

This pair of architectures explores a different philosophy: modeling the spatial relationships between features using recurrent layers [1].

- Methodology:
  - Encoder: An EfficientNet backbone is used for its "parameter efficiency" and strong feature extraction capabilities [1], as once see in Figure 4.
  - Bottleneck: The key innovation resides in the bottleneck. Instead of a standard convolutional block, a ConvLSTM (Convolutional Long Short-Term Memory) module is inserted. A ConvLSTM "models spatial dependencies over feature maps," allowing it to learn contextual relationships across different parts of the image, much like an RNN learns dependencies over time [1].
  - Decoder: A standard Fully Convolutional Network (FCN) decoder with upsampling blocks and skip connections is used to reconstruct the segmentation map [1].
- Methodology (ImprovedEfficientNetFCNConvLSTM, Figure 5):
  - This model enhances the base ConvLSTM architecture with two key additions:
    1. Channel Attention: A channel attention mechanism is added. This module dynamically "adjusts the importance of features" by learning weights for each channel. This is "particularly effective for DTCWT coefficients," as the network could, for example, learn that anomalies are most prominent in the $\pm 15^{\circ}$ bands and thus "prioritize" those features [1].
    2. Optional Wavelet Preprocessing: An additional decomposition into low-pass and high-pass components can be applied before the encoder, providing another layer of frequency-domain analysis [1].
- Strengths: The ConvLSTM module is "crucial for detecting artifacts that may exhibit global or contextual dependencies". The Improved version's attention mechanism adds a valuable layer of intelligent, data-driven feature prioritization, making it more robust [1].
- Limitations: Recurrent layers are notoriously resource-intensive. The ConvLSTM module introduces significant "computational and memory overhead" because it must maintain hidden and cell states for every spatial region in the feature map. Furthermore, while ConvLSTM models general spatial dependencies, it is "less effective at explicitly modeling anisotropic patterns" (i.t., patterns with a strong directional bias), which is a potential weakness given the highly directional nature of the DTCWT input [1].

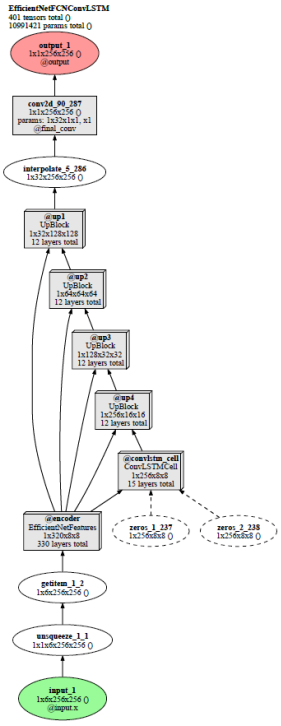


Figure 4 EfficientNet FCN Conv LSTM

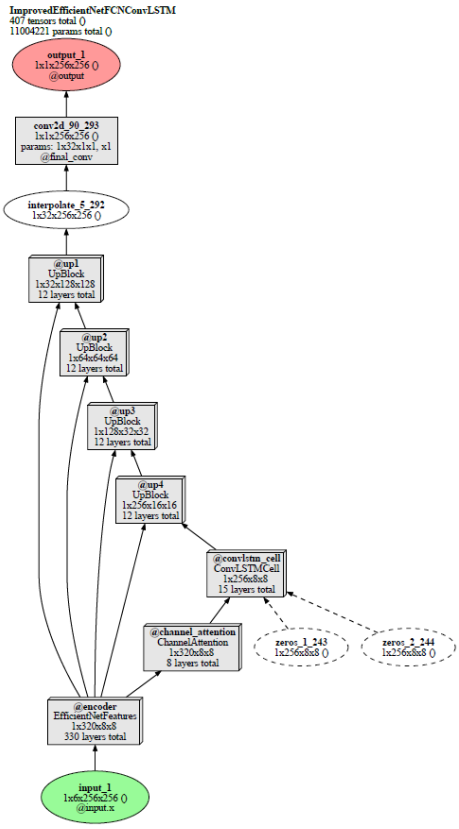


Figure 5 ImprovedEfficientNet FCN Conv LSTM

3.5 Architecture 5: Unsupervised Generative Detection (Variational Autoencoder Architecture)

This architecture represents a fundamental shift in approach, moving from supervised segmentation to unsupervised, reconstruction-based anomaly detection [1].

- Methodology:
  - Feature Extractor: The feature extraction is configurable, supporting either DTCWT or the Shearlet Transform [1]. Shearlets are noted for providing "anisotropic analysis," which is theoretically even more effective for "detecting directional inconsistencies" in inpainted regions [1].
  - Detection Model: A Flexible Variational Autoencoder (VAE) is used as the core model, detailed in Figure 6 (left and right).
  - Detection Principle: The VAE is trained only on authentic (non-forged) features. It learns to compress and then reconstruct the probabilistic distribution of "normal" wavelet or shearlet coefficients. When a forged image's coefficients are passed through, they "deviate from the learned distribution" of normal data. The VAE will fail to reconstruct them accurately, resulting in a high reconstruction error (e.g., Mean Squared Error). This map of reconstruction errors is the final anomaly segmentation mask [1].
- Strengths: The primary strength is that this method is unsupervised. This makes it "particularly useful... where labeled anomaly data is scarce". The use of shift-invariant feature extractors (DTCWT, Shearlets) provides robustness against translation [1].
- Limitations: This approach has two major weaknesses. First, the "computational cost of DTCWT and Shearlet Transforms is significant," creating a preprocessing bottleneck. Second, it is highly susceptible to false negatives. This occurs "when inpainting artifacts closely resemble the surrounding textures". If the forgery is "too good," the VAE's latent space may be general enough to simply learn to reconstruct the forgery as if it were authentic, resulting in a low reconstruction error and a missed detection [1].

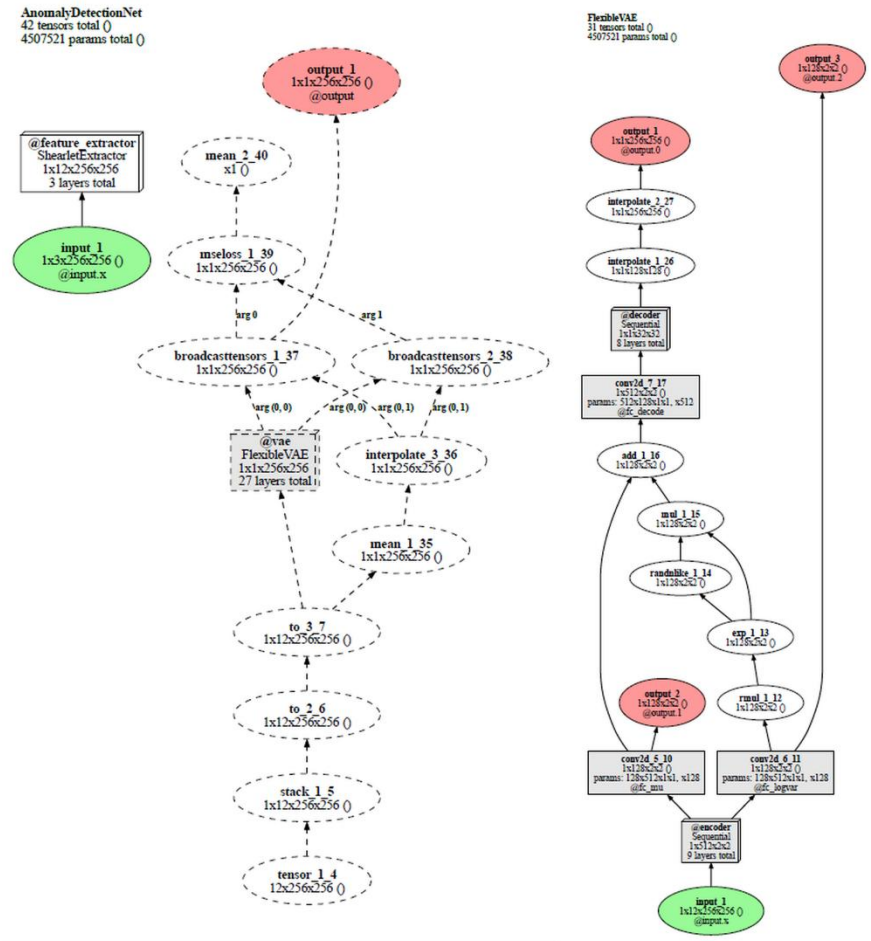


Figure 6 Custom architecture based on Variational Autoencoder and Custom Variation Autoencoder module

3.6 Architecture 6: The Optimal Model (StackDeepAll: A Stacked Unet architecture)

This final architecture is built on a "channel-specific processing" philosophy [1].

- Methodology:
  - Core Design Philosophy: This architecture is a direct response to the limitations of models like MyUnetPlusPlus. Instead of feeding all 12 DTCWT channels into one large encoder (which might "blend" or "average" their unique signals), this model processes each of the 12 channels independently and in parallel.
  - Architecture: The model is composed of 12 separate, parallel UNet-like encoders [1]. Each encoder becomes a specialist for the artifacts present in its single corresponding channel (e.g., one encoder sees only the +15◦ real channel, another sees only the −45◦ imaginary channel, etc.). Each subnet is tasked with enhancing the forged area signal from its channel [1].

  - Fusion: The high-level feature outputs from all 12 parallel encoders are then "fused using a shared decoder" (referred to in the abstract as a "supplementary sub-network" [1]) to generate the final, integrated segmentation map [1].
- Strengths: This "channel-specific processing" is the key to its success. It "ensures that unique features inherent to each coefficient are retained" and prevents "premature feature blending" [1]. This allows the model to learn 12 distinct, highly discriminative feature representations from each domain before fusion, leading to a much richer and more accurate final prediction [1].
- Limitations: The trade-off for this performance is computational cost. The architecture is "significantly high" in cost due to the necessity of processing "12 independent networks" [1]. It also "lacks early-stage cross-channel interaction," meaning it cannot learn correlations between channels (e.g., how the real part of a band relates to its imaginary part). Finally, "scalability... poses a challenge," as adding more feature channels (e.g., from a multi-level wavelet decomposition) would require linearly adding more parallel networks [1].

4 Experimental Evaluation and Results

This section presents the quantitative validation of the six architectures detailed in Section 3. The experiment was designed to directly compare the efficacy of these competing architectural philosophies.

4.1 Experimental Setup

- Dataset: A subset of the author's custom-developed inpainting dataset was used. This dataset comprised 1,000 images for training and 1,000 images for validation.
- Training Protocol: All models were trained for 10 epochs under identical conditions to ensure a fair comparison.
- Input Features:
  - The five supervised architectures (MyUnetPlusPlus, MyUnetPlusPlusWithGA, both ConvLSTM models, and StackDeepAll) used the 12-channel first-level DTCWT coefficients as input.
  - The unsupervised Variational Autoencoder architecture used scattering coefficients as its input, a related but distinct frequency-domain feature set.
- Metric: The primary evaluation metric for this segmentation task was Intersection over Union (IoU), which measures the overlap between the predicted forgery mask and the ground-truth mask.

4.2 Quantitative Results

The comparative performance of all six architectures is presented in Table 1.

Table 1: Comparative Analysis of Proposed Architectures (Training and Validation IoU)

Architecture	Training IoU	Validation IoU
Unet like architecture (MyUnetPlusPlus)	0.3557	0.2806
Unet with Global averaging (MyUnetPlusPlusWithGA)	0.3163	0.2523
EfficientNet FCN with LSTM	0.6631	0.4507
ImprovedEfficientNet FCN with LSTM	0.7201	0.4514
Variational Autoencoder architecture	0.3713	0.3711
Stack Unet architecture (StackDeepAll)	0.9101	0.8126

4.3 Analysis of Results

The quantitative data from Table 1 provides a clear and decisive verdict on the relative success of the six design philosophies.

- StackDeepAll (The Clear Victor): The Stack Unet architecture (StackDeepAll) is the unambiguous winner, achieving a Validation IoU of 0.8126 [1]. This result is not only the highest by a massive margin but also achieves the author's stated performance goal of > 0.8 IoU. This strongly validates the "channel-specific processing" hypothesis. While it shows signs of overfitting (0.9101 training IoU), its generalization performance is still exceptionally high.
- Variational Autoencoder (The Most Stable): The VAE model is remarkable for a different reason. Its Training IoU of 0.3713 and Validation IoU of 0.3711 are virtually identical [1]. This indicates zero effective overfitting. This stability is a direct result of its unsupervised, generative nature; it is not "memorizing" forgery masks but rather learning the fundamental distribution of authentic scattering coefficients. While its absolute performance (0.37 IoU) is mediocre, its stability suggests it is a highly generalizable and robust model, validating the unsupervised approach.
- ConvLSTM Models (High Overfitting): The ImprovedEfficientNet FCN with LSTM achieved the second-best validation performance (0.4514 IoU). However, both LSTM models exhibit severe overfitting, with the improved model dropping from 0.7201 to 0.4514. This suggests that the recurrent layers are "memorizing" spatial-temporal patterns present in the 1,000 training images, but these complex learned patterns do not generalize well to the unseen validation set.
- UNet++ Models (The Clear Failures): The baseline Unet like architecture (MyUnetPlusPlus) performed poorly, achieving only a 0.2806 Validation IoU. This indicates that simply feeding all 12 channels into a standard U-Net design is an ineffective strategy.
  - Failure of Global Averaging: Most strikingly, the Unet with Global averaging model performed even worse, achieving only a 0.2523 Validation IoU [1]. This

quantitatively proves the hypothesis from Section 3.3: the Global Average Pooling (GAP) module is detrimental. By averaging all spatial features, the module "averages out" the subtle, localized wavelet artifacts, effectively destroying the forensic signal and impairing the network's performance.

5 Discussion

The experimental results provide a clear foundation for a deeper discussion of the implications of this work, both for the project itself and for the broader field of forensic detection.

This work's primary contribution is the design, implementation, and empirical validation of six novel neural network architectures, each custom-built for the task of inpainting detection. This involved a thorough "analysis of how different architectures process anomalies" in diverse datasets and "explored an optimal balance between processing efficiency and anomaly detection accuracy". Specific contributions include:

- The pioneering use of a Variational Autoencoder combined with wavelet scattering coefficients for unsupervised, generative anomaly detection, which demonstrated exceptional training stability.
- The validation of the StackDeepAll architecture, proving the efficacy of a "channel-specific" processing pipeline for high-dimensional wavelet features.
- The empirical disproof of a global-context-first approach, demonstrating that Global Average Pooling is detrimental to this specific task.

The results from Table 1 are not just a leaderboard; they tell a clear story about how to process high-dimensional, engineered features.

The 12 DTCWT channels are not 12 arbitrary images; they are 12 highly specific, highfrequency, directional data streams. The MyUnetPlusPlus models, which achieved a ~ 0.28 IoU, fail because their first layer immediately convolves all 12 channels, "muddying the waters" and irretrievably blending the unique signals. The network is forced to find a compromise-feature that is "sort of good" for all 12 channels, but "expert" at none.

The StackDeepAll model, which achieved a 0.81 IoU, succeeds precisely because it does the opposite. It strictly enforces "channel-specific processing." It treats each channel as a separate problem, allowing 12 parallel UNet-like encoders to become experts at finding anomalies only in that specific domain (e.g., an expert for "horizontal-imaginary artifacts," an expert for "diagonalreal artifacts," etc.). Only after these 12 expert-level feature streams are extracted are they fused by the final "supplementary sub-network" [1].

This proves that for this type of feature-centric detection, preventing premature feature-blending is the single most important architectural design choice. The high computational cost is the tradeoff for this specialization, which ultimately yields a vastly superior result.

This paper's work is critically important because it proves the immense, ongoing value of the feature-centric philosophy. While much of the research community chases larger end-to-end models, the StackDeepAll architecture's > 0.8 IoU performance is a powerful demonstration that expert domain knowledge remains a key component of

SOTA performance. It proves that knowing where the artifacts lie (in the complex wavelet domain) and designing an architecture to target that domain can achieve results that are competitive with, or even superior to, purely data-driven, end-to-end spatial methods.

6 Conclusion

This paper has presented a comprehensive investigation into novel neural network architectures for image inpainting detection. We have detailed the methodology, strengths, and limitations of six custom-designed architectures, all predicated on the central hypothesis that inpainting artifacts, while visually subtle, are most evident in the complex wavelet domain.

A rigorous experimental evaluation provided a clear victor: the StackDeepAll architecture. This model, which employs a unique channel-specific processing pipeline with 12 parallel UNetlike encoders, achieved an exceptional Validation IoU of 0.8126. This result empirically validates the hypothesis that for high-dimensional, engineered features like DTCWT coefficients, preventing premature feature-blending and allowing specialized encoders to process each channel independently is a superior design philosophy. Conversely, the experiments also provided a clear disproof of a global-context-first approach, where a Global Average Pooling module was shown to be detrimental to performance, likely by "averaging out" the very localized artifacts being targeted.

Furthermore, we have presented a comprehensive review of the 2023-2025 state-of-the-art, which is currently pivoting towards two new frontiers: the creation of massive, realistic, diffusion-based datasets (e.g., DiQuID, COCOInpaint) [2, 7] and the use of generative models, such as InpDiffusion [4], for detection.

The success of the StackDeepAll model provides a powerful counter-narrative to a purely end-to-end-driven field. It proves that a deep, feature-centric design philosophy, rooted in a forensic understanding of where artifacts manifest, remains a highly effective and robust strategy for combating the "accelerated progress" [1] of generative image manipulation.

References

[1] Barglazan, A. A. (2025). Image inpainting forgery detection (PhD Thesis). "Lucian Blaga" University of Sibiu.
[2] Giakoumoglou, P., et al. (2025). DiQuID: A high-quality inpainting dataset. arXiv:2502.06593v1.
[3] Yan, H., et al. (2025). COCO-Inpaint: A Benchmark for Image Inpainting Detection and Manipulation Localization. arXiv:2504.18361v1.
[4] InpDiffusion. (2025). A new paradigm that treats IIL as a conditional mask generation task utilizing diffusion models. AAAI Conference on Artificial Intelligence.
[5] Eren, E., et al. (2024). DIS25k: A dataset generated using SOTA image composition models. arXiv:2404.02897v1.
[6] Wu, Y., Abdalmageed, W., & Natarajan, P. (2019). ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[7] Wu, H., & Zhou, J. (2021). IID-Net: Image inpainting detection network via neural architecture search and attention. IEEE Transactions on Circuits and Systems for Video Technology, 32(3), 1172–1185.
[8] Z, Zhenfei. (2024). CIMD: A Challenging Image Manipulation Detection benchmark dataset. AAAI Conference on Artificial Intelligence.
[9] UpDM. (2025). Updating the diffusion reconstruction model for the forgery detection task. arXiv:2509.13214v1. [9]
[10] FOCAL. (2023). A method using contrastive learning and unsupervised clustering to differentiate between forged and authentic regions.
[11] MMFusion. (2023). An extension of the TruFor model, acting as a fusion architecture for semantic and low-level artifacts.
[12] Xie, E., et al. (2021). SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. Neural Information Processing Systems (NeurIPS).

DOI: 10.2478/ijasitels-2025-0016

Office Ergonomics and Ambiental Environment Monitoring System

MANTA Aida-Mihaela[1]
[1]*Faculty of Engineering, "Lucian Blaga" University of Sibiu, Romania*
aida.manta@ulbsibiu.ro

Abstract
More and more people suffer from back health issues and other afflictions due to their position at the desk, their working ambiental environment and growing sedentarism. The proposed IoT system aims to help people and companies with such concerns and create a healthier environment to work in. The proposed system is placed on a person's desk. The system uses a camera and a computer vision model to analyse a person's position at the desk. It also uses a set of sensors to assess the ambiental environment's quality in accordance with ESH (Environmental, Safety, and Health) recommendations. The information analysed can be accessed by the end-user through a visual report that emphasizes the deviations from the recommended standards. The possibility to faster detect non-conformities in working environments and bad habits in ergonomics can improve the quality of the working space and delay or prevent health issues in the long term.

Keywords: office ergonomics, office ambiental environment, office microclimate, IoT system

1 Introduction

In recent decades, with the rapid development of the technological world and the continuous digitization of daily activities, humans have become increasingly dependent on technology, both in professional environments and personal life. Most of the time, this technology involves viewing a screen, whether it's a phone, tablet, or computer. An article from January 2023 reveals that, globally, a person spends an average of approximately 6 hours and 58 minutes per day in front of a screen [1]. This average can be even higher for people working a desk job. These prolonged periods of time spent in front of a screen combined with unhealthy posture habits, sedentarism and a non-standard indoor microclimate can lead to different health issues. They can be observed in the near future or in the long-term and are generally related to eyesight, back muscles and back bones problems, headaches, excessive tiredness and even depression [2] [3] [4] [5] [6].
The purpose of this paper is to propose a potential solution to these everyday challenges, enabling employees and employers to continuously monitor the work environment, promote a healthier workspace, and reduce the risk of health issues.
This paper is structured as follows: the second section presents the system prototype and the technologies used, being split in individual subsections for hardware and

software components, and the last section presents the conclusions and future developments of this project.

2 Methodology

This section presents the monitoring system's development in terms of hardware and software used in the project such as TensorFlow library, Node-RED, MySQL database and Microsoft PowerBI. The scope of this section is to present the purpose of the technologies used and how they interact with each other to deliver the end result to the user in a friendly and easy to interpret manner.
The proposed IoT system is placed on the person's desk. An overview of the system is presented in Figure 1.

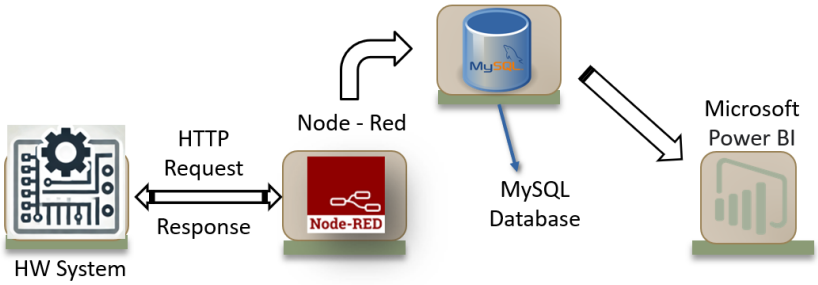


Figure 1. Implementation of main components of the system

2.1 Hardware components

In this subsection the hardware part of the system will be presented. It emphasizes the two purposes of the hardware: the collection of office microclimate data and the picture taken of the person working at the desk.
The data is collected and sent periodically, at the request of the main application, during the usual working hours of a person.

2.1.1 Office microclimate data

As presented in the introduction, the microclimate in which a person is working can have a direct impact on his or her health. Therefore, certain standards regarding the microclimate exist that, if respected, create the healthiest office environment to work in.
For this reason, an important part of the proposed system consists of a set of sensors that collects ambiental environment data (temperature, humidity and light intensity) and sends it to the server (the Node-RED application) to be analysed against the standards established by specialists and stored in the database.

2.1.2 Ergonomics

Another important factor to be monitored is the posture of a person at the desk. As previously presented, a bad posture for prolonged periods of time and unhealthy habits regarding sedentarism can lead to issues at back muscles and back bones level.

The purpose of this proposed solution is to verify the upper-body posture of a person, considering that this is the one visible above a desk
A camera is used to take a picture upon request. The picture taken is then sent to the main application to be analysed.

2.2 Software components

The three main software components of this system are the Node-RED application, the MySQL database and the PowerBI report.

2.2.1 Main application

The application behind this IoT monitoring system is implemented in the Node-RED platform. The main objectives of the application are:

- periodically request data from the hardware;
- analyze the upper-body posture using a model from the Tensorflow library to get important body key points and make relevant observations;
- analyze the microclimate data and calculate the deviation from the standards, if applicable;
- store the collected data in the MySQL database;
- if deviations were observed (regarding either the microclimate or the posture of the person), a request is sent back to the hardware system to display a relevant message for the user on a small display.

For GDPR reasons, the picture taken of the user at the desk is discarded immediately after it is processed. The only information saved is the observation regarding the user's posture, information derived from the taken picture and above-mentioned analysis.
If the person is not at the desk when the data is collected, this scenario is verified and treated accordingly.
The application also takes into account the current season (warm season or cold season), as the microclimate standards have small but important differences depending on the period of the year.

2.2.2 Database

A MySQL database was chosen for this application for its easy-to-use properties and the already implemented connection algorithms in the Node-RED platform.
The database stores the timestamp of data collection, microclimate information, deviations from standards and relevant observations regarding ergonomics.

2.2.3 Microsoft PowerBI

PowerBI is a tool provided by Microsoft that offers a simplified and more visual and interactive method of creating reports. It can connect directly to a database and extract the desired data for the reports, which is helpful for scenarios like this one when new data is constantly added to the database.

The report created for this project is simple in order to provide a friendly and easy to interpret analysis for the end-user. It consists of a filter for the date, a table which showcases the microclimate data collected through that date and relevant observations about the ergonomics, and a visual report which presents the deviation from the standard for each metric.

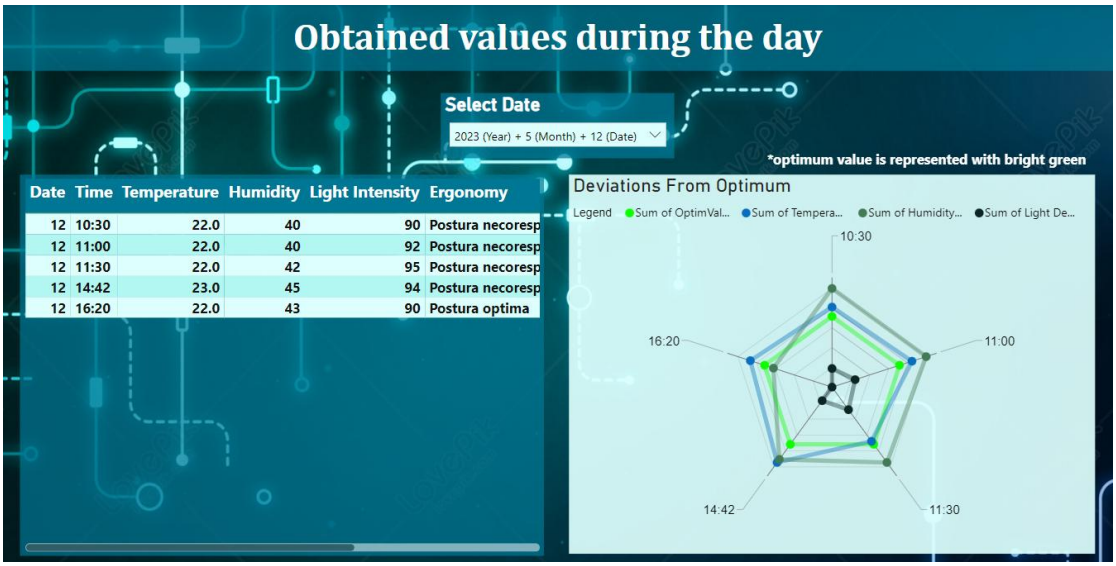


Figure 2. PowerBI visual report

3 Conclusions

In today's society, working at a desk and using a computer have become the new norm, whether it is the case of a student or an employee in a company. This new reality has come with its own challenges when it comes to health issues. More and more people encounter health problems that are often related to their working environment and unhealthy unconscious habits.

The purpose of this IoT prototype is to spread awareness about these important factors that are usually neglected in the working day-to-day life. The purpose is also to address and provide a potential solution to this contemporary problem in a manner that is as simple, cost-efficient, and space-saving as possible.

This system was tested in an environment with one person working at a desk. The results obtained from using this prototype have drawn attention to irregularities in the working space microclimate and generally bad posture, aspects that were rarely observed before.

The future developments considered for this project are: finding solutions to make the system even simpler and more compact while still being cost-efficient, monitoring more microclimate-relevant metrics, scaling the system to have multiple users simultaneously, creating a mobile application for the user interface.

Acknowledgment

I would like to thank Professor Adrian Florea ("Lucian Blaga" University of Sibiu) for his invaluable guidance throughout the development of this project. His constant mentorship,

expertise and insights have been a crucial part of getting this project from just an idea to a working first prototype.

References

[1] J. Howarth, „Alarming Average Screen Time Statistics (2023),” Exploding Topics, [Interactiv]. Available: https://explodingtopics.com/blog/screen-time-stats. [Accesed 07 Feb 2023].
[2] P. J. K., R. M., B. Y., E. J. M. și L. A., „Computer-related visual symptoms in office workers,” Ophthalmic and Physiological Optics, vol. 32, nr. 5, pp. 375 - 382, 2012.
[3] F. F. A., I. Rasdi, E. Z.A., S. N. S.I. și S. D., „Workplace Sedentary Behaviour and Work-related Quality of Life Among Office Workers,” Malaysian Journal of Medicine and Health Sciences, vol. 18, pp. 53 - 61, 2022.
[4] M. Malińska, J. Bugajska și P. Bartuzi, „Occupational and non-occupational risk factors for neck and lower back pain among computer workers: a cross-sectional study,” International Journal of Occupational Safety and Ergonomics, vol. 27, nr. 4, pp. 1108 - 1115, 2021.
[5] H. İkiz și E. Ergin, „Musculoskeletal system problems in office workers: relationship of physical activity levels and quality of life,” International Journal of Occupational Safety and Ergonomics, vol. 29, nr. 1, pp. 321 - 328, 2023.
[6] P. Adrian, „Microclimatul optim pentru locurile de muncă la birou,” euramis, 31 Iulie 2018. [Interactiv]. Available: https://euramis.ro/microclimatul-optim-pentru-locurile-de-munca-la-birou/. [Accesed 04 06 2023].
[7] Tensorflow, „Get started with TensorFlow.js,” Tensorflow, 09 February 2023. [Interactiv]. Available: https://www.tensorflow.org/js/tutorials. [Accesed 11 June 2023].
[8] D. Oved, I. Alvarado și A. Gallo, „Real-time Human Pose Estimation in the Browser with TensorFlow.js,” Tensorflow, 07 May 2018. [Interactiv]. Available: https://blog.tensorflow.org/2018/05/real-time-human-pose-estimation-in.html. [Accesed 11 June 2023].

DOI: 10.2478/ijasitels-2025-0018

Wireless Controller for Modbus RS485

Mihai-Liviu Stroila[1],

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

Abstract

The purpose of this paper is to propose a modern and efficient solution for controlling and analysing RS-485 communication networks, overcoming the limitations of traditional methods. In high-voltage systems, isolating RS-485 nodes is essential to protect both circuits and human operators from high voltages and transient phenomena. Conventional solutions, such as high-speed optocouplers and transformers, are effective but require numerous external components, resulting in bulky devices with high production costs. The proposed solution introduces a rechargeable, battery-powered control device operated wirelessly via a smartphone or laptop. This ensures complete isolation for the operator while simplifying the system's design.

Compared to traditional galvanic isolation methods, this solution significantly reduces production costs and device size, with the total size being comparable to a computer mouse. It offers a cost-effective, compact, and future-ready alternative for industrial applications requiring efficient communication and safety in high-voltage environments. Additionally, the integration of IoT technology enhances functionality by enabling remote monitoring, real-time control, and data analysis. This IoT-driven approach supports predictive maintenance and scalability, making it ideal for modern industrial and automation systems.

Keywords: RS-485, communication networks, IoT, industrial applications.

1.Proposed method

The concept of the developed application consists of creating a control module primarily used for monitoring an RS-485 serial interface. In its final form, the designed equipment should ideally be capable of operating in multiple modes:

- Sniffer mode, acting only as a passive component in the system, intercepting messages transmitted on the line without responding or influencing communication in any way.
- Controller mode, functioning as the master in the network, with the ability to take over the role of the existing master node and send requests to the network nodes to observe how they respond.

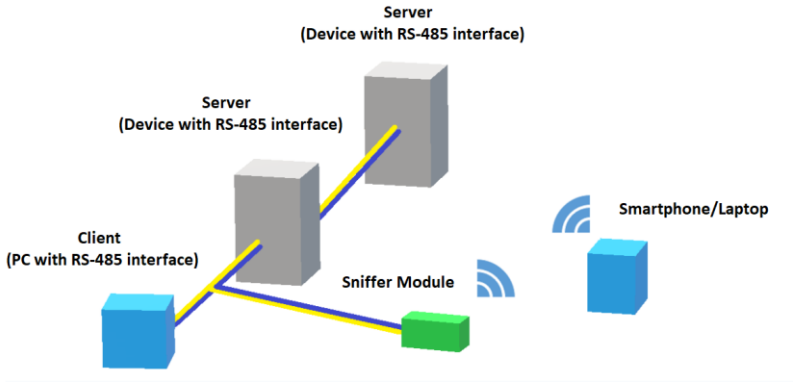


Figure 1 Working Diagra

1.1 The design of the system with a microcontroller from the ESP family.

In order to develop a complex system, the most effective method is to break it down into smaller subsystems with dedicated functions. For each individual block, the desired function is determined, and the components required to achieve this function are selected. In Figure 2 below, the connection method of the component blocks can be observed.

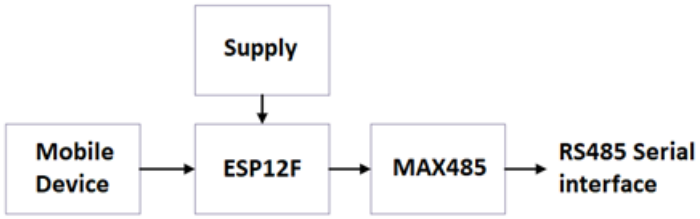


Figure 2 Block Diagram

To establish a connection with the RS-485 serial interface, a driver specifically designed for this interface is required. The MAX485 block consists of this integrated circuit along with the necessary components for its proper functioning, such as capacitors, resistors, and connectors. The data to be transmitted over the interface is received from the microcontroller via a UART-type serial communication.

The ESP12F block serves as the central unit of the entire system and includes the microcontroller of the same name, along with other passive components. This microcontroller runs all the logic required for processing the data that will be sent and received through the serial interface.

The Mobile Device block is represented by a web-based graphical interface that can be accessed through a browser on a mobile device such as a smartphone or any other device with WIFI connectivity and browser access.

The power supply block is divided into two subsystems. The microcontroller is powered by a rechargeable LiPo battery.

- The first subsystem is responsible for regulating the battery voltage to a nominal 3.3V using a linear voltage regulator.
- The second subsystem is in charge of charging the battery. This is achieved using a specialized integrated circuit designed for charging LiPo batteries. The necessary energy for charging is supplied externally via a USB-C connector, which is used exclusively for charging.

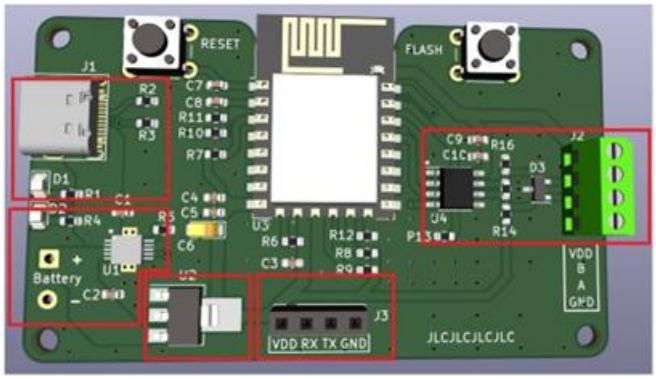


Figure 3 Layout

1.2 Software development of the application

A graphical interface is a means of communication between a user and a system that is meant to be controlled. This allows for easier use of complex systems through various objects such as icons and buttons that replace the old text- or command-based method of communication.

Interaction is carried out via a pointer that acts as a navigator within the interface. The user can enter or select the desired icon or button with a click, which will trigger a series of actions.

The graphical interface used to control the developed application is in the form of a web interface. To create this interface, we used three programming languages: HTML for the structure of the interface, CSS to describe how the elements of the HTML file are displayed and to style the interface, and JavaScript to program the interface's behavior, with its main function being to process requests sent to the backend and display the responses within the interface.

To create the graphical interface, the three corresponding files need to be accessed by the C++ functions. The chosen method was using the LittleFS library.

The files (.html, .css, .js) are stored as data in the microcontroller's flash memory, being flashed separately from the compiled code. These files are accessed using a file access system.

Using specific methods from the LittleFS library, the files are loaded into the microcontroller's RAM. From there, they are retrieved by other functions to initialize the actual web interface.

For the ModBus protocol, the necessary settings are configured using the ModBus RTU library. The choice of this library was explained at the beginning of the section.

The transmission and reception pins are set, along with the baud rate, transmission mode data, and the operating mode, which in this case is master mode.

Once a client is connected, the system waits for a request from the frontend. When a request arrives, it is analyzed and processed by the corresponding backend function, depending on its type.

The communication between front and backend is done using JSON (JavaScript Object Notation). It plays crucial role in this code by serving as the standard format for structuring and transmitting data between the ESP8266 (acting as a backend) and the web client (frontend).

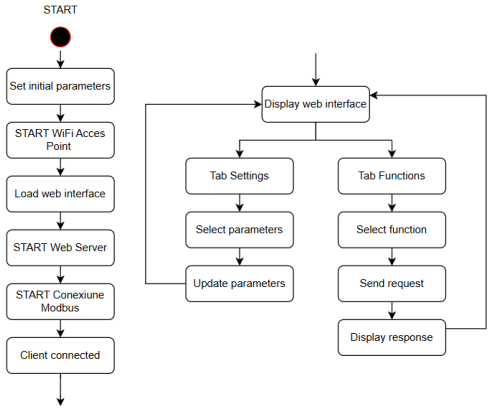


Figure 4 Software workflow

JSON provides a lightweight, human-readable way to organize data into key-value pairs and arrays. In this code, various endpoints (e.g., /settings, /readCoils, /readHolding, etc.) use a global StaticJsonDocument to assemble response data such as transaction statuses, modbus readings, and configuration values. This structure makes it easy for the frontend to parse and understand the response.

The code involves interactions with hardware components via Modbus and then relays the results back to the web interface. By packaging the hardware communication results (like sensor readings, coil states, or error codes) into a JSON object, the data is sent in a format that web clients (typically using JavaScript) can readily process and display. The ArduinoJson library is specifically designed for embedded systems with constrained resources. Its efficient handling of JSON data makes it an ideal choice for an ESP8266 environment where memory and processing power are limited. This allows the microcontroller to handle both Modbus communication and HTTP server tasks without significant overhead.

1.3 Functionality demonstration

To test the project, a system will be used consisting of a device acting as the master (represented by the project itself), an Arduino Uno functioning as a slave, and a device with a dual role—either as a slave or as a monitor—to oversee transmissions on the serial interface using the Windows application ModLink VCL Demo. In Figure 5 below, the block diagram of the system used for the practical demonstration is shown.

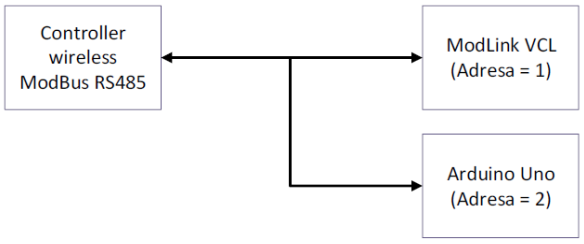


Figure 5 Block diagram of the testing system

To test compatibility with other systems, an Arduino Uno running a ModBus slave code based on the "SimpleModbusSlave.h" library was used, with a slave address set to 2. A TMP36 temperature sensor is connected to the Arduino, monitoring the ambient temperature. The read value is stored in a Holding register at address 0x00.

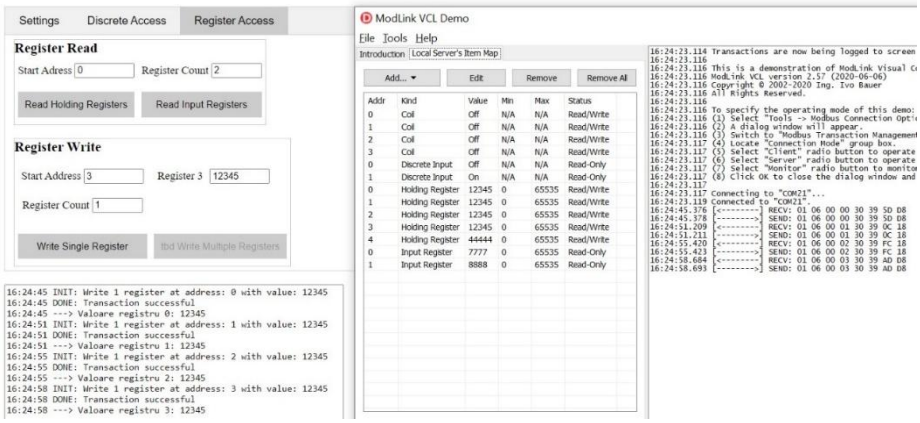


Figure 6 Testing the functionality

2 Comparison with other available solution

In examining the available solutions for establishing wireless connectivity to RS485 servers, it becomes evident that a diverse range of options exists, each designed with specific applications in mind. These alternatives vary in terms of design, performance, and intended use, spanning from simple point-to-point configurations to more sophisticated networked systems optimized for industrial automation, remote monitoring, and control. Each solution addresses the unique challenges of converting

RS485 communication to a wireless medium, ensuring robust and reliable data transfer in different operational environments. Evaluating these options allows for the selection of the most appropriate approach tailored to the particular requirements and constraints of a given project.

	Connectivity	**Power supply**	**Isolation Protection**	**Work mode**
Controller wireless for Modbus RS485	WIFI	5V USBC, 3.7V battery	Power isolation	Web browser interface
Waveshare Rail-Mount Serial Server RS485 to WIFI/Ethernet Module	ethernet, WIFI	5 - 36V DC screw terminal, POE	Power isolation, Signal isolation	Web Browser, MQTT
RS485 to Ethernet converter	ethernet, WIFI	5 - 36V DC	Not mentioned	TCP Server, TCP Client, UDP Server, UDP Client, MQTT Server, MQTT Client
Industrial RS485 IoT Gateway	ethernet, WIFI	6 - 35V DC	Signal isolation	MQTT
Modbus to Lorawan Sensornode for Industrial IoT Application	Lorawan,	USB, Solar Panel, Battery	Not mentioned	MQTT

The alternative projects mentioned above are dedicated to integrating RS485 networks into broader systems for enhanced monitoring capabilities. This approach delivers a permanent, always-connected solution that, once installed becomes a critical component of the overall system architecture. Such a solution is typically intended for continuous, long-term monitoring and control within industrial or building automation environments, where reliability and seamless integration are paramount.

In contrast, my proposed solution is specifically designed as a portable tool for debugging and calibration purposes. This solution is engineered to maximize portability while maintaining a high level of functionality and ease of use. The intent is to provide a flexible, temporary asset that can be deployed quickly for on-site troubleshooting, system verification, and calibration without the need for a fixed installation. Its design emphasizes rapid deployment, allowing technicians to address issues, fine-tune system parameters, and verify network communications directly at the point of need.

Furthermore, the control interface of this portable solution is implemented as a robust web-based platform, accessible via any device equipped with wireless connectivity. This ensures that system diagnostics, updates, and real-time monitoring can be performed from virtually anywhere, significantly enhancing operational efficiency. The web interface is designed to be intuitive and user-friendly, reducing the learning curve for new users and streamlining the maintenance process.

Additionally, the modular design of the proposed solution supports scalability and future enhancements. The system can be easily integrated with other IoT devices, enabling comprehensive monitoring and control across various network segments. This flexibility not only facilitates immediate debugging and calibration tasks but also allows for gradual system upgrades and the incorporation of new functionalities as technological needs evolve.

3 Conclusion

Due to the simplicity and reliability of the protocol—and the fact that it is open source—ModBus remains one of the preferred options in industrial automation and beyond. Almost every industrial automation device with serial communication offers a ModBus option in one of its forms (RTU, ASCII), if not as the standard.

I started with the idea of creating a complete controller, primarily used for monitoring an RS-485 serial interface on which messages are transmitted using the ModBus protocol. Out of the two approaches I considered (a sniffer for intercepting messages, or a controller in which the equipment acts as the master on the network), I managed to implement the latter by finding a solution for implementing a master device on the serial interface. The main objective was to demonstrate that the idea could be implemented, and bringing the project to a functional stage was more important than the number of features and functions the device could perform.

The final result is the successful presentation of a wireless controller for Modbus RS-485, with limited functionality, but the possibility of adding new features through software development. for a few seconds

The final result is the successful demonstration of a wireless controller for Modbus RS-485, with limited functionalities, but with the possibility of adding new functions through software development.

References

[1] Modbus Organisation, Modbus Application Protocol Specification v1.1b3, Modbus.org, 2012.

[2] MODICON, Inc., Industrial Automation Systems, Modbus Protocol - Reference Guide, 1996.

[3] modbus-esp8266, https://github.com/emelianov/modbus-esp8266.

[4] Modlink, https://www.ozm.cz/ivobauer/modlink/introduction.htm.

[5] Espressif Systems Inc, ESP8266EX Datasheet, 2020.

[6] Microchip Technology Inc., Lithium-Ion Battery Charging: Techniques and Trade-offs, 2004.

modbus-esp8266, https://github.com/emelianov/modbus-esp8266.

[7] SimpleModbus NG, https://github.com/angeloc/simplemodbusng.

[8] Rail-Mount Serial Server, RS485 to WIFI/Ethernet Module, Modbus MQTT Gateway https://www.waveshare.com/rs485-to-wifi-eth.htm

[9] Industrial RS485 IoT Gateway https://www.pcbway.com/project/shareproject/Industrial_RS485_IoT_Gateway.html

[10] RS485 to Ethernet converter https://hackaday.io/project/179981-rs485-to-ethernet-converter

[11] Modbus to Lorawan Sensornode for Industrial IoT Application

[12] https://www.hackster.io/dadanugm07/modbus-to-lorawan-sensornode-for-industrial-iot-application-b399e3#toc-hardware-diagram-2

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

VOLUME 15 Number 1 2025

CONTENTS

Ground fault location analysis on transmission lines using MATLAB/Simulink environment 3
Ioan-Dorel Păştina, Claudiu Şolea, Maria Vinţan

Subsummation automata in conceptual graphs 15
Daniel C. Crăciunean

Implementation of Heuristic Algorithms for Simulating Crisis Situations in the Medical System 29
Pătrăuşanu Andrei, Buta Andra-Paraschiva, Florea Adrian

Achieving Clinical Reliability in Suicide Risk Detection: A Low-Resource Benchmark of RoBERTa vs. DistilBERT 43
Stefania-Eliza Berghia, Adrian Barglazan

Sensorless Vector Control of a Surface Permanent Magnet Synchronous Motor 53
Gabriela Crăciunaş, Alina Cristina Viorel, Iulian Bouleanu

Deepfake Voice Detection for Underrepresented Languages: A Romanian Case Study 63
Bălăşoiu Robert-Alin, Ivaşcu Ioana-Daniela, Mihu Cantemir, Muntean Robert-Andrei, Olescu Marco Leon, Resiga Sorana-Ioana

SDR receiver for power measurement with internal noise compensation 73
Iulian Bouleanu, Mădălina Antoneac, Macarie Breazu, Alina Cristina Viorel, Gabriela Crăciunaş

Forecasting and reliability 87
Ioan Lucian Diodiu

GRADIS: A Comprehensive System for Academic Contribution Tracking and Evaluation at Lucian Blaga University of Sibiu 93
Pitic Antoniu Gabriel, Mihu Cantemir, Pitic Elena Alina

IoT System for Milk Dispenser Monitoring and Management 111
Claudia Banciu, Teodora-Artemisa Menchiu, Adrian Florea

Integrating Multi-View features into diagrammatic languages 129
Daniel C. Crăciunean

Optimization of Romanian Identity Documents Processing 141
Delia-Victoria Milaşcon, Constantin Constantinescu, Raluca Brad

Open-loop BLDC motor simulation using the Motor Control blockset from Matlab library 157
Alina Cristina Viorel, Gabriela Crăciunaş

Deep Learning Segmentation for stem cells images 169
Constantin-Cristian DRĂGHICI, Cătălina NEGHINĂ

ISSN 2067 – 354X
eISSN 2067 - 354X

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

VOLUME 15 *Number 1* *2025*

Emergent Strategies in Multi-Agent Systems 179
Vlad OLEKSIK, Cantemir MIHU, Antoniu PITIC
Building a power meter for cycling 191
Raimund TOMA, Macarie BREAZU
A Deep Dive into Wavelet-Based Neural Architectures for Inpainting Forgery Detection 201
Adrian Alin Barglazan, Remus Brad, Pitic Elena Alina, Berghia Ştefania Eliza
Office Ergonomics and Ambiental Environment Monitoring System, 219
MANTA Aida-Mihaela
Wireless Controller for Modbus RS485 225
Mihai-Liviu Stroila

ISSN 2067 – 354X
eISSN 2067 - 354X