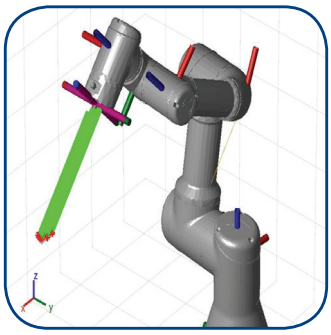
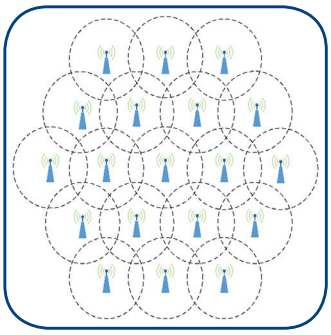
Volume 16, Number 10, 2025 ISSN 1847-6996

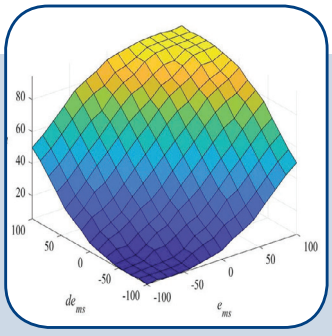
FERIT
FACULTY OF ELECTRICAL ENGINEERING, COMPUTER SCIENCE AND INFORMATION TECHNOLOGY **OSIJEK**

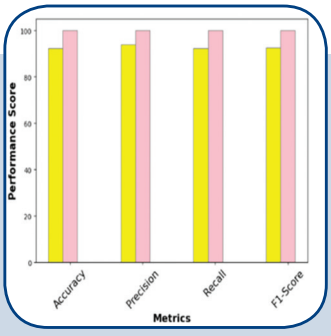
IJECES International Journal of Electrical and Computer Engineering Systems

International Journal of Electrical and Computer Engineering Systems









ISSN: 1847-6996

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

Published by Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Josip Juraj Strossmayer University of Osijek, Croatia

Osijek, Croatia | Volume 16, Number 10, 2025 | Pages 707 - 793

The International Journal of Electrical and Computer Engineering Systems is published with the financial support of the Ministry of Science and Education of the Republic of Croatia

CONTACT

International Journal of Electrical and Computer Engineering Systems (IJECES)

Faculty of Electrical Engineering, Computer Science and Information Technology Osijek, Josip Juraj Strossmayer University of Osijek, Croatia
Kneza Trpimira 2b, 31000 Osijek, Croatia
Phone: +38531224600, Fax: +38531224605
e-mail: ijeces@ferit.hr

Subscription Information

The annual subscription rate is 50€ for individuals, 25€ for students and 150€ for libraries.
Giro account: 2390001 - 1100016777, Croatian Postal Bank

EDITOR-IN-CHIEF

Tomislav Matić
J.J. Strossmayer University of Osijek, Croatia

Goran Martinović
J.J. Strossmayer University of Osijek, Croatia

EXECUTIVE EDITOR

Mario Vranješ
J.J. Strossmayer University of Osijek, Croatia

ASSOCIATE EDITORS

Krešimir Fekete
J.J. Strossmayer University of Osijek, Croatia

Damir Filko
J.J. Strossmayer University of Osijek, Croatia

Davor Vinko
J.J. Strossmayer University of Osijek, Croatia

EDITORIAL BOARD

Marinko Barukčić
J.J. Strossmayer University of Osijek, Croatia

Tin Benšić
J.J. Strossmayer University of Osijek, Croatia

Matjaz Colnarič
University of Maribor, Slovenia

Aura Conci
Fluminense Federal University, Brazil

Bojan Čukić
University of North Carolina at Charlotte, USA

Radu Dobrin
Mälardalen University, Sweden

Irena Galić
J.J. Strossmayer University of Osijek, Croatia

Ratko Grbić
J.J. Strossmayer University of Osijek, Croatia

Krešimir Grgić
J.J. Strossmayer University of Osijek, Croatia

Marijan Herceg
J.J. Strossmayer University of Osijek, Croatia

Darko Huljenić
Ericsson Nikola Tesla, Croatia

Željko Hocenski
J.J. Strossmayer University of Osijek, Croatia

Gordan Ježić
University of Zagreb, Croatia

Ivan Kaštelan
University of Novi Sad, Serbia

Ivan Maršić
Rutgers, The State University of New Jersey, USA

Kruno Miličević
J.J. Strossmayer University of Osijek, Croatia

Gaurav Morghare
Oriental Institute of Science and Technology, Bhopal, India

Srete Nikolovski
J.J. Strossmayer University of Osijek, Croatia

Davor Pavuna
Swiss Federal Institute of Technology Lausanne, Switzerland

Marjan Popov
Delft University, Nizozemska

Sasikumar Punnekkat
Mälardalen University, Sweden

Chiara Ravasio
University of Bergamo, Italija

Snježana Rimac-Drlje
J.J. Strossmayer University of Osijek, Croatia

Krešimir Romić
J.J. Strossmayer University of Osijek, Croatia

Gregor Rozinaj
Slovak University of Technology, Slovakia

Imre Rudas
Budapest Tech, Hungary

Dragan Samardžija
Nokia Bell Labs, USA

Cristina Seceleanu
Mälardalen University, Sweden

Wei Siang Hoh
Universiti Malaysia Pahang, Malaysia

Marinko Stojkov
University of Slavonski Brod, Croatia

Kannadhasan Suriyan
Cheran College of Engineering, India

Zdenko Šimić
The Paul Scherrer Institute, Switzerland

Nikola Teslić
University of Novi Sad, Serbia

Jami Venkata Suman
GMR Institute of Technology, India

Domen Verber
University of Maribor, Slovenia

Denis Vranješ
J.J. Strossmayer University of Osijek, Croatia

Bruno Zorić
J.J. Strossmayer University of Osijek, Croatia

Drago Žagar
J.J. Strossmayer University of Osijek, Croatia

Matej Žnidarec
J.J. Strossmayer University of Osijek, Croatia

Proofreader

Ivanka Ferčec
J.J. Strossmayer University of Osijek, Croatia

Editing and technical assistence

Davor Vrandečić
J.J. Strossmayer University of Osijek, Croatia

Stephen Ward
J.J. Strossmayer University of Osijek, Croatia

Dražen Bajer
J.J. Strossmayer University of Osijek, Croatia

Journal is referred in:

- Scopus
- Web of Science Core Collection (Emerging Sources Citation Index - ESCI)
- Google Scholar
- CiteFactor
- Genamics
- Hrčak
- Ulrichweb
- Reaxys
- Embase
- Engineering Village

Bibliographic Information
Commenced in 2010.
ISSN: 1847-6996
e-ISSN: 1847-7003
Published: quarterly
Circulation: 300

IJECES online
https://ijeces.ferit.hr

Copyright

Authors of the International Journal of Electrical and Computer Engineering Systems must transfer copyright to the publisher in written form.

TABLE OF CONTENTS

Robust and Adaptive Position Control of Pneumatic Artificial Muscles Using a Fuzzy PD+I Controller

Original Scientific Paper

Vinh-Phuc Tran
Can Tho University,
Faculty of Automation Engineering, Automation Lab
3/2 Street, Can Tho, Vietnam
phuctv@vlute.edu.vn

Chi-Ngon Nguyen
Can Tho University,
Faculty of Automation Engineering
3/2 Street, Can Tho, Vietnam
ncngon@ctu.edu.vn

Nhut-Thanh Tran
Can Tho University,
Faculty of Automation Engineering
3/2 Street, Can Tho, Vietnam
nhutthanh@ctu.edu.vn

Chanh-Nghiem Nguyen*
Can Tho University,
Faculty of Automation Engineering, Automation Lab
3/2 Street, Can Tho, Vietnam
ncnghiem@ctu.edu.vn

*Corresponding author

Abstract *– This study evaluates the effectiveness of a fuzzy PD+I (FPD+I) controller for robust and adaptive position control of pneumatic artificial muscles (PAMs), addressing the challenges arising from system nonlinearity and hysteresis. Experiments were conducted under varying loads, setpoints, and actuated distances to assess the robustness and adaptability of the controller under diverse conditions. As part of the evaluation, the results were compared with those obtained using a conventional PID controller. The FPD+I controller consistently demonstrated superior transient response characteristics, improved trajectory-tracking accuracy, and greater adaptability to dynamic operational changes. Notable improvements include a 21% reduction in settling time and a 22% reduction in rise time under constant loads, as well as a 49% improvement in root mean square error and a 24% reduction in rise time during trajectory-tracking tasks. The controller also exhibited enhanced resilience to continuous load disturbances and maintained stable performance under varying signal amplitudes. These findings suggest that the FPD+I controller is a promising solution for precision control applications in robotics and industrial systems employing PAMs, particularly in dynamic and uncertain environments, where both robustness and adaptability are critical.*

Keywords: *Pneumatic artificial muscle (PAM), fuzzy PD+I control, adaptive control, robust control, position control*

Received: May 5, 2025; Received in revised form: July 7, 2025; Accepted: July 24, 2025

1. INTRODUCTION

A pneumatic artificial muscle (PAM) is a type of soft actuator that mimics the function of biological muscles using compressed air to generate force and motion. A PAM consists of a rubber tube encased in a braided mesh, which creates a contraction motion when pressurized [1]. When air is pumped in, the rubber tube expands radially and contracts longitudinally, generating pulling force and torque [2]. PAMs have the advantages of being lightweight, flexible, and capable of producing forces many times greater than their weight [3, 4]. These characteristics make PAMs ideal choices for applications in robotics [5], medical rehabilitation devices [6, 7], and industrial automation systems [8, 9]. However, controlling PAMs presents challenges because of their nonlinear characteristics, variability in pneumatic pressure parameters [10, 11], and dynamic hysteresis [12]–[14], which complicate the control process [15-17]. Additionally, air temperature and pressure also affect the PAM performance [18], [19]. Therefore, developing adaptive and efficient controllers to address these limitations is essential for fully exploiting the potential of PAMs in practical applications. This research area has attracted significant attention from both the scientific and engineering communities.

Over the past decade, various control approaches have been proposed to address these challenges. One major direction involves hysteresis compensation to overcome the system's nonlinearity and dynamic hysteresis. Many hysteresis compensation methods were proposed based on different variants of Prandtl-Ishlinskii (PI) models. Among them, based on the extended unparallel Prandtl-Ishlinskii, an integral inverse-proportional-integral-derivative controller can improve the control precision by 43.86% in comparison with a conventional proportional–integral–derivative (PID) controller without hysteresis compensation [15]. The feedforward and feedback combined control strategies demonstrated a maximum translational error of approximately 0.9 mm (i.e., 0.604% of the full stroke) [20] and a maximum tracking error of the rotation angle of the delta mechanism of ± 0.7° [16] in response to a low-frequency input sinusoidal signal of 0.1 Hz. However, these approaches require complex PAM modeling in real-world systems. Feasible employment of such controllers also requires them to withstand external disturbances and variations in operating conditions [21], which were not comprehensively reported in recent studies.

Although model-based identification methods have also been proposed and demonstrated promising results in accurately modeling the nonlinear dynamics of PAM actuators [22, 23], methods integrating hysteresis modeling and compensation have recently attracted research focus. For instance, Zhang et al. (2024) developed a dynamic model capturing coupled, stiffness- and rate-dependent hysteresis in soft PAM manipulators [24]. Their decoupled inverse compensation strategy improved positioning accuracy and highlighted the complexity of controlling variable-stiffness actuators. Another promising approach is model predictive control (MPC), which can anticipate system behavior and adjust control actions accordingly. Brown and Xie (2025) proposed a PSO-optimized MPC framework that outperformed traditional PID and iterative learning controllers in terms of accuracy and responsiveness in a rehabilitation setting [25]. Zhang et al. (2024) further demonstrated the potential of a disturbance preview-based predictive control strategy, which combined with hysteresis compensation and high-order differentiators, enabled robust trajectory tracking and model simplification for PAM-driven robots [26].

Despite their effectiveness, these methods often require complex system modeling, accurate hysteresis characterization, and high computational resources, making them less practical for real-time control in uncertain and dynamically changing environments. This limitation has led to increasing interest in classical or intelligent control schemes that do not compensate for hysteresis but ensure control effectiveness under certain conditions. Conventional PID controllers have proven to be effective in applications that do not require fast responses or industrial processes, where highly accurate positioning is not very demanding [27, 28]. Phuc et al. (2022) demonstrated an acceptable translational position control of a 25-kg payload with a settling time of only 1 s, insignificant overshoot, and a 0.35-mm steady-state error [27]. A rising time of 8.5 s and steady-state error of 5.129 mm were obtained by a PID controller when controlling the angle of an antagonistic dual-PAM system [28]. When higher positioning accuracy is prioritized, intelligent control systems are generally implemented to tackle PAM nonlinearity due to hysteresis. For example, high-order sliding mode control can be used to control a PAM-actuated robot arm that can follow an objective circular locus with a maximum error of approximately 11.3 mm [29]. To assist rehabilitation exercises, an adaptive sliding controller with a PID compensator can improve the tracking accuracy and reduce the steady-state tracking error of a robotic arm to below 5° [4]. A fuzzy sliding mode controller can be implemented with a steady-state error of 0.003°, satisfying the requirement of high accuracy for lower-limb systems [30]. Based on iterative real-time learning of Generative Adversarial Nets, a PID controller maintains effective tracking performance of antagonistic pneumatic artificial muscles under variations in their initial conditions, target angles, and external disturbances, demonstrating an advancement towards practical utilization of prostheses [31].

Despite advancements in PAM control strategies, several challenges remain in achieving robust and precise control under various operating conditions. A key area of focus is to improve the adaptability of control systems to handle a wider range of uncertainties and disturbances. For instance, the adaptive fixed-time fast terminal sliding mode control proposed by Khajehsaeid et al. (2025) demonstrated robustness against model uncertainties, assuming a 10% difference between actual and nominal link masses [32]. However, real-world applications may involve even greater variations in system parameters and external disturbances. Similarly, the adaptive control strategy presented by Sun et al. (2020) for PAM systems with parametric uncertainties and unidirectional input constraints exhibited resilience to more significant load variations from 0.633 to 1.633 kg and external disturbances applied at specific time intervals [17]. Although these results are promising, further research is required to expand the range of uncertainties and disturbances that can be effectively managed. In the context of lower-limb robotics, Tsai and Chiang (2020) demonstrated the effectiveness of a fuzzy sliding mode controller in achieving high accuracy, with steady-state errors of 0.003° without load and 0.00338° with a 2 kg load [30]. This level of precision is crucial for rehabilitation and assistive device applications. However, the performance of such controllers under more dynamic conditions, particularly when subjected to unexpected external forces, requires further investigation.

Another issue is that an effective control system must adapt to the continuous changes of the system dynamics. This capability enhances the operational performance and broadens the application range of PAM in di-

verse environments and tasks. Possible solutions include adaptive control or fuzzy-based strategies, as they have proven to be effective for external disturbances and varying system parameters [17], [30], [32]. Among fuzzy-based approaches, a combination of a fuzzy and PID controller is a feasible control scheme offering a compelling blend of simplicity, adaptability, good transient response, and accurate control. The fuzzy PD+I (FPD+I) controller, with its inherent adaptability and potential for high accuracy and quick response of position control, has emerged as a promising solution to meet the essential need for control strategies that can adapt to changing conditions in real time while maintaining adequate accuracy. Compared with a conventional PID controller with three branches of proportional, integral, and derivative components, the FPD+I controller comprises an adjustable fuzzy PD branch and a nonfuzzy integral component. The fuzzy PD branch is critical for the control system to dynamically adapt to the inherent nonlinearities and disturbances of the PAM and enable a fast response, whereas the nonfuzzy integral branch helps reduce the steady-state error [33]. This cumulative capacity is more pronounced than that of the PID controller. Thus, high control performance, flexible adaptability, and robustness can be achieved with FPD+I controllers in practical applications [34, 35]. Chan et al. (2003) further demonstrated that coupling an FPD+I controller with a learning control scheme enables accurate tracking control of PAMs [36], reinforcing its potential for robust and adaptive position control.

Despite these advancements, existing studies have not fully explored the potential of FPD+I controllers to address the combined challenges of nonlinearity, complex trajectory tracking, and continuous and complex disturbances in PAM systems. Therefore, this study specifically addresses the challenge of improving the accuracy and adaptability of PAM control under more complex operating conditions, particularly under different initial positions, tracking trajectories, and particularly continuously varying loads in a greater range, by deploying an FPD+I controller without any online learning scheme.

The main contributions of this study are summarized as follows:

- Development and real-time implementation of a fuzzy PD+I controller for PAM position control, eliminating the need for complex dynamic models or online adaptation mechanisms.
- Comprehensive experimental validation under a wide range of practical conditions, including fixed and continuously varying loads, multiple setpoints, and diverse trajectory profiles, to assess controller adaptability and robustness.
- Quantitative comparison with a conventional PID controller, demonstrating that the proposed FPD+I approach significantly improves tracking accuracy, reduces rise and settling times, and maintains performance under dynamic disturbances.
- A simple and practical controller design using a compact fuzzy rule base and no learning phase, enabling straightforward deployment in real-time embedded systems.

The remainder of this paper is organized as follows. Section 2 describes the experimental setup, system configuration, and the design of the proposed FPD+I control algorithm. Section 3 presents the experimental results along with a detailed performance analysis and a comparative evaluation against a conventional PID controller. Finally, Section 4 concludes the study by summarizing the key findings and outlining potential directions for future research.

2. MATERIALS AND METHODS

2.2. EXPERIMENTAL SETUPS

The experiment was conducted on a PAM with a diameter of 20 mm and an initial length of 200 mm (MAS-20-200N, FESTO). The PAM could shrink up to 20% of its original length, that is, a 40-mm distance, although a maximal 25% shrinkage at 6 bar has been reported for a new PAM [2]. The PAM displacement and pressure applied to the PAM were measured using an AccuracyTM KTC 100 rangefinder and Georgin SR13002A pressure sensor, respectively. Compressed air was supplied through a 5/3 proportional-directional control valve (MPYE-5-1/8-HF-101B; FESTO). The control and measurement algorithms were implemented using the MATLAB/Simulink software. All I/O data were transferred by using a Texas Instruments C2000 microcontroller (TMS320F28379D LaunchPAD). An overview of the proposed model is presented in Fig. 1.

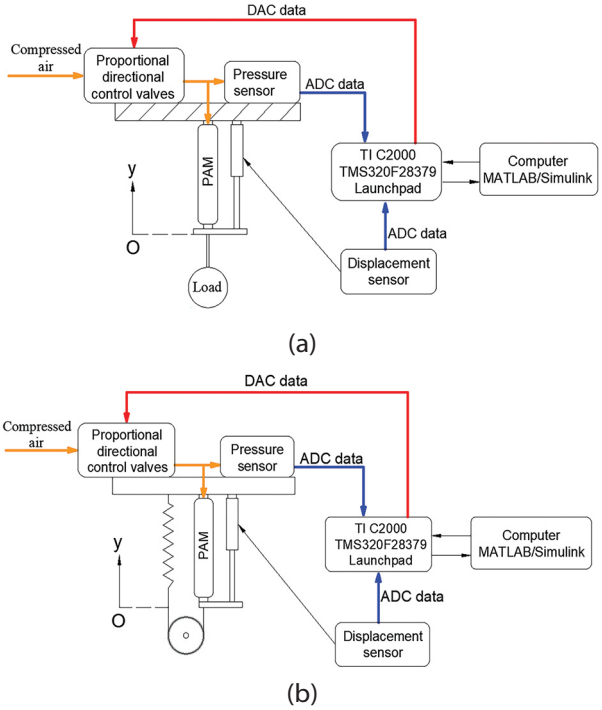


Fig. 1. Experimental setup with **(a)** constant load and **(b)** varying load

Six experiments were conducted in this study with two main goals: to investigate the effect of the load and the effect of different amplitudes and amplitude shifts of the input signal; hence, the setpoints and PAM's actuated distance on the position responses of the system. The experimental conditions and objectives are listed in Table 1. In Experiments 1.2 and 2.2, the robustness of the system was evaluated in response to direction changes in PAM actuation. This was done using 0.1 Hz rectangular pulse trains as reference inputs, formulated as

$$S_{A,c}(t) = A\,\mathrm{sgn}\left(\sin\left(\pi t/5\right)\right) + c \tag{1}$$

where $sgn(.)$ is the sign function, A is the signal amplitude representing half the required actuated distance of the PAM, and c is the amplitude shift defining the initial position around which the bidirectional displacement of the PAM occurs. Both A and c were selected such that the reference input remained within the controllable range of the PAM.

To further explore the robustness of the system to continuously varying disturbances, a spring was added to the PAM's actuating head in Experiments 1.3 and 2.3, as shown in Fig. 1b. This spring was preloaded to create an initial elastic force simulating an equivalent static mass. As the PAM was actuated, the spring displacement increased, generating a continuous load disturbance that tested the robustness of the system under dynamic loading conditions.

Table 1. Conditions and goals of the position control experiments

Experiment No.	Main goal	Conditions		
		Description	Reference input (mm)	Mass (kg)
1.1	Investigate the effect of various disturbances on the load	Step input with various fixed loads	20	{15, 20, 25}
1.2		Rectangular pulse train with various fixed loads	$S_{5,10}(t)$=5 $sgn(sin(\pi t/5))$+10	{15, 20, 25}
1.3		Step input with disturbances to different simulated loads	20	Varying load from an initial equivalent load of {15, 20, 25} kg
2.1	To investigate the effect of different amplitudes and amplitude shifts of the input signal	Step input of different amplitudes	{10, 20, 30}	20
2.2		Rectangular pulse trains of different amplitudes and amplitude shifts	$S_{A,c}(t)$=A $sgn(sin(\pi t/5))$+c where c∈{15,20,25}, A∈{3,5}	20
2.3		Step input of different amplitudes with disturbances to a simulated load	$S_{A,c}(t)$=A $sgn(sin(\pi t/5))$+c where c∈{15,20,25}, A∈{3,5}	Varying load from an initial equivalent load of 20 kg

2.2. CONTROL ALGORITHMS

As shown in Fig. 2, the control system includes two nested control loops. The inner control loop employs a PI controller to regulate the air pressure, ensuring stability for the outer loop to govern the PAM position. The PI controller receives the control signal u and the measured air pressure P from a pressure sensor to generate a pressure control signal, which is then applied to the proportional electro-pneumatic valve, maintaining consistent and stable pressure regulation.

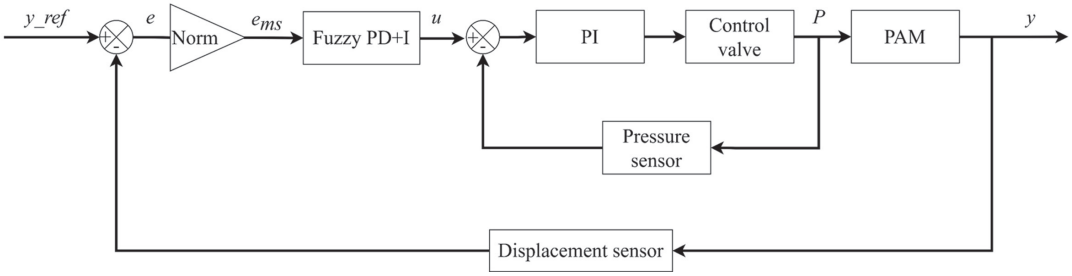


Fig. 3. Structure of the fuzzy PD+I control system

In the outer loop, an FPD+I controller manages the position control. It receives the position error signal e, calculated as the difference between the reference input y_ref and the measured PAM displacement y. The error is normalized by a Norm block to scale its magnitude appropriately. The FPD+I controller then processes the normalized error ems to generate the control signal u for the inner PI control loop. The normalized error is measured as the percent error with respect to the maximal shrinkage from its idle position, calculated as:

$$e_{ms}(k) = \frac{e(k)}{40} \times 100\% \tag{2}$$

The combined action of the outer FPD+I controller and the inner PI controller enables precise control of the PAM position while compensating for pressure fluctuations and external disturbances. The displacement sensor continuously measures the actual position y, thereby closing the outer feedback loop. A summary of the signals shown in Fig. 2 is presented in Table 2.

Table 2. Main signals in the control system

Parameter	Description	Unit
y_{ref}	Reference displacement input	mm
y	Measured displacement output	mm
e	Position error ($e = y_{ref} - y$)	mm
e_{ms}	Normalized error	%
u	Control signal applied to the inner control loop	%
P	Measured air pressure inside PAM	bar

2.2.1. PI controller

The pressure inside the PAM depends on the nonlinearity of the airflow through the proportional directional control valve, volume variation, air temperature, and air leakage through the valve. Therefore, to ensure a stable air pressure, a PI controller was implemented. The controller gains were empirically determined and kept constant in all experiments at $K_P = 2$ and $K_I = 4.6$.

2.2.2. PID controller

The PID controller computes the control signal $u_{PID}(k)$ at each discrete time step k based on the time-domain error between the desired position and actual measured position of the PAM, as follows:

$$u_{PID}(k) = K_P \cdot e_{ms}(k) + K_I \cdot \sum_{i=0}^{k} e_{ms}(k) \cdot T_s + K_D \cdot \frac{e_{ms}(k) - e_{ms}(k-1)}{T_s} \quad (3)$$

where K_P, K_I, and K_D are the proportional, integral, and derivative gains of the PID controller, respectively, and $e_{ms}(k)$ is the percent error at time step k.

The gains of the PID controller were empirically determined to be $K_P = 0.2$, $K_I = 1.0$, and $K_D = 0.01$. The sampling period Ts was 0.01 s.

2.2.3. Fuzzy PD+I controller

An FPD+I controller, whose structure is shown in Fig. 3, was built to control the position of the PAM. A Mamdani fuzzy inference system (FIS) was implemented using the Fuzzy Logic Toolbox, MATLAB.

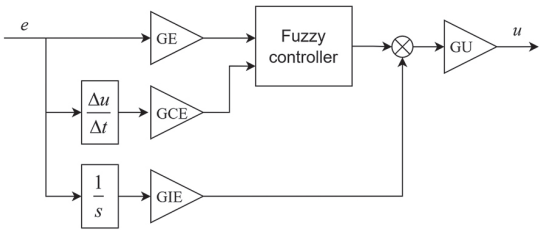


Fig. 3. The FPD+I structure

The output from the fuzzy PD branch of the FPD+I controller at time k is calculated as:

$$u_{FPD}(k) = F\left(e_{ms}(k), de_{ms}(k)\right) \quad (4)$$

where $F(.)$ denotes the mapping realized by the fuzzy rule base with the Gaussian membership functions.

The output from the integral branch of the FPD+I controller component is defined as

$$u_I(k) = GIE \sum_{i=0}^{k} e(i) \cdot T_s \quad (5)$$

The control output of the FPD+I controller is a combination of the outputs from the fuzzy PD and integral branches, calculated as follows:

$$u(k) = GU * \left[u_{FPD}(k) + u_I(k)\right] \quad (6)$$

where GU is the overall gain-scaling factor.

Seven Gaussian membership functions were defined based on the percent error e_{ms} and the derivative of the percent error. Fuzzy control rules were established, as listed in Table 3. The response surface based on these control rules is shown in Fig. 4. The crisp output of the FPD+I controller obtained from the fuzzy variables agrees with that of a previous report [37].

The range of the control output u was [0, 100%], corresponding to the airflow capacity of the pneumatic valve. The proportional gains GE, GCE, GIE, and GU (Fig. 2) were determined to satisfy the operating limit of the controller's input so that [38]

$$GE * GU = K_P \quad (7)$$

$$GCE/GE = T_D \quad (8)$$

$$GIE/GE = 1/T_I \quad (9)$$

where $K_P = 0.2$, $T_I = 0.2$, and $T_D = 0.05$, which were empirically selected and remained unchanged during the tests.

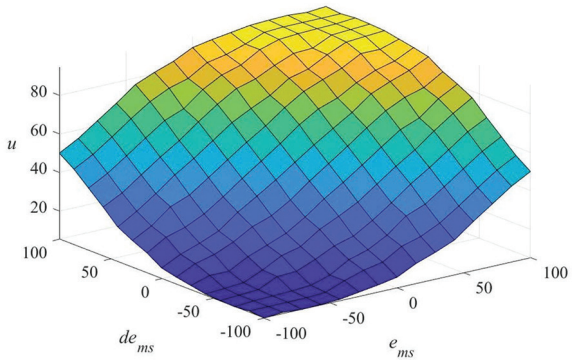


Fig. 4. The response surface

Table 3. Fuzzy rules for controlling PAM position

e_{ms} ∈ [-100, 100%]	de_{ms} ∈ [-100, 100%]						
	NL	NM	NS	Z	PS	PM	PL
VVS	VVS	VVS	VVS	VVS	VS	S	M
VS	VVS	VVS	VVS	VS	S	M	L
S	VVS	VVS	VS	S	M	L	VL
M	VVS	VS	S	M	L	VL	VVL
L	VS	S	M	L	VL	VVL	VVL
VL	S	M	L	VL	VVL	VVL	VVL
VVL	M	L	VL	VVL	VVL	VVL	VVL

VVS: very very small; VS: very small; S: small; M: medium; L: large; VB: very large; VVB: very very large; NL: negatively large; NM: negatively medium; NS: negatively small; Z: zero; PS: positively small; PM: positively medium; PL: positively large.

A conventional PID controller was designed through trial and error for comparison with the FPD+I controller. The initial proportional, integral, and derivative gains of the PID controller were empirically determined and adjusted to balance the rise time, settling time, and overshoot, thereby ensuring its effectiveness under typical operating conditions without excessive oscillations or delays.

2.3. EVALUATION OF THE MODEL PERFORMANCE

Root mean squared error (RMSE) is selected to quantitatively evaluate the response of the position controller, calculated as

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{\left(y_i - \hat{y}_i\right)^2}{N}} \tag{10}$$

where $\hat{y}_i$ is the response of the corresponding set-point value $\hat{y}_i$ in the dataset of N samples. In this study, the PAM displacement and RMSE were computed in millimeters.

The system performance was also evaluated based on the transient response characteristics, including the percent overshoot (POT), rise time, and settling time, under various experimental conditions and setups. Because of the low response characteristic of PAM, the rise time was calculated as the time required for the response to increase from 0% to 90% of its reference value (rather than its final value). For better reference, the percent of improvement (PoI) was calculated to show the relative performance improvement in a certain performance metric, including RMSE, POT, rise time, and settling time, as

$$PoI = \frac{m_{PID} - m_{FPD+I}}{m_{PID}} \times 100\% \tag{11}$$

where m_{FPD+I} and m_{PID} are the values of a performance metric (i.e., RMSE, POT, rise time, or settling time) obtained using the FPD+I and PID controllers, respectively. A positive PoI indicates an enhancement in control performance compared to the baseline PID controller, whereas a negative PoI implies a deterioration of the desired response.

To evaluate the effect of load disturbances, the percent of degradation (PoD) was formulated to show the performance degradation in a performance metric as follows:

$$PoD = \frac{\left|\bar{m}_{disfree} - \bar{m}_{dis}\right|}{\bar{m}_{disfree}} \times 100\% \tag{12}$$

where $\bar{m}_{dis}$ and $\bar{m}_{disfree}$ are the mean values of a performance metric (i.e., POT, rise time, or settling time) obtained with and without the load disturbances, respectively. The PoD reflects the extent to which external disturbances degrade the system with respect to a performance metric. Therefore, PoI and PoD provide a systematic and quantitative basis for assessing both the effectiveness of the proposed control strategy and its robustness under varying operating conditions.

3. RESULTS AND DISCUSSION

3.1. EFFECT OF LOAD ON THE POSITION RESPONSE

3.1.1. Experiment 1.1: Step response analysis under different loads

In Experiment 1.1, a set-point position of $r = 20$ mm was applied to the system under loads of 15 kg, 20 kg, and 25 kg. The corresponding step responses are shown in Fig. 5. The performance of the FPD+I and PID controllers are listed in Table 4. Compared with the PID controller, the FPD+I controller had a lower POT, shorter rise time, and shorter settling time of at least 14%, 22%, and 21%, respectively, than the PID controller. These results demonstrate its superior performance compared with the PID controller in transient responses under different loads.

3.1.2. Experiment 1.2: Trajectory tracking with a rectangular pulse train under varying loads

In this experiment, a 0.1-Hz rectangular pulse train $S5{,}10\,(t)=5\,sgn(sin(\pi t/5))+10$ was applied as the reference input to the system under different loads of 10, 20, and 25 kg. In addition to evaluating the load effects on the position control performance, this experiment was conducted to partly evaluate the robustness of the control system in response to a tracking trajectory of higher complexity.

As shown in Fig. 6, the FPD+I controller exhibits quicker responses than the conventional PID controller. Compared with the PID controller, the FPD+I controller reduced the RMSE and rise time by 49% and 24%, respectively (Table 5), demonstrating that the FPD+I controller was more accurate and robust for complex trajectory-tracking applications under significant load variations of 15–25 kg.

It should be noted that an insignificant overshoot was observed with the FPD+I controller, whereas the system exhibited an underdamped response, that is, the PAM could not reach the desired setpoint when using the PID controller.

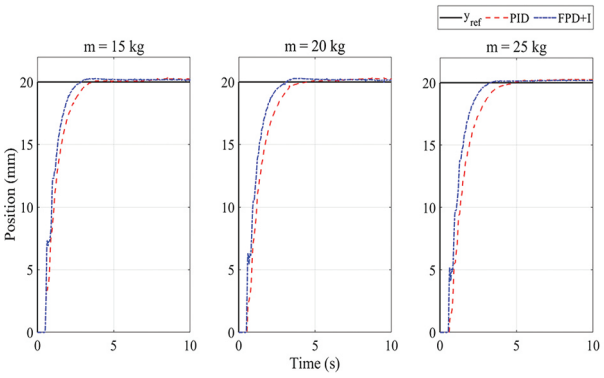


Fig. 5. Step responses under different loads (Experiment 1.1)

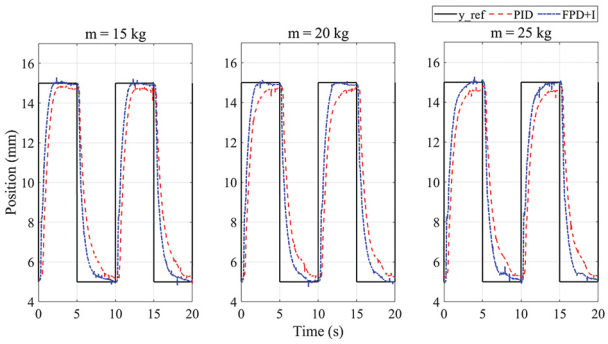


Fig. 6. Position responses of the PAM with different loads to a 0.1 Hz rectangular pulse train (Experiment 1.2)

3.1.3. Experiment 1.3: Step response analysis under varying simulated loads with continuous load disturbances

Experiment 1.3 was conducted in alignment with Experiment 1.1. However, the corresponding loads were simulated using the equivalent elastic force of a string with a proper displacement. Based on the setup shown in Fig. 1b, a continuous load disturbance can be applied to the system to evaluate the robustness of the FPD+I controller. As the PAM reached setpoint r = 20 mm, the elastic force increased, leading to equivalent load increases of 130%, 131%, and 155% from the initial simulated loads of 15, 20, and 25 kg, respectively. Fig. 7 shows the step responses of the PAM system under different simulated loads with continuous disturbances. The detailed system performance is listed in Table 6. Because of these significant and continuous load disturbances, minor POTs were observed, and the settling and rise times were greater than those of the corresponding disturbance-free cases in Experiment 1.1 (Table 4).

Based on the average performance metrics summarized in Table 7, the FPD+I controller consistently outperformed the PID controller under continuous load disturbances. Although both controllers exhibited performance degradation in the presence of dynamic loads, the FPD+I controller demonstrated greater robustness, as reflected by the smaller increases in rise and settling times. Specifically, the FPD+I controller experienced increases of only 0.5 s in rise time and 0.8 s in settling time, compared to larger increases of 0.7 s and 1.0 s, respectively, for the PID controller. Furthermore, the slightly lower percent of degradation (PoD) values observed for the FPD+I controller (Table 8) further confirm its superior resilience to continuous external disturbances.

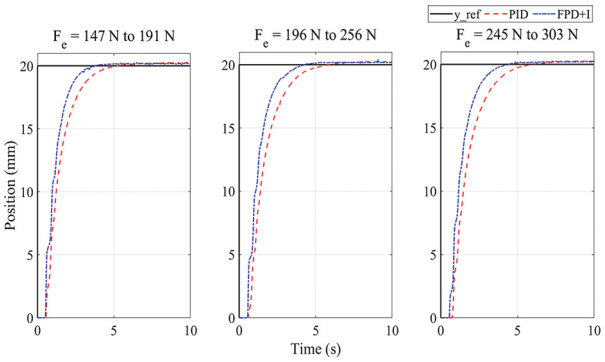


Fig. 7. Position response of PAM with varying elastic forces simulated by springs (Experiment 1.3)

Table 4. Performance of FPD+I and PID controllers under different loads (Experiment 1.1)

	POT			Rise time			Settling time		
Mass (kg)	FPD+I (%)	PID (%)	*Pol* (%)	FPD+I (s)	PID (s)	*Pol* (%)	FPD+I (s)	PID (s)	*Pol* (%)
15	1.5	1.9	21	2.1	2.7	22	2.4	3.1	23
20	1.4	2.2	36	2.3	3.0	23	2.7	3.4	21
25	1.8	2.1	14	2.5	3.2	22	2.8	3.9	28

Table 5. Performance of FPD+I and PID controllers in response to a 0.1 Hz rectangular pulse train under different loads (Experiment 1.2)

	RMSE			Rise time		
Mass (kg)	FPD+I (mm)	PID (mm)	*Pol* (%)	FPD+I (s)	PID (s)	*Pol* (%)
15	1.7	3.7	54	1.6	2.1	24
20	1.9	3.9	51	1.6	3.1	48
25	2.0	3.9	49	2.2	3.3	33

Table 6. Performance of the FPD+I and PID controllers under different simulated loads with continuous load disturbances (Experiment 1.3)

	POT			Rise time			Settling time		
Equivalent mass (kg)	FPD+I (%)	PID (%)	*Pol* (%)	FPD+I (s)	PID (s)	*Pol* (%)	FPD+I (s)	PID (s)	*Pol* (%)
15	1.4	1.7	18	2.5	3.4	26	3.0	4.0	25
20	1.9	2.1	10	2.9	3.8	24	3.6	4.6	22
25	1.3	1.7	24	3.0	4.0	25	3.6	4.8	25

Table 7. Average performance of FPD+I and PID controllers with and without continuous load disturbances

	FPD+I controller		PID controller	
Metrics	**Exp. 1.1**	**Exp. 1.3**	**Exp. 1.1**	**Exp. 1.3**
RMSE	3.6 mm	3.9 mm	4.0 mm	4.3 mm
POT	1.6%	1.5%	2.1%	1.8%
Settling time	2.6 s	3.4 s	3.5 s	4.5 s
Rise time	2.3 s	2.8 s	3.0 s	3.7 s

Table 8. Performance degradation of FPD+I and PID controllers due to load disturbances

Metrics	FPD+I	PID
POT	-2,1%	-11,3%
Settling time	29,1%	28,8%
Rise time	21,7%	25,8%

PoD was calculated without rounding up the mean values of a certain metrics

In addition to robustness, the FPD+I controller exhibited better adaptability to dynamic load variations, as indicated by its smaller performance degradation relative to the conventional PID controller, without the need for controller retuning. Quantitatively, the FPD+I controller reduced the percent degradation in the rise and settling times by approximately 29% and 20%, respectively, compared to the PID controller.

3.2. EFFECT OF THE AMPLITUDE OF THE INPUT SIGNAL

3.2.1. Experiment 2.1: Step responses analysis with varying setpoints

The objective of this experiment was to evaluate the ability of the controller to maintain performance across different setpoints compared with a conventional PID controller. A fixed load of 20 kg was applied, and step responses were obtained at setpoint positions of 10, 20, and 30 mm. It should be noted that the step response for r = 20 mm was obtained in Experiment 1.1.

As shown in Table 9 and Fig. 8, the FPD+I controller generally exhibited faster transient responses than the PID controller, particularly at larger setpoint changes. For the short actuated distance of r = 10 mm, the FPD+I controller did not outperform the PID controller in terms of the rise time. The rise time of the FPD+I controller was slightly longer by approximately 2.4%. Although it achieved a significantly faster settling time, reducing it by approximately 2.4 seconds compared to the PID controller, the overshoot resulted in oscillations, causing the final settling time to increase substantially to 5.9 seconds. This value was at least 200% longer than those observed at larger setpoints (r = 20 mm and 30 mm, where settling times were 2.5 s and 2.8 s, respectively).

At longer actuated distances (r = 20 and 30 mm), the FPD+I controller demonstrated substantial improvements in rise time, achieving enhancements of more than 31% compared to the PID controller. However, no significant improvement was observed in the POT compared to the PID controller.

In all cases, the FPD+I controller consistently outperformed the PID controller in terms of settling time. These results highlight the superior transient performance of the FPD+I controller, particularly for larger signal amplitudes. Overall, the findings underscore the controller's superior adaptability to varying setpoint commands and its robustness in maintaining the control performance under different operating demands.

It should also be noted that actuating the PAM initially from the zero position introduced a baseline delay because the muscle needed to inflate before contraction could generate an actuated force. This inflation delay averaged approximately 0.49 seconds and was included in the calculated rise and settling times.

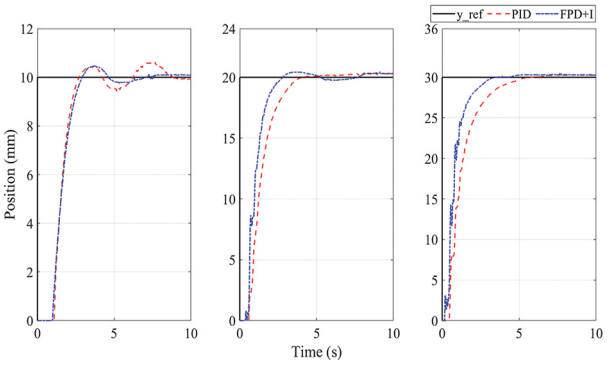


Fig. 8. Position control performance across different setpoints under a 20 kg load (Experiment 2.1)

Table 9. Characteristics of transient responses of FPD+I and PID controllers across varying setpoints (Experiment 2.1)

	POT			Rise time			Settling time		
Position setpoint (mm)	**FPD+I (%)**	**PID (%)**	***PoI* (%)**	**FPD+I (s)**	**PID (s)**	***PoI* (%)**	**FPD+I (s)**	**PID (s)**	***PoI* (%)**
10	4.8	6.2	23	2.61	2.55	-2.4	5.9	8.3	29
20	2.3	2.2	-5	2.12	3.09	31.4	2.5	3.6	31
30	1.7	1.6	-6	2.27	3.5	35.1	2.8	4.5	38

3.2.2. Experiment 2.2: Trajectory tracking with rectangular pulse trains of different amplitudes

This experiment aimed to assess the control system's robustness and adaptability in tracking more complex trajectories. A 0.1-Hz rectangular pulse train was applied as the reference input, with varying amplitudes (A = 3 mm and 5 mm) and amplitude shifts (c = 10, 15, and 25 mm) to evaluate the system performance under different actuated distances of the PAM. The setup was similar to that of Experiment 1.2, but with additional variations to increase the tracking difficulty.

As shown in Fig. 9 and Tables 10–11, the FPD+I controller consistently achieved faster responses than the PID controller, which is consistent with the observations from Experiment 1.2. Specifically, the FPD+I controller achieved notable reductions in rise time, with average improvements of approximately 41% and 32% for amplitudes of 3 and 5 mm, respectively (Table 10).

In terms of positional accuracy, the FPD+I controller outperformed the PID controller, as indicated by the lower RMSE values. The average improvements in the RMSE were approximately 21% and 19% for amplitudes of 3 mm and 5 mm, respectively (Table 11).

It is worth noting that the FPD+I controller exhibited minimal overshoot, indicating good transient behavior, even under complex reference signals. In contrast, the PID controller produced an underdamped response, where the PAM failed to effectively reach the desired setpoints. These results demonstrate the superior robustness and adaptability of the FPD+I controller for managing varying trajectory profiles and actuation distances.

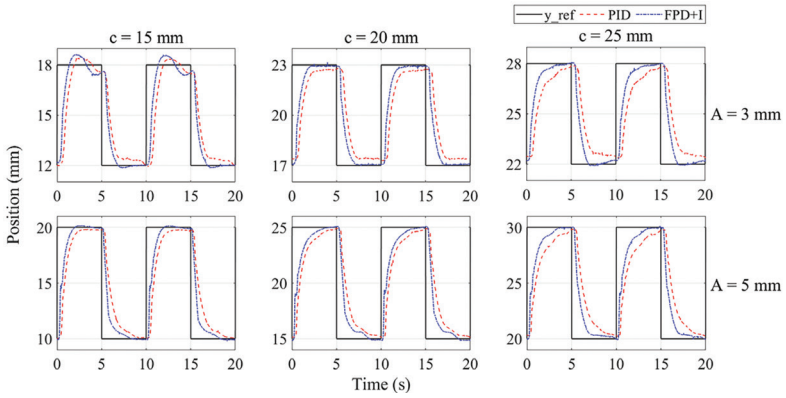


Fig. 9. PAM position control performance under varying amplitudes (A = 3 mm and 5 mm) and amplitude shifts (c = 10–25 mm) using FPD+I and PID controllers (Experiment 2.2)

Table 10. Rise time comparison for the FPD+I and PID controllers with varying amplitudes and amplitude shifts (Experiment 2.2)

Amplitude shift *c* (mm)	Amplitude A = 3 mm			Amplitude A = 5 mm		
	FPD+I	PID	*PoI* (%)	FPD+I	PID	*PoI* (%)
10	1.2	1.7	29	1.5	2.2	32
15	1.7	3.7	54	2.6	3.6	28
25	2.5	4.2	40	2.9	4.6	37
Average	1.8	3.2	41	2.3	3.5	32

Table 11. RMSE comparison for the FPD+I and PID controllers with varying amplitudes and amplitude shifts (Experiment 2.2)

Amplitude shift *c* (mm)	Amplitude A = 3 mm			Amplitude A = 5 mm		
	FPD+I	PID	*PoI* (%)	FPD+I	PID	*PoI* (%)
10	2.0	2.4	17	3.0	3.7	19
15	2.0	2.6	23	3.3	4.0	18
25	2.1	2.7	22	3.3	4.2	21
Average	2.0	2.6	21	3.2	4.0	19

3.2.3. Experiment 2.3: Evaluation of system robustness in trajectory tracking under continuous load disturbances

Similar to Experiment 1.3, this experiment used a spring with an initial deformation to simulate a 20 kg load, introducing a continuous load disturbance due to the changing opposing elastic force during PAM contraction. However, rectangular pulse trains similar to those used in Experiment 2.2 were applied as reference inputs to evaluate the robustness of the system in complex trajectory tracking with varying actuated distances of the PAM under continuous disturbances.

The results in Tables 12 and 13 demonstrate that the FPD+I controller outperformed the conventional PID controller in terms of both the tracking accuracy and response speed under continuous load disturbances. Specifically, the FPD+I controller achieved lower RMSE values across all tested conditions, with average improvements of 18% and 21% for amplitudes A of 3 mm and 5 mm, respectively (Table 12). Similarly, the rise time was significantly reduced, with the FPD+I controller achieving improvements of 34% and 35% compared to the PID controller for amplitudes of 3 mm and 5 mm, respectively (Table 13).

In addition to the improved accuracy and faster responses, it is noteworthy that the FPD+I controller exhibited only an insignificant overshoot, indicating good transient performance (Fig. 10). Conversely, the PID controller resulted in an underdamped response, where the PAM struggled to reach the desired setpoints under the influence of varying load disturbances, particularly for longer actuated distances. These results further confirm the robustness and effectiveness of the FPD+I controller in complex trajectory-tracking scenarios.

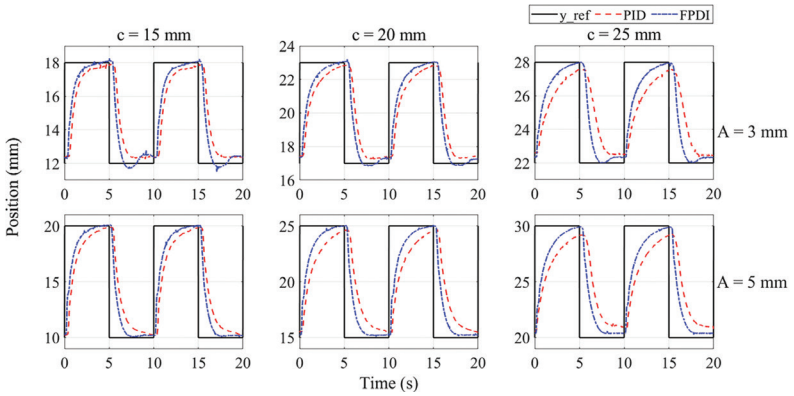


Fig. 10. PAM position response under elastic force with 0.1 Hz square wave input (Experiment 2.3)

Table 12. RMSE comparison for FPD+I and PID controllers with varying amplitudes and amplitude shifts (Experiment 2.3)

Amplitude shift *c* (mm)	Amplitude A = 3 mm			Amplitude A = 5 mm		
	FPD+I	PID	*PoI* (%)	FPD+I	PID	*PoI* (%)
10	2.1	2.5	16	3.2	3.9	18
15	2.2	2.6	15	3.3	4.3	23
25	2.2	2.8	21	3.5	4.5	22
Average	2.2	2.6	18	3.3	4.2	21

Table 13. Rise time comparison for the FPD+I and PID controllers with varying amplitudes and amplitude shifts (Experiment 2.3)

Amplitude shift *c* (mm)	Amplitude A = 3 mm			Amplitude A = 5 mm		
	FPD+I	PID	*PoI* (%)	FPD+I	PID	*PoI* (%)
10	2.4	3.9	38	2.5	3.7	32
15	2.9	4.4	34	3	4.8	38
25	3.5	4.9	29	3.5	5.4	35
Average	2.9	4.4	34	3.0	4.6	35

3.3. ADAPTABILITY ANALYSIS

To further evaluate the performance of the FPD+I controller beyond individual experiments, its adaptability under dynamic and uncertain operating conditions was analyzed. In this context, adaptability refers to the controller's ability to maintain high tracking accuracy and stable transient responses despite variations in load, setpoint amplitudes, and external disturbances. As summarized in Table 14 and illustrated in Fig. 11, the FPD+I controller consistently maintained a lower RMSE, faster rise times, and shorter settling times across a wide range of operating conditions including fixed and varying loads, different setpoints (10–30 mm), and continuous load disturbances (up to 155% equivalent mass increase). Compared to the conventional PID controller, the FPD+I controller exhibited significantly smaller performance variations, demonstrating superior robustness and adaptability.

Fig. 11 shows that while the PID controller's tracking accuracy and transient behavior degraded noticeably under changing conditions, the FPD+I controller maintained a stable performance with minimal degradation. These findings confirm that the FPD+I controller can dynamically adjust to real-time variations in system behavior without requiring model-based compensation or parameter

Table 14. Performance consistency of FPD+I controller across different operating conditions

Condition type	Variation range	Performance (RMSE, rise time, settling time)	Evaluation remarks
Load variations	15–25 kg (fixed and varying)	Consistent; low RMSE; rise/settling times slightly affected	Robust to large static and dynamic load changes
Setpoint variations	10–30 mm	Minimal impact on transient response; minor overshoot for the smaller setpoints	Good adaptability to different target positions; significant increase in settling time for small setpoint (r = 10 mm)
Signal amplitude and amplitude shift	Amplitude: 3–5 mm; Amplitude shift: 15–25 mm	Maintained fast response and low error	Well adapted to varying amplitude shifts
Continuous load disturbance	130%–155% equivalent mass increase	Performance degradation <30%	High resilience under dynamic external forces

retuning. Overall, the results validated the effectiveness of the FPD+I controller in providing both robust and adaptive position control for pneumatic artificial muscles operating in dynamic and uncertain environments.

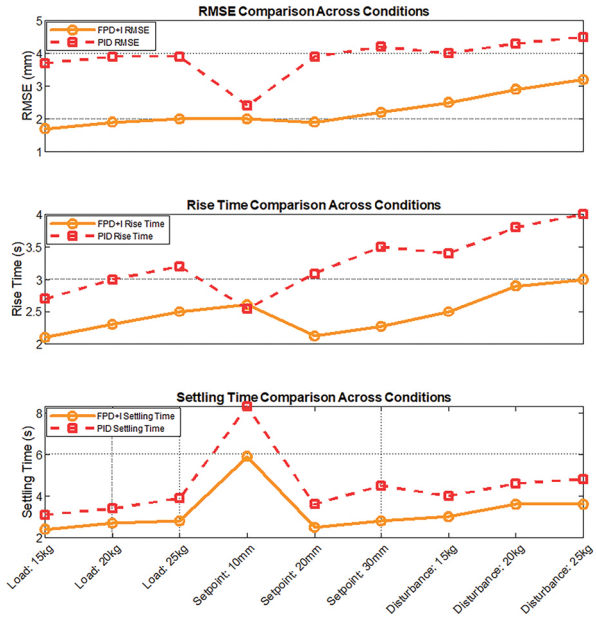


Fig. 11. Performance comparison of FPD+I and PID controllers across varying operating conditions in terms of RMSE, rise time, and settling time

4. CONCLUSIONS

This study investigated the robust position control of pneumatic artificial muscles (PAMs) using a fuzzy PD+I (FPD+I) controller to address the challenges arising from system nonlinearities, hysteresis, and external disturbances. The proposed controller was evaluated through a series of experiments involving various loads, setpoints, and actuated distances. The key results demonstrate that the FPD+I controller significantly outperformed a conventional PID controller, achieving superior transient response characteristics, including reductions of at least 21% in settling time and 22% in rise time and improvements in trajectory tracking accuracy (up to 49% reduction in RMSE). The FPD+I controller consistently maintained strong performance under diverse operating conditions, including continuous and dynamic load disturbances, thereby validating its robustness against uncertainties and external variations. Simultaneously, its ability to sustain high tracking accuracy and fast response across different setpoints, actuated distances, and complex trajectories confirmed its high adaptability without the need for model-based hysteresis compensation or retuning.

To contextualize these results, relevant comparisons with findings from previous studies are discussed. For instance, Phuc et al. (2023) reported steady-state errors of approximately 0.35 mm and settling times around 1 s using PID control under fixed load conditions [27]. In comparison, the present study demonstrated that similar or better levels of accuracy and responsiveness can be achieved without the need for retuning, even under dynamic and varying loads. Other approaches, such as the sliding mode controller used by Lin et al. (2021), showed maximum position errors of about 11.3 mm, whereas the proposed FPD+I controller achieved average RMSE values as low as 3.9 mm [29]. Moreover, while adaptive and fuzzy sliding mode controllers [4, 17, 30] have achieved high tracking accuracy, they often involve complex system modeling and extensive parameter tuning. The FPD+I controller, by contrast, provided competitive performance with a simpler design and implementation. These qualitative comparisons highlight the practical advantages of the proposed method in achieving robust and adaptive position control of PAMs across a wide range of operating conditions.

5. REFERENCES:

[1] M. Chavoshian, M. Taghizadeh, "Recurrent neuro-fuzzy model of pneumatic artificial muscle position", Journal of Mechanical Science and Technology, Vol. 34, No. 1, 2020, pp. 499-508.

[2] A. Festo, "Fluidic Muscle DMSP", 2007.

[3] D. H. Plettenburg, "Pneumatic actuators: a comparison of energy-to-mass ratio's", Proceedings of the 9th International Conference on Rehabilitation Robotics, Chicago, IL, USA, 28 June - 1 July 2005, pp. 545-549.

[4] H. T. Nguyen, V. C. Trinh, T. D. Le, "An Adaptive Fast Terminal Sliding Mode Controller of Exercise-Assisted Robotic Arm for Elbow Joint Rehabilitation Featuring Pneumatic Artificial Muscle Actuator", Actuators, Vol. 9, No. 4, 2020, p. 118.

[5] R. M. Robinson, C. S. Kothera, N. M. Wereley, "Variable Recruitment Testing of Pneumatic Artificial Muscles for Robotic Manipulators", IEEE/ASME Transactions on Mechatronics, Vol. 20, No. 4, 2015, pp. 1642-1652.

[6] A. Merola, D. Colacino, C. Cosentino, F. Amato, "Model-based tracking control design, implementation of embedded digital controller and testing of a biomechatronic device for robotic rehabilitation", Mechatronics, Vol. 52, 2018, pp. 70-77.

[7] S. Hussain, P. K. Jamwal, M. H. Ghayesh, S. Q. Xie, "Assist-as-needed control of an intrinsically compliant robotic gait training orthosis", IEEE Transactions on Industrial Electronics, Vol. 64, No. 2, 2017, pp. 1675-1685.

[8] G. Andrikopoulos, G. Nikolakopoulos, S. Manesis, "A Survey on applications of Pneumatic Artificial Muscles", Proceedings of the 19th Mediterranean Conference on Control & Automation, Corfu, Greece, 20-23 June 2011, pp. 1439-1446.

[9] T.-C. Tsai, M.-H. Chiang, "A Lower Limb Rehabilitation Assistance Training Robot System Driven by an Innovative Pneumatic Artificial Muscle System", Soft Robotics, Vol. 10, No. 1, 2023, pp. 1-16.

[10] S. Jamian et al. "Review on controller design in pneumatic actuator drive system", Telkomnika (Telecommunication, Computing, Electrical & Electronics, and Instrumentation & Control), Vol. 18, No. 1, 2020, pp. 332-342.

[11] B. Tondu, S. Ippolito, J. Guiochet, A. Daidie, "A Seven-degrees-of-freedom robot-arm driven by pneumatic artificial muscles for humanoid robots", The International Journal of Robotics Research, Vol. 24, No. 4, 2005, pp. 257-274.

[12] M. Al Saaideh, M. Al Janaideh, "On Prandtl-Ishlinskii Hysteresis Modeling of a Loaded Pneumatic Artificial Muscle", ASME letters in dynamic systems and control, Vol. 2, No. 3, 2022, pp. 1-12.

[13] S. Shakiba, M. Ourak, E. Vander Poorten, M. Ayati, A. Yousefi-Koma, "Modeling and compensation of asymmetric rate-dependent hysteresis of a miniature pneumatic artificial muscle-based catheter", Mechanical Systems and Signal Processing, Vol. 154, 2021, p. 107532.

[14] S. L. Xie, H. T. Liu, Y. Wang, "A method for the length-pressure hysteresis modeling of pneumatic artificial muscles", Science China Technological Sciences, Vol. 63, No. 5, 2020, pp. 829-837.

[15] L. Hao, H. Yang, Z. Sun, C. Xiang, B. Xue, "Modeling and compensation control of asymmetric hysteresis in a pneumatic artificial muscle", Journal of Intelligent Material Systems and Structures, Vol. 28, No. 19, 2017, pp. 2769-2780.

[16] P. Geng, Y. Qin, L. Zhong, F. Qu, D. Bie, J. Han, "Direct Inverse Hysteresis Compensation of a Pneumatic Artificial Muscles Actuated Delta Mechanism", Proceedings of the 10th Institute of Electrical and Electronics Engineers International Conference on Cyber Technology in Automation, Control, and Intelligent Systems, Xi'an, China, 10-13 October 2020, pp. 18-23.

[17] N. Sun, D. Liang, Y. Wu, Y. Chen, Y. Qin, Y. Fang, "Adaptive Control for Pneumatic Artificial Muscle Systems With Parametric Uncertainties and Unidirectional Input Constraints", IEEE Transactions on Industrial Informatics, Vol. 16, No. 2, 2020, pp. 969-979.

[18] P. A. Laski et al. "Study of the effect of temperature on the positioning accuracy of the pneumatic muscles", EPJ Web of Conferences, Vol. 143, 2017, p. 02065.

[19] T. Tagami, T. Miyazaki, T. Kawase, T. Kanno, K. Kawashima, "Pressure Control of a Pneumatic Artificial Muscle Including Pneumatic Circuit Model", IEEE Access, Vol. 8, 2020, pp. 60526-60538.

[20] Z. Song, L. Zhong, N. Sun, Y. Qin, "Hysteresis Compensation and Tracking Control of Pneumatic Artificial Muscle", Proceedings of the IEEE 9th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems, Suzhou, China, 29 July - 2 August 2019, pp. 1412-1417.

[21] Y. Qin, H. Zhang, X. Wang, J. Han, "Active Model-Based Hysteresis Compensation and Tracking Control of Pneumatic Artificial Muscle", Sensors, Vol. 22, No. 1, 2022, p. 364.

[22] C. Van Kien, N. N. Son, and H. P. H. Anh, "Identification of 2-DOF pneumatic artificial muscle system with multilayer fuzzy logic and differential evolution algorithm", Proceedings of the 12th IEEE Conference on Industrial Electronics and Applications, Siem Reap, Cambodia, 18-20 June 2017, pp. 1264-1269.

[23] N. N. Son, C. Van Kien, H. P. H. Anh, "A novel adaptive feed-forward-PID controller of a SCARA parallel robot using pneumatic artificial muscle actuator based on neural network and modified differential evolution algorithm", Robotics and Autonomous Systems, Vol. 96, 2017, pp. 65-80.

[24] Y. Zhang, Q. Cheng, W. Chen, J. Xiao, L. Hao, Z. Li, "Dynamic modelling and inverse compensation for coupled hysteresis in pneumatic artificial muscle-actuated soft manipulator with variable stiffness", ISA Transactions, Vol. 145, 2024, pp. 468-478.

[25] D. F. Brown, S. Q. Xie, "Model Predictive Control with Optimal Modelling for Pneumatic Artificial

Muscle in Rehabilitation Robotics: Confirmation of Validity Though Preliminary Testing", Biomimetics, Vol. 10, No. 4, 2025, p. 208.

[26] X. Zhang, N. Sun, G. Liu, T. Yang, J. Yang, "Disturbance Preview-Based Output Feedback Predictive Control for Pneumatic Artificial Muscle Robot Systems With Hysteresis Compensation", IEEE/ASME Transactions on Mechatronics, Vol. 29, No. 5, 2024, pp. 3936-3948.

[27] V. P. Tran, T. H. T. Nguyen, H. N. Ngo, M. K. Nguyen, C. N. Nguyen, C. N. Nguyen, "Điều khiển vị trí cơ nhân tạo khí nén sử dụng bộ điều khiển PID [Position control of a pneumatic artificial muscle using a PID controller]", Can Tho University Journal of Science, Vol. 59, 2023, pp. 45-49.

[28] J. Takosoglu, "Angular position control system of pneumatic artificial muscles", Open Engineering, Vol. 10, No. 1, 2020, pp. 681-687.

[29] C.-J. J. Lin, T.-Y. Y. Sie, W.-L. L. Chu, H.-T. T. Yau, C.-H. H. Ding, "Tracking Control of Pneumatic Artificial Muscle-Activated Robot Arm Based on Sliding-Mode Control", Actuators, Vol. 10, No. 3, 2021, p. 66.

[30] T. C. Tsai, M. H. Chiang, "Design and control of a 1-DOF robotic lower-limb system driven by novel single pneumatic artificial muscle", Applied Sciences, Vol. 10, No. 1, 2020, p. 43.

[31] Z. Zhou, Y. Lu, S. Kokubu, P. E. Tortós, W. Yu, "A GAN based PID controller for highly adaptive control of a pneumatic-artificial-muscle driven antagonistic joint", Complex & Intelligent Systems, Vol. 10, No. 5, 2024, pp. 6231-6248.

[32] H. Khajehsaeid, A. Soltani, V. Azimirad, "Design of an Adaptive Fixed-Time Fast Terminal Sliding Mode Controller for Multi-Link Robots Actuated by Pneumatic Artificial Muscles", Biomimetics, Vol. 10, No. 1, 2025.

[33] A. M. El-Nagar, A. Abdrabou, M. El-Bardini, E. A. Elsheikh, "Embedded Fuzzy PD Controller for Robot Manipulator", Proceedings of the International Conference on Electronic Engineering, Menouf, Egypt, 3-4 July 2021, pp. 1-6.

[34] I. N. Al-Obaidi, "Design Of Fuzzy-Pd Controller For Heating System Temperature Control", Bilad Alrafidain Journal for Engineering Science and Technology, Vol. 1, No. 1, 2022, pp. 16-20.

[35] S. Dan, H. Cheng, Y. Zhang, H. Liu, "A Fuzzy Indrect Adaptive Robust Control for Upper Extremity Exoskeleton Driven by Pneumatic Artificial Muscle", Proceedings of the IEEE International Conference on Mechatronics and Automation, Guilin, Guangxi, China, 7-10 August 2022, pp. 839-846.

[36] S. W. Chan, J. H. Lilly, D. W. Repperger, J. E. Berlin, "Fuzzy PD+I learning control for a pneumatic muscle", Proceedings of the 12th IEEE International Conference on Fuzzy Systems, St. Louis, MO, USA, 25-28 May 2003, pp. 278-283.

[37] H. Chaudhary, V. Panwar, N. Sukavanum, R. Prasad, "Fuzzy PD+I based Hybrid force/ position control of an Industrial Robot Manipulator", IFAC Proceedings Volumes, Vol. 47, No. 1, 2014, pp. 429-436.

[38] C.-N. Nguyen, C.-N. Nguyen, "Fuzzy Control: A Textbook", Can Tho: Can Tho University Publishing House, 2020.

Privacy-First Mental Health Solutions: Federated Learning for Depression Detection in Marathi Speech and Text

Original Scientific Paper

Priti Parag Gaikwad*
Dr. D. Y. Patil Institute of Technology, Electronics and Telecommunication Engineering
COEP Technological University, Pune, Maharashtra
pritigaikwad071@gmail.com

Mithra Venkatesan
Dr. D. Y. Patil Institute of Technology, Pimpri, Pune, Maharashtra
mithra.v@dypvp.edu.in

*Corresponding author

Abstract *– Federated Learning (FL) is a cutting-edge approach that allows machines to learn from data without compromising privacy, making it especially valuable in sensitive areas like mental health. This research focuses on using FL to detect depression through speech and text data from a Marathi-speaking population. Depression, a widespread mental health issue, often leaves subtle clues in the way people speak and write, making speech and text analysis a powerful tool for early identification. However, mental health data is highly personal, and protecting it is crucial. FL addresses this by enabling training across multiple devices without ever sharing the raw data. In this study, we introduce a federated learning framework designed specifically to detect depression in Marathi speakers. The framework combines Natural Language Processing (NLP) for analyzing text and audio processing techniques for studying speech patterns. Using a Marathi dataset that includes both speech and text samples from individuals with and without depression, we train local models on individual devices. These models are then combined into a global model, which is continuously improved through a process called federated averaging.*

Our findings show that this FL-based approach performs well in detecting depression while keeping the data private and secure. This highlights the potential of FL in mental health applications, especially for languages like Marathi, where gathering and processing data centrally can be difficult. By prioritizing privacy, this work opens the door for future research into using federated learning for other regional languages and mental health challenges. The Federated Learning model outperforms the non-FL model, achieving around 97.9% across accuracy, precision, recall, and F1-score, compared to 97.4% without FL. with speech dataset the model demonstrates high parameter values of above 96.0%., This demonstrates FL's effectiveness in improving performance on the text and speech based depression detection task.

Keywords: *Federated Learning, Depression Detection, Mental Health, Speech Analysis, Text Analysis, Marathi Dataset, Privacy-Preserving Machine Learning*

Received: March 14, 2025; Received in revised form: May 27, 2025; Accepted: July 13, 2025

1. INTRODUCTION

Depression is a mental illness from which approximately 264 million individuals suffer worldwide. The subjective self-evaluations and clinical interviews are essential parts of conventional diagnostic techniques, though they are not always reliable and accessible. Nowadays, the use of artificial intelligence in mental health screening, especially when using speech and text data, has opened up new path for early detection and intervention. The utilization of personal and behavioral data, however, raises significant privacy concerns, especially when it is collected and managed in centralized systems [1].

Marathi, a language that is widely spoken in western India, has over 83 million native speakers and is still underrepresented in speech-based mental health research, despite improvements in the diagnosis of depression for high-resource languages. Rich emotional and acoustic clues, including pitch change, pauses, and tone shifts, are present in Marathi speech data and can reveal psychological suffering. However, the centralized collection of such data is frequently impeded by sociocultural and privacy limitations. Privacy-sensitive speech analysis [2, 3] is crucial for assessing mental health issues, but using only privacy-preserving features can lead to information loss and data-sharing risks. Decentralized training meth-

ods like Federated Learning (FL) have been explored to reduce privacy concerns, but their accuracy and overhead can be concerns, especially when deployed on mobile devices. No work has been conducted to establish the accuracy and performance overhead of FL for privacy-preserving speech analysis. In this work, we provide a federated learning system that uses convolutional architectures that have been locally trained on client devices for Marathi speech-based depression diagnosis. This method improves both ethical norms and model robustness by maintaining language diversity and auditory nuance while guaranteeing that the user's data never leaves their device. The data privacy gap is closed through the use of Federated Learning. We trained a model using Federated Learning (FL) that maintains user privacy without passing user-specific raw data to the central server.

Similarly, textual data also provides helpful linguistic indicators of depression, including negative sentiment, self-relatable phrases, and a reduction in syntactic complexity [4]. In this study, Marathi, a low-resource language, has been selected. We trained localized text-based models on dispersed clients using the federated learning paradigm, which enables each instance to contribute to a global, privacy-preserving model while adapting to region-specific language patterns. Considering the multilingual and dialectally diverse user base of Indians, this strategy effectively achieves a balance between generalization and personalization.

This study discusses the importance of Federated Learning (FL) in addressing privacy concerns and improving model robustness by training on distributed data without compromising individual privacy. The proposed framework represents a promising advancement in depression detection, offering a robust and privacy-preserving solution that could significantly impact clinical practice and research in the field of mental health [5]. Ensuring decentralized model training without compromising client data privacy. This is the novel combination of deep learning, attention mechanisms, and privacy-preserving. FL offers a robust and scalable solution for accurate depression detection across low-resource datasets. Here the users' data privacy concern is addressed because their private data is never posted and cannot be accessed by the server. Comparing the effectiveness of a model built using centralized and distributed machine learning techniques is another goal of the research [6].

1.1. OBJECTIVES

- To protect user privacy by implementing Federated Learning (FL), ensuring that sensitive personal data remains on users' devices and is not shared with centralized databases.
- To strengthen security against cyber threats by decentralizing data storage and reducing the risk of breaches and unauthorized access.
- To train AI models without sharing raw data by enabling decentralized learning, allowing models to learn from multiple sources while preserving data confidentiality.
- To maintain high model accuracy across different data sources by ensuring effective and reliable learning even in a decentralized and diverse data environment.
- To make data more usable without compromising privacy by allowing organizations to collaborate on AI model improvements without accessing private user information.
- To encourage safe and secure participation by designing a system where users can contribute to model training while ensuring their data remains protected

1.2. CONTRIBUTION

- This study incorporated Federated Learning (FL) to both speech and text data in Marathi language, which has remained unexplored in existing research work. Hence, this addresses a significant gap in the current literature. FL has been used in the past for privacy-preserving depression detection, mostly in high-resource languages such as English.
- This study employed a novel hybrid DepreLex-BERT model to identify depressive language patterns in Marathi text. Additionally, to detect depression in Marathi speech, this study combines a hybrid feature selection algorithm with an attention-driven CNN architecture.
- This Federated Learning model yields impressive results above 96% accuracy, precision, recall, and F1-score. This illustrates FL's powerful capacity to significantly enhance performance on tasks requiring the identification of depression in both text and speech.

The remainder of the article is structured as follows: Section 2 reviews the related work based on existing research. Section 3 outlines the proposed methodology. Section 4 represents the system implementation. Section 5 presents the results and provides a comparative analysis, and Section 6 concludes the article.

2. RELATED WORKS

The literature highlights gaps and shortcomings in centralized databases, including privacy and security concerns, limited datasets, limited training performance, vulnerability to attacks, advanced encryption, communication issues, and integration of multi-view data with federated learning in existing studies. To address these challenges, Federated Learning (FL) is a distributed machine learning approach designed to handle the decentralization of privacy-sensitive personal data. It is a networked machine learning framework that maintains client confidentiality by training on data from multiple clients without exposing it. The process starts with the creation of an initial global model, which

is then shared with each client, followed by updates using local data. This method minimizes the risk of data breaches by preventing access to raw data.

Y. Cui *et al*. [7] have developed a new method to diagnose depression using speech data while maintaining user privacy by employing federated learning (FL), which ensures that sensitive voice recordings remain on users' devices, with only encrypted model updates being shared with a central server. To address challenges like variations in individual speech patterns and the need for efficient communication, they implement secure aggregation to enhance privacy. This method offering stronger privacy protections. However, the study acknowledges challenges in scaling the system for real-world use and highlights the potential benefits of integrating speech analysis with other data types.

B. S. Reddy *et al*. [8] have proposed a multimodal approach to detecting depression by combining both speech and text data. The system analyzes acoustic features such as pitch, tone, and speech rate alongside linguistic aspects like sentiment, word choice, and syntactic patterns, which may indicate depressive symptoms. By integrating these two modalities, the goal is to improve the accuracy and reliability of depression detection. Some challenges in the proposed work are complexity of data preprocessing, the need for large, high-quality datasets, and ensuring that the system remains effective across diverse populations.

J. Li *et al*. [09] have proposed a privacy-focused method for detecting depression through speech analysis using asynchronous federated optimization. This approach improves traditional federated learning by allowing devices to update the model at different times, reducing. The study shows that this method enhances training efficiency and maintains diagnostic accuracy comparable to existing techniques.

L. Zhang *et al*. [10] have developed a model to detect adolescent depression by analyzing both speech and text from interviews. By combining vocal and linguistic features, the system improves accuracy in identifying subtle depression indicators. Machine learning helps recognize patterns, but challenges like data variability and emotional cue interpretation remain. This study highlights the potential of multimodal methods for reliable mental health assessments.

A. B. Vasconcelos *et al*. [11] have explored using FL to detect depression in social media posts while preserving user privacy. By combining FL with Natural Language Processing (NLP), the approach analyzes linguistic markers of depression—such as negative sentiment and writing style changes—without centralizing sensitive data. FL enables collaborative model training across distributed data, ensuring strong detection performance while maintaining privacy.

A. Kim *et al*. [12] have presented an automatic system for detecting depression using speech signals from smartphones and deep Convolutional Neural Networks (CNNs). By analyzing speech patterns such as rate, tone, and intonation, the model identifies acoustic markers linked to depression. Trained on a large dataset, the deep CNN achieves high accuracy in recognizing depressive states. This research highlights the potential of leveraging everyday smartphone data for mental health monitoring, offering a convenient and low-effort approach to mobile health applications.

L. Liu *et al*. [13] have assessed the accuracy of deep learning models in detecting depression through speech samples. By analyzing studies that use deep neural networks, the authors evaluate model performance, highlighting both strengths and limitations. The findings show that models focusing on acoustic features like pitch, intensity, and speech rate achieve high diagnostic accuracy.

J. Ye *et al*. [14] have presented a multimodal approach to detecting depression by combining emotional audio with evaluative text. By analyzing speech for emotional cues like tone and rhythm alongside text-based sentiment and word usage, the system improves accuracy in identifying depressive indicators. Machine learning techniques help uncover patterns that might be missed when using a single data source. Results show that this combined method outperforms single-modality models, emphasizing the advantage of integrating multiple data types for more reliable depression detection.

L. He and C. Cao [15] have explored the use of Convolutional Neural Networks (CNNs) to detect depression through speech analysis. By examining acoustic features like pitch, energy, and speech rate, the CNN model is trained to identify signs of depression with high accuracy. The findings highlight the potential of advanced speech analysis for real-time mental health monitoring and assessment.

D. Low *et al*. [16], have explored how speech analysis can assist in the automated assessment of psychiatric disorders, with a focus on depression. It examines various studies that use speech processing techniques to identify mental health conditions by analyzing vocal traits like prosody, speech rate, and tone. The review highlights the benefits of speech-based assessments, including their non-invasive nature and ease of use, while also addressing challenges such as variations in speech patterns, background noise, and the need for diverse datasets. The authors discuss potential clinical applications and future research directions for speech-based diagnostic tools.

W. Wu *et al*. [17] have introduced a self-supervised learning approach for detecting depression through speech analysis. The method uses unsupervised learning to extract meaningful patterns from large amounts of unlabeled speech data, which can then be fine-tuned for identifying depression. By leveraging self-supervised techniques, the model overcomes the challenge of limited labeled data, a common issue in mental health research. Findings show that this ap-

proach improves detection performance compared to traditional supervised methods, making it a promising tool for mental health assessment.

X. Xu *et al*. [18] have introduced "FedMood," a federated learning framework designed to detect mood disorders like depression using mobile health data. The system analyzes information from activity logs, speech patterns, and physiological signals while maintaining user privacy by keeping data on local devices. Researchers address challenges such as inconsistent data distribution and propose optimization techniques to enhance model performance. The findings highlight FedMood's potential for scalable, privacy-focused mental health monitoring, though the study notes limitations related to dataset diversity and real-world application.

E. L. Campbell *et al*. [19] have demonstrated how multimodal inputs are complementary; let's look at the diagnosis of Major Depressive Disorder (MDD) utilizing both textual and speech modalities. They illustrate the necessity of sophisticated fusion techniques and show how language and auditory information enhance detection capabilities. However, the cross-lingual environment and lack of privacy issues restrict the applicability of their findings in multilingual, real-world situations.

L. Yang *et al*. [20] have introduced feature-augmenting networks to improve the assessment of depression severity through speech analysis. By enhancing acoustic features like prosody and energy with deep neural network representations. The research highlights the benefits of combining traditional speech features with deep learning for more precise detection.

M. Ahmed *et al*. [21] have explored the use of on-device FL used to detect depression by analyzing text from Reddit posts. To protect user privacy, the approach keeps data on individual smartphones while allowing models to be trained collaboratively, challenges remain, including scalability issues and the need for a substantial amount of labeled training data.

Q. Deng *et al*. [22] have introduced a framework for detecting depression by analyzing both speech signals and textual transcripts using a speech-level transformer model with hierarchical attention mechanisms. By leveraging hierarchical attention, the model highlights crucial cues in both modalities, offering insights into how these features contribute to classification. The research underscores the potential of advanced AI for mental health applications while acknowledging challenges like real-world data variability and the need for effective multimodal alignment. W. Wei *et al*. [23] have introduced a federated contrastive learning method (FedCPC) that can also be applied to identifying depression. The approach uses contrastive pre-training to develop meaningful speech representations from decentralized data while maintaining user privacy. By allowing collaborative model training without sharing raw data, FedCPC enhances both accuracy and security. The study emphasizes the potential of combining federated learning with contrastive techniques to create scalable and privacy-focused healthcare solutions.

2.1. GAPS IDENTIFIED

Following are the gaps identified from the analysis.

- Limited integrated frameworks that assess the ethical effects of employing decentralized learning in mental health.
- Decentralized training methods like Federated Learning (FL) have been explored to reduce privacy concerns, but their accuracy and overhead have been concerns.
- Limited work has been conducted to establish the accuracy and performance overhead of decentralized learning for privacy-preserving speech analysis.
- Another problem is the diversity in the language of the content. People love to talk and express their feelings in their mother tongue. Most of the research in natural language processing (NLP) deals with the English language

3. PROPOSED METHODOLOGY

In mental health studies, large amounts of EHR (electronic health records) may expose owners' sensitive information, such as disease history, medical records, and personal details. In our work, we have proposed a model in which federated learning (FL) will be applied to e-health records of mental health patients. Instead of sending their original data to the server, data owners in the FL train the model locally and provide only the generated gradients. In this study, a practical and privacy-preserving FL architecture is proposed that protects the privacy of all EHR owners and is robust to their training dropout. A novel framework of acoustic, linguistic, and channel attention is proposed, which is shown in Fig. 1.

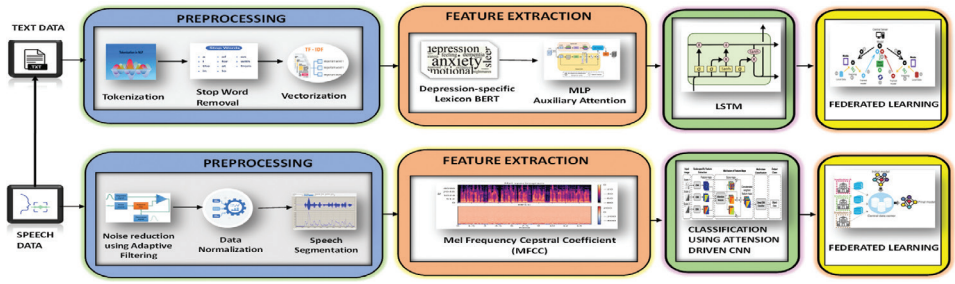


Fig.1. Proposed Methodology

The proposed work focuses on creating a Marathi speech and text dataset. The database contains recordings from various speakers, each differing in gender and age. Ultimately, it is the sentences, rather than individual words, that often convey the emotions present in speech. This speech data is then transformed into a text dataset. Consequently, a key challenge in current research is to enhance emotion recognition accuracy by utilizing Low-Level Descriptors (LLDs) alongside sentence-level features. Traditional methods for speech emotion recognition typically involve three main approaches. Data preprocessing for the speech dataset includes noise reduction, data normalization, and speech segmentation, while the text dataset undergoes tokenization, stop word removal, and vectorization. These processes are essential for analysing the diverse speakers and the emotions they express in their speech. Initially, noise reduction is a crucial step in speech data preprocessing to enhance the clarity of audio signals by minimizing unwanted background noise. In the text procession, the unwanted text is removed, and each word is important for depression-specific words. Text processing comes with several challenges, such as handling noisy data, managing different text formats, resolving ambiguity, and working with large vocabularies to improve language understanding. To tackle these issues, text preprocessing is a vital step in NLP, as it cleans and structures raw text so that machine learning models can analyse it effectively. One common technique used in this process is TF-IDF, which measures word importance by considering both how often a term appears and its relevance across documents. This helps filter out stop words, allowing models to focus on meaningful content. Additionally, FastText-based vectorization represents words as character n-grams, mapping them into high-dimensional vectors. This approach captures sub-word information, making it useful for understanding semantic similarities. Before being processed by an artificial neural network, each word must be converted into a single vector to ensure consistency and enhance model accuracy. Following the initial pre-processing stage, many feature sets were extracted for every audio recording and text dataset together with their corresponding statistical measurements. The spectral characteristics associated with the spectral centroid, MFCCs, which are cepstral characteristics linked to cepstrum analysis, and prosaic components. The initial harmonious signal, the abecedarian frequency, and these features represent the unique features of voice that represent the depression. In the text database, the depression-specific lexicons are extracted from Depression-Specific Lexicons-BERT (DepreLex-BERT), which combines depression-specific lexicons, N-grams, and BERT algorithms. Extensive research has explored word embedding techniques using statistical methods to derive meaningful word representations from text corpora. More recently, deep learning algorithms based on transformers, such as BERT, have been widely adopted for word embedding. This approach significantly enhances the performance of various natural language processing tasks by preserving nuanced meanings and improving contextual understanding. The next block is classification to support the diagnosis of depression; it is therefore necessary to develop depression classification methods. Acoustic feature extraction utilizes a 1D CNN to capture temporal patterns in sound signals, while linguistic feature extraction relies on an LSTM layer to process sequential language data. To improve performance, an attention mechanism is integrated to highlight the most relevant parts of the input, along with a channel attention mechanism that identifies interactions across CNN channels. Finally, the model's performance is evaluated through parameter validation, where metrics such as balanced accuracy, precision, recall, and F1 score are computed using the confusion matrix. The literature identifies several limitations of centralized databases, including privacy and security risks. To address these issues, Federated Learning (FL) has emerged as a distributed machine learning approach designed to handle the decentralization of privacy-sensitive personal data. FL operates within a networked architecture that safeguards client confidentiality by enabling model training across multiple clients without exposing their raw data. A networked machine learning system called federated learning (FL) enables training on data from various clients without jeopardizing client confidentiality. FL's key component is preserving privacy because data collection is unnecessary. The recurrence of the following processes usually results in FL progressing. An initial global model is created and distributed to each client via a centralized server. With local data, each client trains a copy of the model. The model's local modifications are transmitted from each client to the server. The server collects the updates from the clients, updates the global model, and then sends the revised global model back to the clients. These processes are repeated until the model converges or a stopping requirement is satisfied. This reduces the attack surface because training is carried out solely through local model updates without access to raw data.

3.1. FEDERATED LEARNING

We are unable to use patient health data kept in hospitals for centralized learning due to privacy concerns. Federated learning is a new framework that Google has suggested. Each hospital involved in the model training cooperation may keep its own data locally during the federated learning model's training process, eliminating the need for uploading. Every hospital downloads the model from the server for training using its own data. It then uploads the trained model or gradient to the server for aggregation. Finally, the server notifies each hospital of the aggregated model or gradient information. We use the model average approach for training, taking into account the load on communication, connection dependability, and other factors. When the global model parameters are updated in the t round, assuming that K hospitals are involved in federated learning, the K-th participant uses n (1) to cal-

culate the local data average gradient of the current model parameters. The server then aggregates these gradients and uses the updated model parameters to update the global model using equation (2).

$$g_k = \nabla F_k(\omega_t) \quad (1)$$

$$\omega_{t+1} \leftarrow \omega_t - \eta \sum_{k=1}^{K} \frac{n_k}{n} g_k \quad (2)$$

where g_k is the local data's average gradient for the current model parameter, ω_t. The learning rate is denoted by η, where $\sum_{k=1}^{K} \frac{n_k}{n} g_k = \nabla f(\omega_t)$.

Every hospital performs one or more gradient descent steps using local data, updates the model parameters locally in accordance with equation (3), and transmits the modified local model parameters to the server. The server then provides the aggregated model parameters to each hospital after computing the weighted average of the model results using equation (4). The enhanced system may select an alternate gradient update optimizer in addition to SGD and can minimise communication requirements by 10-100 times when compared to the simply distributed SGD.

$$\forall k, \omega_{t+1}^{(k)} \leftarrow \overline{\omega_t} - \eta g_k \,, \quad (3)$$

$$\overline{\omega}_{t+1} \leftarrow \sum_{k=1}^{K} \omega_{t+1}^{(k)}, \quad (4)$$

where $\bar{\omega}_t$ is the existing model parameter of the local client.

The literature identifies several gaps and challenges associated with centralized databases, such as concerns around privacy and security, limitations in dataset availability, restricted training performance, vulnerability to attacks, communication inefficiencies, and issues with integrating multi-view data. Additionally, advanced encryption techniques and privacy safeguards often struggle to fully address these challenges. To tackle these limitations, Federated Learning (FL) has emerged as a promising distributed machine learning paradigm. FL enables decentralized processing of privacy-sensitive data by training models directly on clients' devices without transferring raw data to a central server. In this approach, an initial global model is created and distributed to multiple clients. Each client then updates the model using its local data, after which these updates are aggregated to improve the global model. Local and global training are the primary components of the asynchronous federated learning framework. Every device updates its local model during local training using the global model that the server sends. When any device uploads parameters, the server instantly updates the global model. In contrast to the federated average algorithm, noise is introduced before global model updating to prevent parameter leakage. Take federated learning with clients (devices) into consideration. is the local data on the ith device in a horizontal federated learning system. A certain instance is from the ith device. The ultimate goal is to use the distributed local models from every device to train a global model. To achieve this, perceive Eq. (5) as the optimization's ultimate objective.

$$\min F(w) = min\left(\frac{1}{n}\sum_{i\in[n]}^{n} E_{Z^i \sim D^i}(W;\, Z^i)\right) \quad (5)$$

The basic idea behind federated optimization is that there are global epochs, and worker i sends a local model with new information to the server using eqs. (6) and (7).

$$G_t = W_{back}^i - W_{new}^i \quad (6)$$

$$W_{t+1}^k = W_t^k - \partial_k g_t^k \quad (7)$$

Gradients cannot be uploaded straight to the server because the majority of devices are edge devices. Extract the file after uploading the device's most recent model to the server. To enable the server to determine the gradient that the ith device updated, save each model that the ith device uploads to the server in. Is the model on the other device prior to the new one being updated? To prevent the negative effects of the direct upload gradient on the server, use Eq. (6) to indirectly calculate the gradient of the device due to the unreliable connection and poor communication efficiency of edge devices. Following the server's receipt of the local model, equation (7) shows the updated global model. This decentralized process enhances data security by minimizing exposure to potential breaches and attacks, as raw data remains on individual devices throughout the learning process.

Algorithm1:
Federated Learning for Speech and Text Data

Input: Global model parameters W_o, number of global communication rounds T, learning rate η, number of devices n, local dataset D^i on each device i.

Output: Optimized global model W_T

Initialize Global Model:

$Wo \leftarrow$ Random Initialization

For each global round $t \in [1,T]$:

Send current global model W_t to all participating devices {1,2,…,n}.

For each device i in parallel

Download global model W_t.

Perform local training using the speech/text data D^i for E epochs:

$$W_{t+1}^i \leftarrow W_t - \eta \nabla F_i(W_t; D^i)$$

Calculate the local model update:

$$G_t^i = W_t^i - W_{t+1}^i$$

Apply differential privacy (e.g., Gaussian noise) if necessary:

$$\tilde{G}_t^i = G_t^i + noise(\mu, \sigma)$$

Send $\tilde{G}_t^i$ to the server.

Server Aggregation:

Update the global model using Federated Averaging (FedAvg):

$$W_{t+1} = W_t - \eta \sum_{i=1}^{n} \frac{|D^i|}{\sum_{j=1}^{n} |D^i|} \tilde{G}_t^i$$

Repeat Steps 2a to 2c until convergence or maximum rounds T are reached.

Return Optimized Global Model: W_T

4. RESULTS AND DISCUSSION

The following section provides a comparison of baseline models, along with detailed descriptions of the evaluation metrics, outcomes, and datasets. Each subsection delves into these components to offer a comprehensive understanding of the results. The proposed deep learning-based method is evaluated against a benchmark filter using simulations, data preprocessing, and Python 3 libraries, including NumPy, pandas, seaborn, and Scikit-learn. The model is developed with TensorFlow 2.10 and the Keras library.

The dataset's training and testing versions were divided, with 80% going toward training and 20% toward testing. Considering their suitability for classifying the depression for 50 epochs, the technique was trained using the AdamW optimizer with verbose 2, batch size 16, a learning rate of 0.001, and a cross-entropy loss function.

4.1. RESULT OBTAINED FROM THE PROPOSED METHOD

4.1.1. Dataset Description and Transcription

The dataset was collected from 54 non-professional Marathi speakers (both genders) using a microphone at a consistent distance, recorded in a clinic in Pune, Maharashtra, under the supervision of a psychologist. The Beck Depression Inventory (BDI), a widely used self-report tool with 21 items assessing depression symptoms and severity, was used to validate the dataset. The BDI assesses emotional, cognitive, and physical symptoms of depression, scoring each item from 0 to 4 to determine overall depression severity for individuals aged 13 and older. More severe depressed symptoms are indicated by higher overall scores. [24]

The speech Marathi dataset is transcribed into the text Marathi dataset using manual transcription to maintain the 99% accuracy of the dataset. The Marathi speech audio files of 54 individuals are transcribed into the Marathi text dataset and used while implementing the algorithm. All the datasets are labelled according to the BDI tool score and stored in the different folders according to their labelled class for normal, mood disturbance, slightly depressed, moderately depressed, and depressed in different class folders 0, 1, 2, 3, and 4, respectively.

This section presents a comparison of the proposed methodology, which is evaluated both with and without the federated learning approach. The aim is to determine which method performs better based on various parameters. We analyse the effectiveness of the proposed method against current techniques, using metrics such as Accuracy, Precision, Recall, and F1-score. The figure illustrates a comprehensive comparison of the proposed procedures in the context of Centralized and Decentralized Learning.

4.2 PERFORMANCE EVALUATION RESULTS

The section presents the performance evaluation of the model's training and validation results, both with decentralized (Federated Learning) and Centralized (without Federated Learning), using a confusion matrix and ROC curve for assessment.

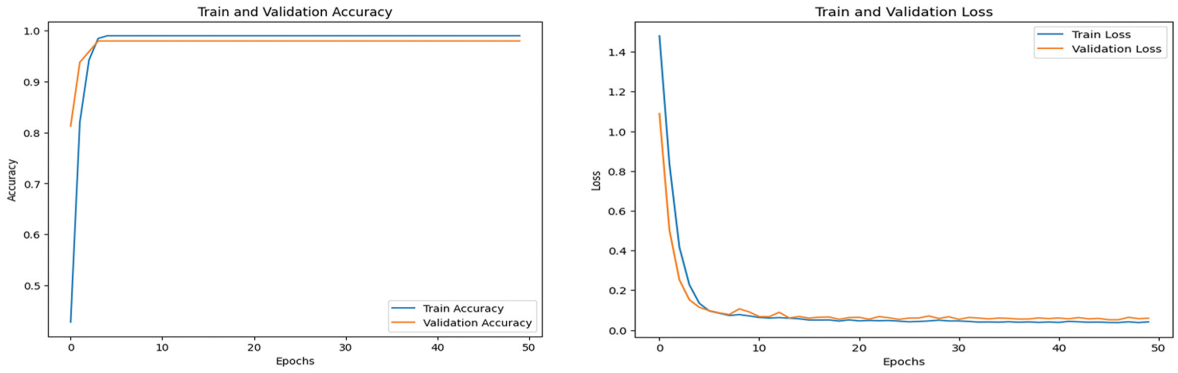


Fig. 2. Accuracy- Loss for training and validation with decentralized Learning for text Moel

Fig. 2 illustrates the training and validation accuracy over 50 epochs for the depression detection text model with federated learning applied. The training accuracy increases quickly, converging at around 99% by the 10th epoch, while the validation accuracy stabilizes close to 98.5%, reflecting strong learning performance with minimal overfitting. Simultaneously, the loss plot shows a sharp decrease in training loss, reaching approximately 0.02 by epoch 10, with the validation loss levelling off around 0.05. Both accuracy and loss exhibit early improvements, indicating rapid convergence and effective optimization using federated learning.

Fig. 3 illustrates the training and validation accuracy over 50 epochs for the depression detection. Speech model with decentralized learning: the training accuracy rises sharply to nearly 98% within the first 5 epochs and reaches 100% after 10 epochs. However, the validation accuracy levels off at approximately 80%, revealing a significant gap between the training and validation performance, suggesting that the model may be overfitting to the training data. The training loss decreases significantly, dropping to around 0.1 by epoch 10, whereas the validation loss exhibits a different behavior, increasing after initially reaching its lowest point of 0.8. This divergence between the training and validation losses further supports the indication of overfitting, where the model performs well on data in the training set but struggles to generalize to unseen data.

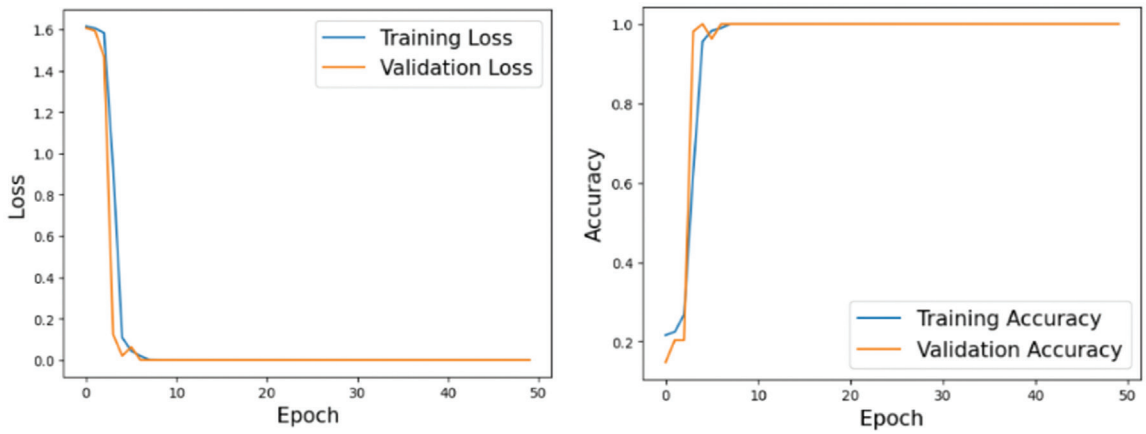
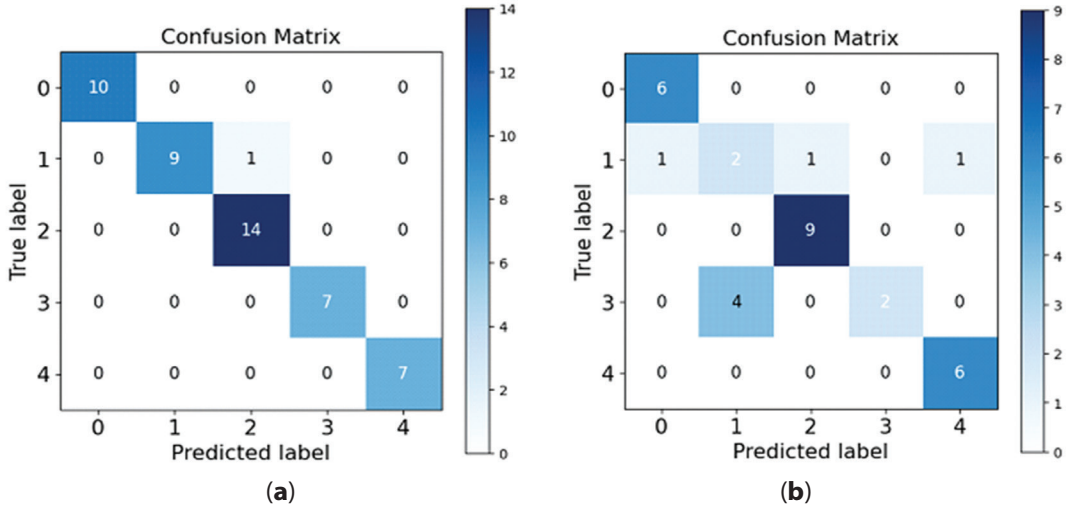


Fig. 3. Accuracy- Loss for training and validation with decentralized learning for Speech Model

Fig. 4 illustrates the confusion matrix for the text model with centralized and decentralized learning and demonstrates strong performance across all classes, with Class 0 correctly predicted 10 times and no misclassifications, Class 1 accurately classified 9 times with only 1 misclassification into Class 2, Class 2 having 14 correct predictions and no errors, and Classes 3 and 4 both showing 7 correct classifications with no misclassifications. This absence of significant errors indicates high accuracy and consistency, suggesting the model generalizes well to unseen data. In contrast, the confusion matrix for the model without federated learning reveals more misclassifications. Class 0 is correctly predicted 6 times but has 1 misclassification into Class 1, while Class 1 has 2 correct predictions with misclassifications into Classes 0, 2, and 4. Class 2 maintains good accuracy with 9 correct predictions, but Class 3 has increased errors, with only 2 correct predictions and 4 misclassifications into Class 4. Similarly, Class 4 is correctly predicted 6 times.

The FL model's test evaluation combines predictions from several customers, each of whom contributes a distinct number of samples per class. In contrast, the centralized (non-FL) model is evaluated using a unified, potentially more balanced dataset.

The higher number of errors, especially in Classes 0, 1, and 3, indicates that the model without Federated Learning struggles more with generalization, leading to more prediction inaccuracies compared to the Federated Learning approach.

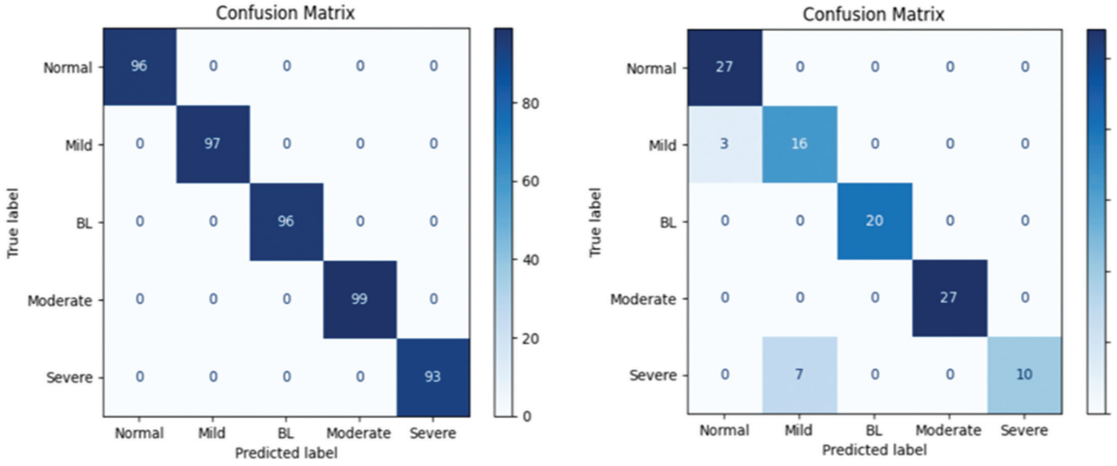


(**a**) (**b**)

Fig. 4. Confusion matrix for the text model with Centralized and Decentralized Learning. (**a**) with FL, (**b**) without FL

The higher number of errors, especially in Classes 0, 1, and 3, indicates that the model without Federated Learning struggles more with generalization, leading to more prediction inaccuracies compared to the Federated Learning approach.

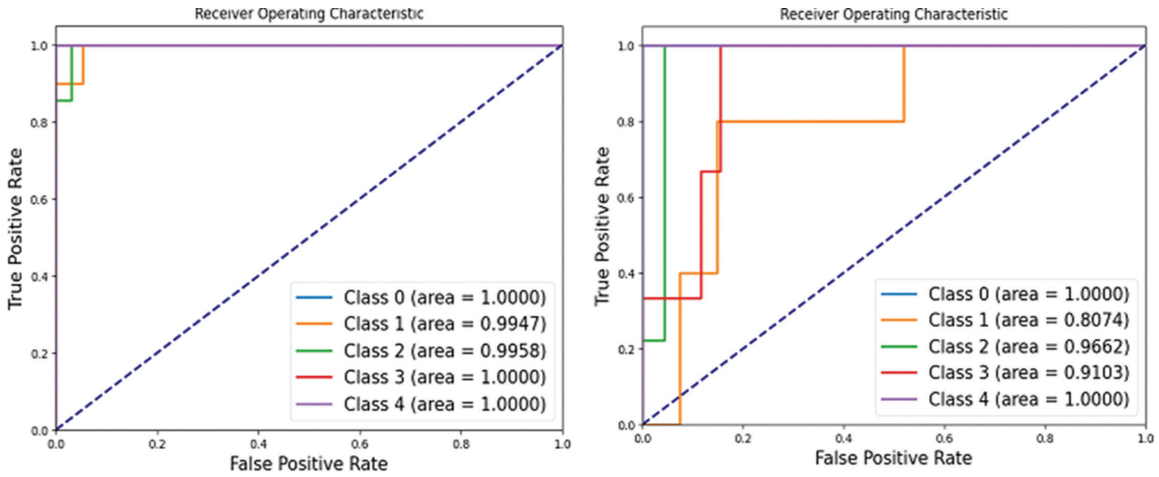
Fig.5 compares the decentralized approach and the centralized approach. shows the confusion The decentralized matrix shows the performance of a classification model for speech across five classes: normal, mild, BL, moderate, and severe. The matrix indicates perfect classification, with all instances correctly predicted for each class (e.g., 96 Normal, 97 Mild, 96 BL, 99 Moderate, and 93 Severe). There are no misclassifications, suggesting 100% accuracy. While centralized, the confusion matrix evaluates a classification model's performance across five classes: Normal, Mild, BL, Moderate, and Severe. The model correctly predicted 27 normal, 16 mild, 20 BL, 27 moderate, and 10 severe instances. However, there are misclassifications: 3 mild instances were incorrectly predicted as normal, and 7 severe instances were misclassified as mild. While the model performs well for normal, BL, and moderate classes, the errors in mild and severe classes indicate room for improvement, particularly in distinguishing between mild and severe cases. Instead of being centrally pooled, data in federated learning is dispersed among several clients, each of which represents a distinct local dataset. The total class distribution in the test phase may therefore be different from that in the centralized (non-FL) assessment, where the dataset is uniformly available, when combining predictions from all clients in the FL configuration. These differences in class-wise sample counts reflect the variability introduced by FL's decentralized nature and demonstrate the model's performance under realistic, heterogeneous conditions.



(**a**) (**b**)

Fig. 5. Confusion matrix for the speech model with Centralized and Decentralized Learning (**a**) with FL, (**b**) without FL

Fig. 6 illustrates the ROC curve for the text model with centralized and decentralized learning.



(**a**) (**b**)

Fig. 6. ROC curve for the model with Centralized and Decentralized Learning. (**a**) with FL, (**b**) without FL

With decentralized learning, the model demonstrates near-perfect classification performance across all classes. The Area Under the Curve (AUC) values are exceptionally high, with Class 0, Class 3, and Class 4 achieving a flawless AUC of 1.0000, indicating perfect classification. Class 1 reaches an AUC of 0.9947, while Class 2, though slightly lower, still achieves a highly accurate AUC of 0.9958. These high values reflect the model's strong generalization ability and superior predictive performance with federated learning, minimizing false positives and ensuring high accuracy across all classes. Conversely, without federated learning, the ROC curve reveals a decline in classification performance. Although Class 0 and Class 4 retain perfect AUC scores of 1.0000, other classes exhibit noticeable drops in performance. Class 1 shows an AUC of 0.8074, a significant reduction compared to the FL model, highlighting the model's difficulty in correctly classifying instances of this class. Class 2 and Class 3 also experience lower AUC values of 0.9662 and 0.9103, respectively, compared to their Federated Learning counterparts. The increase in false positives and lower AUC values for several classes indicate that the model without federated learning struggles to generalize to unseen data and introduces more classification errors.

4.4. COMPARATIVE ANALYSIS

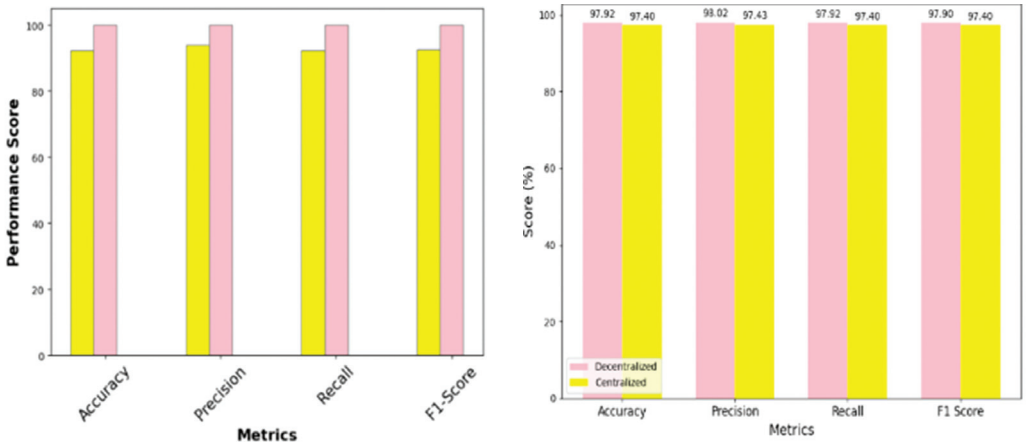
To compare the performance metrics—such as accuracy, precision, recall, and F1-Score—of the proposed technique with existing methods, a dataset comparison is conducted, along with an evaluation of the performance metrics both with and without FL.

Fig. 7 a) and Table 1 illustrates the comparison of performance metrics—accuracy, precision, recall, and F1-score—between the model with FL shows significantly better performance across all metrics. Accuracy reaches a near-perfect score of 0.979 with FL, while the model without FL drops to around **.9740**. Precision also improves with FL, scoring approximately 0.98 compared to 0.9773 without FL. Similarly, recall is higher with FL, achieving about 0.979, while without FL it falls to 0.974. The F1-score follows the same trend, reaching 0.979 with FL, in contrast to 0.974 without FL. Overall, these results demonstrate the enhanced generalization and predictive capabilities of the model when FL is applied.

Fig. 7 b) and Table 1 shows a comparison of the depression lexicon-based BERT with the Attention LSTM model using Federated Learning (FL) and without FL. The proposed method with federated learning, evaluated through various parameters such as Accuracy, Precision, Recall, and F1-score, which yield different values. When comparing the proposed method with federated learning, the model demonstrates high parameter values of around 97.90%.

Table 1. Comparison of Performance Metrics for Decentralized and Centralized approach on Text and Speech Datasets

Metrics	Text Dataset		Speech Dataset	
	With Decentralized	With Centralized	With Decentralized	With Centralized
Accuracy (%)	97.92	97.40	97.20	97
Precision (%)	98.06	97.73	96.40	96
Recall (%)	97.92	97.40	97.20	96
F1-Score (%)	97. 90	97.40	96.20	96



(a) (b)

Fig. 7. Performance comparison of the model with and without Federated Learning (FL). (**a**) Speech Model, (**b**) Text Model

The model consistently performs well across both modalities, according to a comparison of the performance metrics for the depression lexicon-based BERT with the Attention LSTM model using FL for both text and speech datasets as shown in Table 1.

The model achieves 97.92% accuracy, 98.06% precision, 97.92% recall, and 97.90% F1-score for the text dataset. In contrast, the model performs somewhat worse but still well on the speech dataset, achieving 97% accuracy, 96% precision, 97% recall, and 96% F1-score.

With slightly higher metrics in the text dataset than in the speech dataset, these findings demonstrate the model's resilience and capacity for generalization across various data types.

Table 2. Comparison of Decentralized approach with existing work

Ref No	Accuracy	Precision	Recall	F1 Factor
[19] FedAVg	0.65	0.69	0.46	0.47
[20] FEDCPC	84.29	79.5	74.7	75.3
[17] FL	0.6588	0.5037	0.435	
[4] Text CNN	73.33	81.82	81.82	81.82
Proposed (Text)	97.92	98.06	97.92	97.90
Proposed (Speech)	97.20	96.40	97.20	96.20

Table 2 shows comparison of decentralized approach(proposed) with existing work. The Proposed model stands out as the best-performing approach, achieving over 90% in accuracy, precision, recall, and F1 score, significantly outperforming existing methods in federated learning. Traditional techniques like FedAvg (Ref. [19]) struggle with imbalanced data, resulting in lower accuracy (0.65) and recall (0.46), making it less reliable. Similarly, General FL (Ref. [17]) faces similar challenges, with an accuracy of 0.6588 and the lowest recall (0.435), indicating difficulties in capturing diverse patterns across distributed datasets. More advanced techniques, such as FEDCPC (Ref. [20]), introduce contrastive predictive coding to enhance learning, improving accuracy (84.29%) and F1 score (75.3%). While it performs better than basic federated learning models, it still falls short compared to your model. Text CNN (Ref. [4]), a deep learning-based approach for text analysis, achieves high recall and precision (81.82%) but has a lower overall accuracy (73.33%), suggesting that while it captures patterns well, it may struggle with generalization.

In contrast, the proposed model achieves a perfect balance across all metrics, making it both highly accurate and reliable. This significant improvement could be due to better model design, optimized learning techniques, or improved data handling. The results show that the proposed approach not only addresses the weaknesses of existing federated learning models but also sets a new benchmark for privacy-preserving, high-performance AI in distributed environments

5. CONCLUSION

Federated Learning (FL) is transforming mental health research through the ability to identify disorders like depression without compromising the privacy and security of personal information of individuals. This work has investigated the use of FL for text and speech data analysis from Marathi speakers to identify symptoms of depression. FL models yield promising results without sacrificing privacy by teaching models separately on various devices and combining their knowledge without sharing raw data. It becomes especially important for languages such as Marathi, where centralized data collection may become problematic. Our work shows how technology can be made flexible to both help and safeguard the privacy of various groups. The work emphasizes that FL can increase the accessibility and quality of mental health services, particularly in poor resourced communities like Marathi. The results show a precision rate of 97.92%, recall of 97.90%, and F1-score 97.9 for the text dataset. For the speech dataset, the model obtains similar high-performance levels, over 96%, meaning that the suggested model performs better than current models while maintaining security and privacy to clients. Thus, usage of Federated Learning goes beyond technological innovation; it aims towards a world where mental health interventions work and uphold every person's rights. Prioritizing privacy, such tools can be created enabling people feel safe while using, which will ultimately enable more individuals to seek assistance that they need.

ACKNOWLEDGMENT

We would like to sincerely thank for the guidance given by Shivnarayan Khandre – Founder & Director, Divine Life Counselling Clinic, Pune, Clinical Psychologist, Dip. Counselling Psychology, NLP Practitioner, Hypnotist (U.S.A.), whose expertise enabled us to collect the Marathi Speech and Text database that was used in this work.

6. REFERENCES

[1] S. Bn, S. Abdullah, "Privacy sensitive speech analysis using federated learning to assess depression", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Singapore, 23-27 May 2022, pp. 6272-6276.

[2] L. Campanile, M. Iacono, F. Marulli, M. Mastroianni, "Privacy regulations challenges on data-centric and IoT systems: A case study for smart vehicles", Proceedings of the 5th International Conference on Internet of Things, Big Data and Security, Vol. 1, May 2020, pp. 507-518.

[3] L. Campanile, M. Iacono, A. H. Levis, F. Marulli, M. Mastroianni, "Privacy regulations, smart roads, blockchain, and liability insurance: Putting technologies to work", IEEE Security & Privacy, Vol. 19, No. 1, 2020, pp. 34-43.

[4] A. H. Orabi, P. Buddhitha, M. H. Orabi, D. Inkpen, "Deep learning for depression detection of Twitter users", Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, New Orleans, LA, USA, June 2018, pp. 88-97.

[5] P. M. Mammen, "Federated learning: Opportunities and challenges", arXiv:2101.05428, 2021.

[6] X. Xu, H. Peng, M. Z. A. Bhuiyan, Z. Hao, L. Liu, L. Sun, L. He, "Privacy-preserving federated depression detection from multisource mobile health data", IEEE Transactions on Industrial Informatics, Vol. 18, No. 7, 2021, pp. 4788-4797.

[7] Y. Cui, Z. Li, L. Liu, J. Zhang, J. Liu, "Privacy-preserving Speech-based Depression Diagnosis via Federated Learning", Proceedings of the 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, Glasgow, Scotland, UK, 11-15 July 2022, pp. 1-5.

[8] B. S. Reddy, K. Jishitha, V. Akshaya, B. Bhavani, P. Aishwarya, "Development of a Depression Detection System using Speech and Text Data", Proceedings of the 14th International Conference on Computing Communication and Networking Technologies, Delhi, India, 6-8 July 2023, pp. 1-5.

[9] J. Li, R. Zhang, M. Cen, X. Wang, M. Jiang, "Depression Detection Using Asynchronous Federated Optimization", Proceedings of the IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications, Shenyang, China, 20-22 October 2021, pp. 1-5.

[10] L. Zhang, Y. Fan, J. Jiang, Y. Li, W. Zhang, "Adolescent Depression Detection Model Based on Multimodal Data of Interview Audio and Text", International Journal of Neural Systems, Vol. 32, No. 1, 2022.

[11] A. B. Vasconcelos, L. Drummond, R. Brum, A. Paes, "Exploring Federated Learning to Trace Depression in Social Media with Language Models", Proceedings of the International Symposium on Computer Architecture and High Performance Computing Workshops, Porto Alegre, Brazil, 17-20 October 2023, pp. 1-5.

[12] A. Kim, E. Jang, S.-H. Lee, K.-Y. Choi, J. Park, H.-C. Shin, "Automatic Depression Detection Using Smartphone-Based Text-Dependent Speech Signals: Deep Convolutional Neural Network Approach", Journal of Medical Internet Research, Vol. 23, No. 8, 2021, p. e34474.

[13] L. Liu, L. Liu, H. A. Wafa, F. Tydeman, W. Xie, Y. Wang, "Diagnostic accuracy of deep learning using speech samples in depression: A systematic review and meta-analysis", Journal of the American Medical Informatics Association: JAMIA, Vol. 31, No. 1, 2024, pp. 1-10.

[14] J. Ye, Y. Yu, Q. Wang, W. Li, H. Liang, Y. Zheng, G. Fu, "Multi-modal depression detection based on emotional audio and evaluation text", Journal of Affective Disorders, Vol. 290, 2021, pp. 99-105.

[15] L. He, C. Cao, "Automated depression analysis using convolutional neural networks from speech", Journal of Biomedical Informatics, Vol. 87, 2018, pp. 1-10.

[16] D. Low, K. Bentley, S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review", Laryngoscope Investigative Otolaryngology, Vol. 4, No. 1, 2019, pp. 1-10.

[17] W. Wu, C. Zhang, P. Woodland, "Self-Supervised Representations in Speech-Based Depression Detection", Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4-10 June 2023, pp. 1-5.

[18] X. Xu, H. Peng, L. Sun, M. Z. A. Bhuiyan, L. Liu, L. He, "FedMood: Federated Learning on Mobile Health Data for Mood Detection", arXiv:2102.09342, 2021.

[19] E. L. Campbell, L. Docío Fernández, N. Cummins, C. García-Mateo, "Speech and Text Processing for Major Depressive Disorder Detection", Proceedings of IberSPEECH, Granda, Spain, 14-16 November 2022.

[20] L. Yang, D. Jiang, H. Sahli, "Feature Augmenting Networks for Improving Depression Severity Estimation From Speech Signals", IEEE Access, Vol. 8, 2020, pp. 24033-24045.

[21] M. Ahmed, A. Muntakim, N. Tabassum, M. A. Rahim, F. M. Shah, "On-device Federated Learning in Smartphones for Detecting Depression from Reddit Posts", arXiv:2410.13709, 2024.

[22] Q. Deng, S. Luz, S. de la Fuente Garcia, "Hierarchical attention interpretation: an interpretable speech-level transformer for bi-modal depression detection", arXiv:2309.13476, 2023.

[23] W. Wei, Z. Yang, Y. Gao, J. Li, C. Chu, S. Okada, S. Li, "FedCPC: An Effective Federated Contrastive Learning Method for Privacy Preserving Early-Stage Alzheimer's Speech Detection", Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop, Taipei, Taiwan, 16-20 December 2023, pp. 1-5.

[24] P. P. Gaikwad, M. Venkatesan, "KWHO-CNN: A Hybrid Metaheuristic Algorithm Based Optimized Attention-Driven CNN for Automatic Clinical Depression Recognition", International Journal of Computational and Experimental Science and Engineering, Vol. 10, No. 3, 2024.

Offloading of mobile – cloud computing based on Lion Optimization Algorithm (LOA)

Original Scientific Paper

Ammar Mohammed Khudhaier*
Mustansiriya University, College of Science, Department of Computer Science
Baghdad, Iraq
promail2010@uomustansiriyah.edu.iq

Jamal Nasir Hasoon
Mustansiriya University, College of Science, Department of Computer Science
Baghdad, Iraq
jamal.hasoon@uomustansiriyah.edu.iq

*Corresponding author

Abstract *– Many problems and limitations appeared with the spread of mobile cloud computing (MCC) technology, due to the battery life of the mobile, limited resources, storage space and processors. In addition, the biggest challenge is to make decisions about executing tasks between local devices and cloud servers (edge) to save energy and reduce delay time by reducing the energy consumption (or cost) of the system concerned. In this paper, an optimization method based on the Lion Optimization Algorithm (**LOA**) is proposed for task- offloading of mobile cloud computing. Task- offloading is the transfer of tasks from mobile devices to servers that handle high-level computational operations, such as cloud servers or Mobile-Edge Computing (MEC) servers, to reduce task execution time and energy consumption. The proposed algorithm has a high performance in reducing the cost for offloading tasks. This method can be used to improve the performance of 5G, 6G and later mobile technologies to implement applications that require a large amount of computing, also, multi-objectives could be applied for more cost reduction. The results showed that the proposed algorithm **LOA** is better than other traditional algorithms such as the genetic algorithm in terms of calculating the costs to find the best solutions.*

Keywords*: task offloading, cloud servers, mobile-edge computing, mobile-cloud computing*

Received: April 26, 2025; Received in revised form: N/A; Accepted: July 14, 2025

1. INTRODUCTION

Due to rapid progress and development of communication and internet technology, and the widespread adoption of scalable, fault-tolerant, and highly available cloud computing technology led to the emergence of diverse applications used by many users across millions of edge devices which produces a big data [1], therefore request resources allocation and services in cloud computing with lack of coordination in resource utilization causes delays in internet services [2, 3]. Therefore, it is necessary to distribute cloud computing resources equally, ensuring high performance for internet services. This is achieved by load balancing on the all-cloud computing servers without burdening any server, improving resource utilization and reducing response time [4, 5]. Load balancing is a major challenge, requiring the use of unique hardware and software that consume high energy, leading to an increase in operating cost, as well as handling numerous applications that require rapid execution [6, 7].

The traditional methods used to improve performance do not provide the most appropriate solution to these devices and applications, therefore, over the past few years, researchers have developed many meta-heuristic algorithms to solve these problems [8], to reach algorithms with satisfactory performance, including algorithms inspired by nature [9]. An example on these algorithms, the algorithm inspired by improving the colony of ants searching for food [10] and the particle swarm optimization algorithm that simulates the social behavior of migratory birds searching for an unknown destination [11] and the bacterial food search algorithm that simulates the search for the best food for bacteria [12] and other algorithms. However, after analyzing these algorithms, one of the defects was revealed, which is in the rate of convergence of explorations or independence [5]. In 2012, Rajakumar and Ramakrishnan and Sankara Gomathi 2017 were able to mathematically design and formulate the Lion Optimization Algorithm (LOA) [13], which is a class of

powerful inferential models inspired by the solitary and cooperative behaviors of lions that differ from other algorithms, which has proven high performance in implementation [14]. The working principle of the lion optimization algorithm is based on defense and territory reserve with the aim of finding and replacing the worst solutions with the best ones, where the strongest pride lion shows its dominance compared to the territorial lion that will migrate or die [15]. The strongest pride lion represents the global minimum solution, and the territorial lion represents the local minimum solution. A group of lionesses hunts the prey by surrounding it from several axes and quickly attacks it through organized group hunting to succeed in hunting and continue life. Lions show other behaviors such as the method of hunting alone, placing territorial marks, and the difference between residents and nomad lions, as well as migration [9, 14].

2. RELATED WORK

Since the mobile cloud computing system has been used, many researchers have worked on finding optimal solutions to offload tasks to the cloud by reducing the energy consumed by the cloud infrastructure and maximizing resources by reducing the processing time of user tasks [16-18]. Accordingly, many nature-simulation algorithms have been proposed, including the proposed lion optimization algorithm, which is one of the Swarm Intelligence (SI) algorithms to solving optimization problems [19]. It has been employed in this paper for the process of tasks offloading [20], which have not been used for this purpose in the past. Some of the algorithms used for mobile cloud computing offloading are Chirag et.al. 2024 [21] Proposed an algorithm that combines elements of both the Chimpanzee and Whale Optimization Algorithms (CWOA). In the population initialization phase, to ensure that the population is evenly distributed over the solution space, a Sobol sequence is used, which increases the accuracy of the algorithm. The most important features of the chimpanzee-whale optimization algorithm are the neglect of false positives and the increase in computational speed. Several metrics are used to evaluate the efficiency of the algorithm, most notably task duration, delay, energy usage, and cost. Shuyue et.al. 2021 [22] Proposed the Particle Swarm Optimization (PSO) algorithm that is based on queueing. In the optimization process, Pareto optimality relationship defines the mobility probability to obtain the optimal solution. PSO has proven that the task offloading strategy balances energy consumption and reduces the server MEC delay as well as efficient resource allocation. Lina et.al. 2024 [23] Proposed the Ant Colony Optimization (ACO) algorithm that mimics the behavior of ants in searching for the shortest path to finding food and housing, by leaving pheromone trails that strengthen over time as they are used more. ACO reduces energy consumption and response time by searching for the most efficient paths. The experimental results of this algorithm demonstrated the methodology's effectiveness, significant improvements, and high-quality performance compared to traditional methods. Manal et.al. 2023 [24] Proposed the Binary Cuckoo Search (BCS) algorithm that mimics the behavior of cuckoos that replace weaker eggs with stronger ones by laying their eggs outside their nests, where the algorithm replaces less efficient solutions with better ones. Offloading decisions depend on several parameters, the most important of which are bandwidth and the number of tasks and mobile devices interacting with the edge server. The priority of each task is taken into account when making decisions. The binary cuckoo search algorithm is very efficient in reducing execution time when there are many mobile devices. Mohammad et.al. 2022 [25] Proposed Bacterial Foraging Optimization (BFO) algorithm, it is a nature-inspired algorithm which is based on multi-objective bacterial search optimization. The BFO algorithm is evaluated in comparison with the ant colony optimization (ACO), particle swarm (PSO) and RR algorithms, as it has proven superior to these algorithms in terms of communication cost, response time, and load management effectiveness.

3. LION OPTIMIZATION ALGORITHM (LOA)

Lions are a unique species of cat, characterized by their cooperative and competitive social relationships. There are two types of lions: resident lions, which live in groups called prides, and nomadic lions, which live separately and travel individually or in pairs. There are many behaviors that lions follow, the most important ones were chosen and represented mathematically to simulate nature [9, 26], which are as follows:

First behavior (hunting), Females hunt in three subgroups: two wings (left and right), and the central group. They move toward the prey, surrounding it from several directions. To simulate this behavior, the positions of both the prey and the hunters are updated. Initially the prey is dummy prey and is defined by Equation 1, and its position is updated by Equation 2. The positions of the hunters on the wings are updated by Equation 3, and those of the central hunters are updated by Equation 4.

$$PREY = \sum \frac{hunters(x1, x2, x3, \ldots, xn)}{the\ number\ of\ hunters} \quad (1)$$

Where $PREY$ is the position of dummy prey in the center of hunters, and hunters(x_1, x_2, …, x_n) is the summation position of hunters that will hunt.

$$PREY' = PREY + rand(0,1) \times PI \times (PREY - Hunter) \quad (2)$$

Where $PREY'$ is the new position of prey, $PREY$ is the current position of prey, $rand(0,1)$ is to add a random value between 0 and 1 to make the process of updating lion position more diverse and random, PI is the percentage of improvement in cost of hunter, and $Hunter$ is the new position of hunter who attack to prey.

$$Hunter' = \begin{cases} rand(Hunter, PREY), & Hunter < PREY \\ rand(PREY, Hunter), & Hunter > PREY \end{cases} \quad (3)$$

Where $Hunter'$ the new position of hunter who attacks prey, $Hunter$ is the new position of center hunter, and $PREY$ is the current position of prey.

$$Hunter' = \begin{cases} rand(2 \times PREY - Hunter, PREY), & (2 \times PREY - Hunter) < PREY \\ rand(PREY, 2 \times PREY - Hunter), & (2 \times PREY - Hunter) > PREY \end{cases} \quad (4)$$

Where $Hunter'$ is the new position of left or right hunter, and $Hunter$ is the current position of left or right hunter.

Second behavior (moving toward safe place), the remaining females move towards one of the territory's areas to search for the best one and record it as the best position visited. The females' positions are updating by equation (5).

$$Female\ Lion' = Female\ Lion + 2D \times rand(0,1)\{R1\} + U(-1,1) \times \tan(\theta) \times D \times \{R2\} \quad (5)$$

Where $Female\ Lion'$ is the new position of female lion, $Female\ Lion$ is the current position of female lion, D is the distance between the female lion's position and the selected point chosen by tournament selection among the pride's territory, U(-1, 1) is the uniformly distributed random number is used to prevent female lions from moving in a repetitive manner while moving to a safe place, $\{R1\}$ is a vector which its start point is the previous location of the female lion, and its direction is toward the selected position, $\{R2\}$ is the perpendicular to $\{R1\}$, and $tan(\theta)$ is the angle tangent is used to adjust the random distribution of female lions' movement while moving toward safe place.

Third behavior (Roaming) To search for the best position and record them as the best position visited, the resident males and the nomad males and females are roaming each according to their region. The positions of the resident males are updating by equation (6) and the nomad lions by equation (7).

$$x \sim U(0,2 \times d) \quad (6)$$

Where x is a random number with a uniform distribution that represents the amount of random movement of resident male lions while roaming, U(0,2×d) is a random number generated by uniform distribution that controls the amount of change as the lions' roam, and d is the distance between the male lion's position and the selected area of territory.

$$Lion_{ij}' = \begin{cases} Lion_{ij}, & if\ rand_j > pr_i \\ RAND_j, & otherwise \end{cases} \quad (7)$$

Where $Lion_{ij}'$ is the new position of ith nomad lion, j is dimension, $Lion_{ij}$ is the current position of ith nomad lion, j is dimension, $rand_j$ is a uniform random number within [0, 1], to check transition, $RAND_j$ is random generated vector in search space to generate new random location for lion, and pr_i is the probability of moving for each nomad lion independently.

Fourth behavior (Mating) Resident female mate with one or more resident males, while nomad female mate with only one nomad male. Each type of resident and nomad lions produces two cubs by equations (8) and (9).

$$\text{Offspring}_j 1 = \beta \times Female\ Lion_j + \sum \frac{(1-\beta)}{\sum_{i=1}^{NR} S_i} \times MaleLion_j^i \times S_i \quad (8)$$

$$\text{Offspring}_j 2 = (1-\beta) \times Female\ Lion_j + \sum \frac{\beta}{\sum_{i=1}^{NR} S_i} \times MaleLion_j^i \times S_i \quad (9)$$

Where $Offspring_j 1$ is the first cub, produced from mating parent female lion and male, $Offspring_j 2$ is the second cub, produced from mating parent female lion and male, β is a random number with a normal distribution with mean value 0.5 and standard deviation 0:1, to control the diversity of genetic traits, $Female\ Lion_j$ is the current position of female lion, $MaleLion_j^i$ is the current position of male lion, S_i is represent 1 for select male i for mating, and 0 for not select the male i, and NR is the number of resident males in a pride.

Fifth behavior (Defense) resident males defend the pride against new mature males and nomad males. If the resident lion is weaker than the new mature or nomad lion will either be they killed or migrating to become nomad.

Sixth behavior (Migration) to enhance population diversity and exchange of information between prides, resident females migrate from one pride to another or change their lifestyle to become nomad or vice versa.

4. PROPOSD METHOD

The proposed method relies on AI to solve the problem of offloading tasks from edge devices (mobile devices) to the cloud computing, this problem is NP-Hard problem [27, 28]. This method organizes the execution of operations consisting of multiple tasks by balancing these executions, some of which are on the edge devices and others on the cloud associated with the application executed in mobile applications that contain a cloud server. This is done by generating a set of random solutions called population and then selecting the best among them. The best solution, which represents the execution decision, depends on the cost resulting from a combination of energy consumption and latency. The lower the cost, the less energy in consumed to execute a task in the shortest time. This is achieved by applying the Lion Optimization Algorithm (LOA), which is a meta-heuristic algorithm [29], and which is represented by the diagram in (Fig. 1).

4.1. INITIAL POPULATION

As mentioned earlier in this paper, the Lion Optimization algorithm manages and organizes the offloading of tasks between the cloud and mobile devices. The algorithm first prepares the initial data for the tasks arriving from the cloud server. Let's assume we have a set of tasks to be executed, say ten, with each task divided among twelve workers. A decision is then made for ex-

ecuting tasks by each worker on the cloud (called the edge server) or on the mobile device (called the local device), based on the energy consumption of executing each task (cost). The proposed algorithm generates a set of random solutions, say 80 solutions. Each solution is a two-dimensional matrix consisting of 10 rows representing the tasks and 12 columns representing the workers who will execute the tasks. The values of these rows and columns are randomly generated zeros and ones. Ones represents the task will execute by the worker on the cloud, and zero on the local or mobile device. Table 1 is an example of one solution.

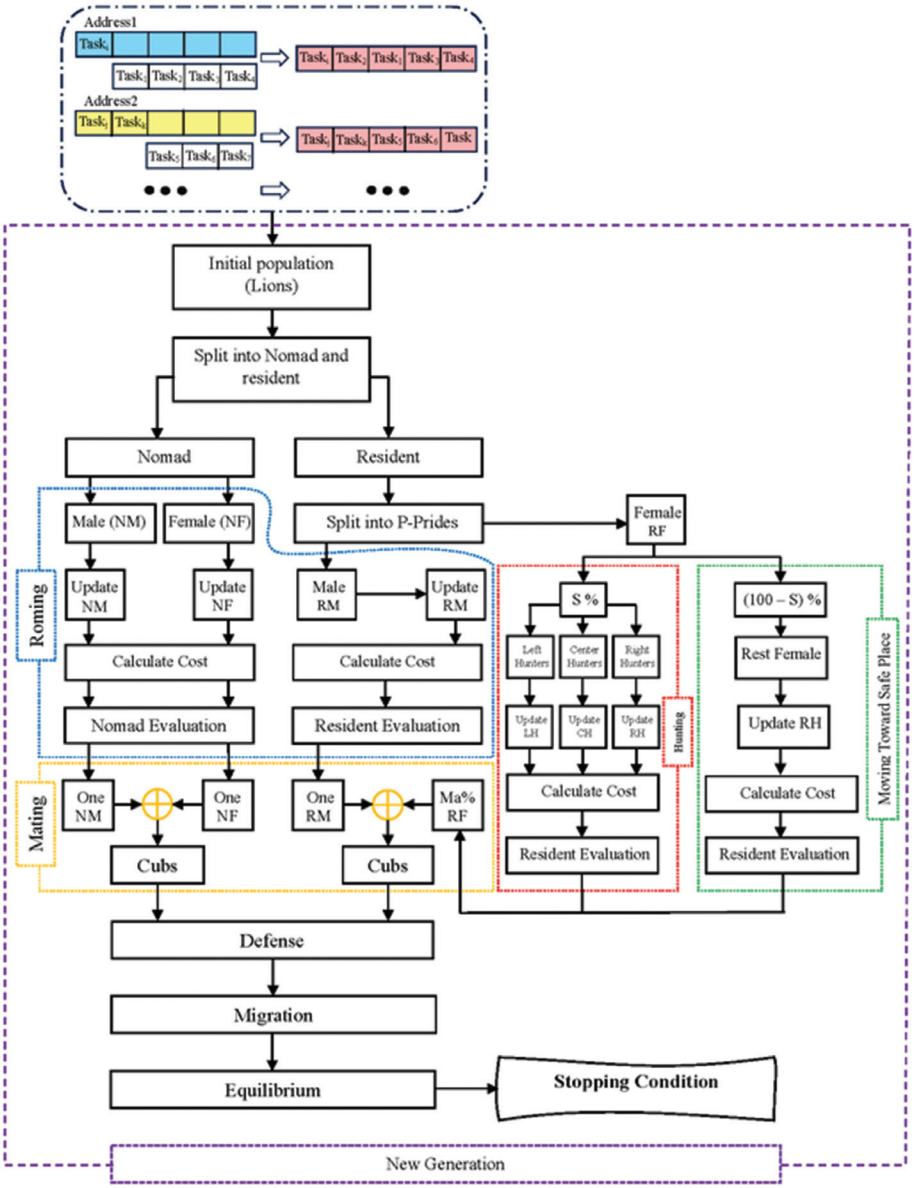


Fig. 1. Diagram of Lion Optimization Algorithm (**LOA**)

Table 1. Example of one solution for population

	0	1	2	3	4	5	6	7	8	9	10	11
0	1	0	1	1	0	0	0	1	0	0	1	0
1	0	0	0	1	1	0	1	1	1	1	1	0
2	0	1	1	0	0	0	0	0	1	0	0	1
3	1	1	0	1	0	0	1	0	1	1	0	0
4	1	0	1	0	0	1	1	1	1	1	0	0
5	1	1	1	1	0	0	1	0	1	0	1	1
6	0	0	1	1	1	1	0	0	1	1	1	1
7	0	1	1	1	0	0	0	0	0	1	1	0
8	0	1	1	1	0	1	0	1	0	1	0	1
9	0	1	0	1	1	1	0	0	1	0	1	1

The set of solutions is called a population, and each solution is called a lion (male or female). Each lion (solution) is evaluated by calculating the cost of each solution to perform tasks based on initial values calculated mathematically.

For example, When calculating the cost of the solution in the Table 1, the result is (12923157959.379).

4.2. LIONS SPLITING

Lions are raandomly split into nomads and resident: nomads represent N% of the total population, and the rest are resident lions.

4.3. THE RESIDENT LIONS ARE SPLIT INTO P-PRIDES

Each pride is randomly split into females (lionesses) S% (between 75-90) of each pride, and the rest are males. The proportion of females in nomad lions is the inverse of the proportion in resident lions, i.e., 1-S, and males are the rest of nomad lions.

4.4. HUNTING

Hunters are females. Through the lion optimization algorithm, they are randomly divided into three groups: the center, which is the best solution, and the left and right wings. The algorithm selects one hunter at a time, and their positions are updated according to the group to which the hunter belongs. If the hunter belongs to the center group, their position is updated according to Equation (3), and Equation (4) if they are on either the left or right wings. Then, the new cost for each location is calculated and the costs are evaluated. If the cost of the new position is better (lower) than the cost of the previous position, the new cost is stored with its position in the list of best solutions, the prey'szposition is also updated according to Equation (2), whose position was previously calculated using Equation (1), but, if the previous position of hunter is better than the new position, the hunter is moved to another location.

Example 1: Hunting
Asumed the same solution example in the Table 1 is the current selected hunter with the following values:
Current hunter's position = 35 in the center group,
Current hunter's cost = **12923157959.379**
Current prey' position = 12
The **LOA** calculates the following **output** of new values:
New hunter's position = 17, New hunter's cost = **12552743255.394**
In this case hunter improved its own position and **LOA** will select the lowest cost for new position and update prey's position equal to 11.

4.5. MOVE TOWARD SAFE PLACE

After select female hunters for hunting, the rest females (100-S) from each pride move toward safe place. The LOA updating the position to get the new position by the equation (5), then calculate the cost of new position and evaluatting the cost of new position and previous position to save the best cost with it position in the best cost list.

Example 2: Move toward safe place
Asume the females number of pride = 9, The number of hunters = 7, then the number of rest female in the pride = 2 wich represent the females which are moving towared safe place.
After applying **LOA**, the solution at the current position (35), which represents (*Female Lion*) according to equation (5), its cost (**12923157959.379**) is computed using a fitness function.
After execution equation (5), a new solution (represent*Female Lion'* according to equation (5)) is created and its cost is computed.
In this example, the new cost is equal to (**12220873013.407**) which is less than the cost at the current position (35). Therefore, the LOA replaces the new solution with the currrent one at location 35.

4.6. ROAMING

During the roaming process of resident lions, the LOA randomly selects specific places within the pride's territory at a rate of R% based on the number of males. Each time, a male is selected from the resident lions to move from one place to another in search of the best solution. Using Equation (6), the lion's position is updated to calculate its cost and evaluate it with the cost of the previous position visited by each lion. As for the nomad lions, the lions are selected one by one, males and females, and their positions are updated according to Equation (7), then the costs are calculated and evaluated in comparison with the costs of the previous positions, then the best is chosen and stored in the list of the best solutions.

Example 3: lions roaming
Let's assume the following values:
$R = 90\%$
Number of males in the pride = 5
Locations to be selected from the pride's territory = $5 * R = 4$
The lion selected for roaming = 51
The cost of lion selected = **145776.230**
The **LOA** updating these values to be the following output results:
The cost of new position = **221497.173**
The **LOA** don't store the cost of new position, due to its lowest than cost of previous position.

4.7. Mating

To simulate the mating process among resident lions, the **LOA** selects a percentage (Ma%) of the total females in each pride to mate with one or more males, randomly selected by the **LOA**. In nomad lions, mating occurs between one female and one male. Based on Equations (8) and (9), two cubs are produced in both resident and nomad lions, whose sexes are randomly determined later. Both sexes are also genetically mutated at a rate of Mu%. The new cubs are then added to the population.

Example 4: lions mating
Population size = 100, number of prides = 4 for resident lions, the number of females to mate = 3
The output results:
• Each female mates with two males, producing two

cubs of lions. This means that there are 7 matings and two cubs per mating, producing 14 cubs for each pride of resident lions, as follows:
3 matings * 2 cubs * 4 prides = 24 cubs. In this example, this is for resident lions only.
However, in nomad lions, only two cubs are produced, which are added to the total number of cubs.
Total number of cubs = 26.
The final population size = 126 lions, or solutions.

4.8. DEFENSE

In the **first** instance of defense against new mature resident males, as mentioned above, the **LOA** first merges the positions of old males with new mature males, then sorts them in descending order of cost per position. Finally, the weakest male from the pride, the one with the highest cost, is removed from the list and becomes a nomad male. The **second** defense is against nomad male lions. **LOA** works by randomly selecting the prides to be attacked by nomad male (nomad male are selected one by one). The prides to be attacked are represented by a randomly selected list of zeros and ones. One means the pride will be attacked, and zero means it will not attack. If the nomad male is less cost than the resident male, the resident male is drive out from the pride, becoming a nomad lion, and the nomad male becomes a resident lion.

Example 5: First instance of defense
old resident males of the first pride from four prides = [92, 43, 70, 84, 90]
Cubs males of the first pride = [100, 102, 104, 106, 108, 110, 112]
Nomad male lions = [11, 29, 69, 97, 35, 28, 36, 34, 51, 30, 14, 24, 89, 21, 13, 156]
Number = 16
Output
List of resident males after merging = [92, 43, 70, 84, 90, 100, 102, 104, 106, 108, 110, 112]
The same list arranged in ascending order according to costs =
[43, 100, 102, 104, 106, 84, 90, 92, 112, 70, 108, **110**]
After the drive out of the weakest lion, which is position **110**, the new list of resident lions becomes =
[43, 100, 102, 104, 106, 84, 90, 92, 112, 70, 108]
List of nomad lions after the weakest lion moves to it = [11, 29, 69, 97, 35, 28, 36, 34, 51, 30, 14, 24, 89, 21, 13, 156, **110**]
Final list of nomad lions after the four weakest resident males move = [11, 29, 69, 97, 35, 28, 36, 34, 51, 30, 14, 24, 89, 21, 13, 156, 110, 126, 138, 45]
Nomad males number = 20 lions or solutions

4.9. MIGRATION

To simulate the migration behavior of lions, the algorithm configures the number of females to migrate (*Number of migratory females* = [*Number of surplus females per pride + Number of resident females allowed per pride * I%*] * *Number of prides*). The surplus females represent the females produced during the mating among resident lions. After selecting the females designated for migration, the algorithm moves them to nomad females by removing them from the resident females list and adding them to the nomad females list. The nomad females list is sorted in ascending order according to the best female, i.e., from the lowest cost to the higher one. The number of females removed from the resident females list is replaced by the lowedt cost females from nomad females list as (*Number of females required to replace the resident females = Number of resident females allowed per pride – Number of migratory females per pride*).

4.10. EQUILIBRIUM

To equilibrium the number of lions in the population, the algorithm works by returning it to the allowed original population size (for example, 100 solutions or lions), by eliminating the solutions with the highest costs. If the population size reaches, for example, 120 solutions or lions, the 20 solutions with the highest costs are eliminated, meaning that 100 solutions with costs lower than the original 20 solutions are retained.

4.11. CONVERGENCE

In this paper, the search for the best solutions stops after 200 iterations are completed. The best solution is selected from among 100 solutions (100 solutions = population size) from each iteration resulting 200 best solutions, which are compared with other 200 best solutions generated by the Genetic Algorithm **GA** [30-32].

5. EXPERIMENTAL RESULT

For the purpose of evaluating **LOA** by comparing it with **GA** the same initial data is used, the result of costs in all tables (3, 4, 5) and all diagrams in the figures (7-10) are generated by the implementing each of the **LOA** based on single objective and Implementing **GA**. First, the data is described which is used for this purpose in section 5.1, then comparing the results for evaluating **LOA** through the cost values of tables in section 5.2 and diagrams in section 5.3 with fixed some values and changing others.

Example 7: one of the 12 cases.
[array([[0.000749, 0.00106, … , 0.000676, 0.000996],
[0.000551, 0.00049, … , 0.000807, 0.000801],
. .
. .
[0.000393, 0.000816, … , 0.000493, 0.00796],
[0.000513, 0.000729, … , 0.000633, 0.000353]]),
array([……..……..…]),
array([……..…..……]),
array([……..…..……])]

5.1. DATA DESCRIPTION

To evaluate **LOA** by comparing it with **GA**, randomly generated data was used which matching the real dataset that is currently unavailable. Twelve cases from this random data were recorded and saved in a PKL file in the form of a three-dimensional matrix. Each case contains four arrays, each one includes four variables (T_{local}, E_{local}, T_{edge}, *and* E_{edge}).

Where T_{local} is the local time consumption, E_{local} is the local energy consumption, T_{edge} is the marginal time consumption, and E_{edge} is the marginal energy consumption, all for the worker w_j task c_i, ($i \in 1, 2, 3, \ldots, n$) and ($j \in 1, 2, 3, \ldots, m$).. Each variable consists of a matrix of 10 rows and 12 columns, as in Table 2. The data of the four variables are created based on the r_{local} = 656725 (Local data processing rate), q_{local} = 356713 (Consumption per bit of data processed locally), r_{edge} = 127571 (The rate at which edge server data is processed), q_{edge} = 852433 (Transmission energy consumption per bit data of edge server), x_{edge} = 469182 (Edge server data transfer rate), and yedge = 114295 (Energy consumption per bit of data processed by edge servers).

5.2. EVALUATING LOA BY COMPARING IT WITH GA IN DEFERENT CASES

All cases for **LOA** have constant values which are the roaming (R) = 30%, resident females (S) = 75%, prides number (P) = 4, mating (Ma) = 20%, resident females' surplus (I) = 10%, and nomad (N) = 10%.

a. Mutation (Mu)= 3% and iteration= 100, Table 3.

b. Population size=200 and iteration= 100, Table 4.

c. Population size=200 and mutation (Mu)=3%, Table 5.

Table 2. Complete values of array for the variable T_{local}

	0	1	2	3	4	5	6	7	8	9	10	11
0	0.00074	0.00106	0.00046	0.00096	0.00096	0.00062	0.00034	0.00076	0.00037	0.00106	0.00067	0.00099
1	0.00055	0.00049	0.00071	0.00064	0.00060	0.00069	0.00094	0.00051	0.00053	0.00091	0.00080	0.00080
2	0.00102	0.00044	0.00070	0.00040	0.00051	0.00106	0.00101	0.00084	0.00035	0.00103	0.00102	0.00080
3	0.00077	0.00047	0.00087	0.00105	0.00045	0.00081	0.00076	0.00088	0.00095	0.00096	0.00103	0.00046
4	0.00055	0.00060	0.00062	0.00046	0.00098	0.00069	0.00091	0.00084	0.00037	0.00036	0.00098	0.00083
5	0.00064	0.00087	0.00077	0.00059	0.00098	0.00035	0.00101	0.00088	0.00096	0.00043	0.00072	0.00046
6	0.00076	0.00105	0.00086	0.00070	0.00106	0.00080	0.00092	0.00073	0.00039	0.00099	0.00053	0.00071
7	0.00051	0.00058	0.00064	0.00103	0.00085	0.00080	0.00098	0.00104	0.00096	0.00106	0.00090	0.00041
8	0.00039	0.00081	0.00037	0.00047	0.00042	0.00092	0.00091	0.00033	0.00056	0.00077	0.00049	0.00079
9	0.00051	0.00072	0.00062	0.00065	0.00094	0.00066	0.00090	0.00058	0.00055	0.00103	0.00063	0.00035

Table 3. Three different Population sizes

	Population size, Mu = 3%, iteration = 100					
Data Case	**100**		**150**		**200**	
	LOA-cost	GA-cost	LOA-cost	GA-cost	LOA-cost	GA-cost
0	8723034721.51	9720584791.47	8578311175.52	10019971021.46	8372255875.52	10468444321.44
1	8132422510.49	9954436198.41	7793764564.50	9715896886.42	7522013986.51	9349845706.44
2	8985631163.52	10328626883.47	8794848197.53	10486441001.46	8245286591.55	10666072739.45
3	8505810941.52	10517637923.44	8387753375.52	10036923029.45	8026308137.53	9242034407.49

Table 4. Three different Mutation ratios

	Mutation, Population size= 200, iteration = 100					
Data Case	**3%**		**5%**		**7%**	
	LOA-cost	GA-cost	LOA-cost	GA-cost	LOA-cost	GA-cost
0	8372255875.52	10468444321.44	8434072465.52	10040818969.46	8657339443.51	9940942753.46
1	7522013986.51	9349845706.44	7886610658.50	9440267620.44	8218965736.48	9967284352.41
2	8245286591.55	10666072739.45	8459584103.54	10928369015.44	8774242667.53	10124510927.48
3	8026308137.53	9242034407.49	8662655387.51	9487361423.48	8826045119.50	10269644309.45

Table 5. Three different Iterations

	Iteration, Population size= 200, Mu = 3%,					
Data Case	**100**		**200**		**300**	
	LOA-cost	GA-cost	LOA-cost	GA-cost	LOA-cost	GA-cost
0	8372255875.52	10468444321.44	7698576253.55	9501923755.48	8044264321.54	10040576551.46
1	7522013986.51	9349845706.44	7922246104.50	9206576668.44	7944548560.49	9730441966.42
2	8245286591.55	10666072739.45	8475098855.54	9900516695.48	8398009931.54	10492259033.46
3	8026308137.53	9242034407.49	8371511369.52	9481785809.48	7868326601.57	9670387013.47

5.3. EVALUATING LOA BY DIFFERENT CURVES OF LOA WITH CHANGING SOME VALUES

The constant values are population size = 150, Iteration= 100, prides number (P) = 4, resident females (S) = 75%, mating (Ma) = 20%, mutation (Mu) = 3%, and resident females' surplus (I) = 10%.

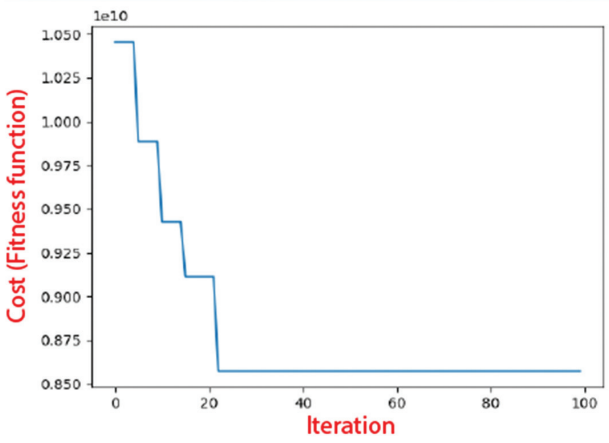


Fig. 2. $N = 10\%$, $R = 30\%$

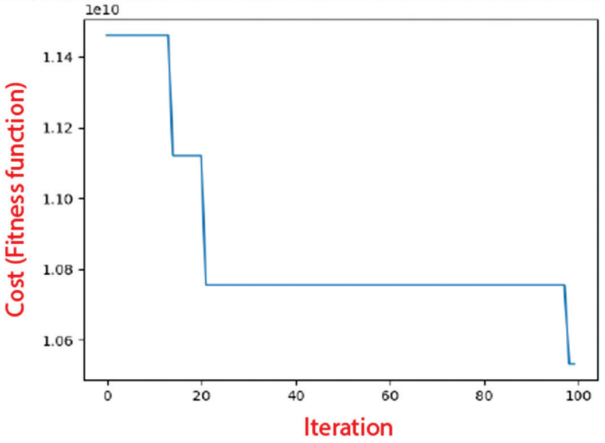


Fig. 3. $N = 10\%$, $R = 5\%$

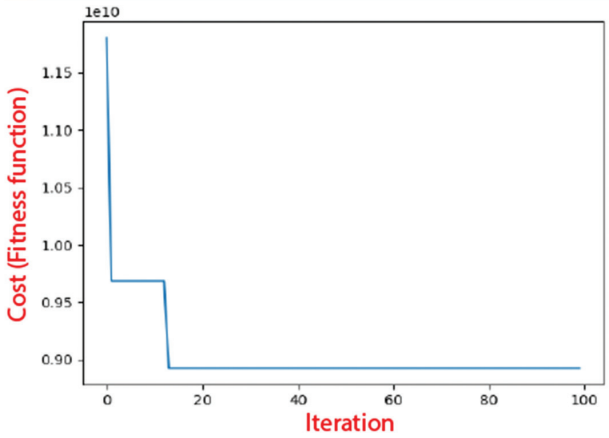


Fig. 4. $N = 40\%$, $R = 30\%$

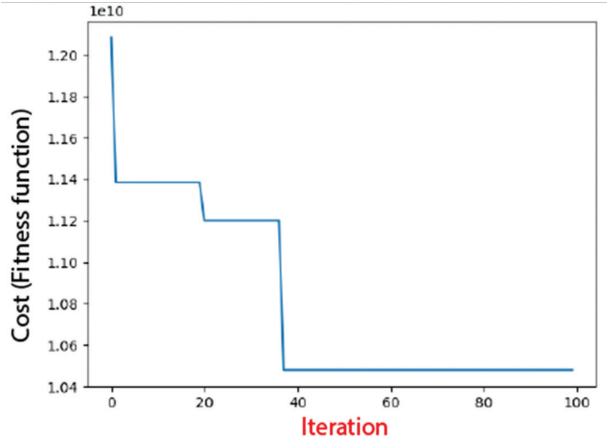


Fig. 5. $N = 40\%$, $R = 5\%$

6. CONCLOSION

In mobile cloud computing task-offloading, an optimization technique based on the **LOA** is suggested. In order to decrease task execution time and energy consumption, tasks are transferred from mobile devices to servers that manage complex computational operations, such as cloud servers or Mobile-Edge Computing servers. The suggested approach performs well in lowering offloading work costs. Multi-objectives could be utilized for further cost reduction. **LOA** has several groups, each of them attracted to produced new solutions that enhances the algorithm's ability to adapt in different environments. the experimental tests when increasing the mutation rates. The performance of **LOA** in offloading is butter than **GA** when applied in same sample data. The evaluation process shows through comparing the **LOA** with **GA** in the all tables that the results of the costs by **LOA** are always less than **GA**. For example, in the record (0) of Table 3, the costs of two algorithms are as follows: **LOA** cost = 8723034721.51, **GA** cost = 9720584791.47. The unit of cost is a combination of joules and second.

7. ACKNOWLEDGEMENT

The authors would like to thank Mustansiriyah University (www.uomustansiriyah.edu.iq Baghdad Iraq) for its support in the present work.

8. REFERENCES

[1] J. Hasoon, R. Hassan, "Big Data Techniques: A Survey", Iraqi Journal of Information Technology, Vol. 9, No. 4, 2019.

[2] M. Pai, S. Rajarajeswari, D. P. Akarsha, S. D. Ashwini, "Analytical Study on Load Balancing Algorithms in Cloud Computing", Expert Clouds and Applications, Vol. 209, 2022, pp. 631-646.

[3] S. Dong, J. Tang, K. Abbas, R. Hou, J. Kamruzzaman, L. Rutkowski, R. Buyya, "Task offloading strategies for mobile edge computing: A survey", Computer Networks, Vol. 254, 2024, p. 110791.

[4] M. Hashemi, A. Masoud, "Load Balancing Algorithms in Cloud Computing Analysis and Performance Evaluation", 2020, https://philarchive.org/archive/HASLBA (accessed: 2025)

[5] R. Kaviarasan, G. Balamurugan, R. Kalaiyarasan, Y. Reddy, "Effective load balancing approach in cloud computing using Inspired Lion Optimization Algorithm", e-Prime - Advances in Electrical Engineering, Electronics and Energy, Vol. 6, 2023, p. 100326.

[6] J. Adaikalaraj, C. Chandrasekar, "To improve the performance on disk load balancing in a cloud environment using improved Lion optimization with min-max algorithm", Measurement: Sensors, Vol. 27, 2023, p. 100834.

[7] A. Gaikwad, K. Singh, S. Kamble, M. Raghuwanshi, "A comparative study of energy and task efficient load balancing algorithms in cloud computing", Physics: Conference Series, Vol. 1913, 2021.

[8] B. Gerald, D. R. P. Geetha, "Metaheuristic algorithm-based load balancing in cloud computing", Theoretical and Applied Information Technology, Vol.102, No. 5, 2024, pp. 2099-2115.

[9] M. Yazdani, F. Jolai, "Lion Optimization Algorithm (LOA): A nature-inspired metaheuristic algorithm", Computational Design and Engineering, Vol. 3, No. 1, 2016, pp. 24-36.

[10] A. Alaidi, C. Soong Der, Y. W. Leong, "Systematic Review of Enhancement of Artificial Bee Colony Algorithm Using Ant Colony Pheromone", International Journal of Interactive Mobile Technologies, Vol. 15, No. 16, 2021, pp. 172-180.

[11] Q. You, B. Tang, "Efficient task offloading using particle swarm optimization algorithm in edge computing for industrial internet of things", Cloud Computing, Vol. 10, No. 41, 2021.

[12] C. Guo, H. Tang, B. Niu, C. P. Lee, "A survey of bacterial foraging optimization", Neurocomputing. Vol. 452, 2021, pp. 728-746.

[13] B. Rajakumar, "Lion Algorithm and Its Applications", Frontier Applications of Nature Inspired Computation, 2020, pp. 100-118.

[14] M. Mansor, M. Kasihmuddin, S. Sathasivam, "Modified Lion Optimization Algorithm with Discrete Hopfield Neural Network for Higher Order Boolean Satisfiability Programming", Malaysian Journal of Mathematical Sciences, Vol. 14, 2020, pp. 47-61.

[15] S. Almufti, "Lion algorithm: Overview, modifications and applications", international research journal of science, technology, education and management, Vol. 2, No. 2, 2022, pp. 176-186.

[16] F. Saeik, M. Avgeris, D. Spatharakis, N. Santi, D. Dechouniotis, J. Violos, A. Leivadeas, N. Athanasopoulos, N. Mitton, S. Papavassiliou, "Task offloading in Edge and Cloud Computing: A survey on mathematical, artificial intelligence and control theory solutions", Computer Networks, Vol. 195, 2021, p. 108177.

[17] X. Chen, G. Liu, "Federated Deep Reinforcement Learning-Based Task Offloading and Resource Allocation for Smart Cities in a Mobile Edge Network", Sensors, Vol. 22, 2022, p. 4738.

[18] X. Shichao, Y. Zhixiu, W. Guangfu, L. Yun, "Distributed Offloading for Cooperative Intelligent Transportation Under Heterogeneous Networks", IEEE Transactions on Intelligent Transportation Systems, Vol. 23, No. 9, 2022, pp. 16701-16714.

[19] A. Al-Obaidi, H. Abdullah, Z. Ahmed, "Camel Herds Algorithm: a New Swarm Intelligent Algorithm to Solve Optimization Problems", International Journal on Perceptive and Cognitive Computing, Vol. 3, No. 1, 2017.

[20] L. Meng, Y. Wang, H. Wang, X. Tong, Z. Sun, Z. Cai, "Task offloading optimization mechanism based on deep neural network in edge-cloud environment", Cloud Computing, Vol. 12, No. 76, 2023.

[21] C. Chandrashekar, P. Krishnadoss, V. Kedalu, B. Anantharakrishnan, "MCWOA Scheduler: Modified Chimp-Whale Optimization Algorithm for Task Scheduling in Cloud Computing", Computers, Materials and Continua, Vol. 78, No. 2, 2024, pp. 2593-2616.

[22] M. Shuyue, S. Song, L. Yang, J. Zhao, F. Yang, L. Zhai, "Dependent tasks offloading based on particle swarm optimization algorithm in multi-access edge computing", Applied Soft Computing, Vol. 112, 2021, p. 107790.

[23] L. Ibrahim, O. Fadare, F. Al-Turjman, A. ALwhelat, "Ant Colony Optimization with Lagrangian Relaxation for Cloud Computing Offloading Optimiza-

tion", Dijlah Journal of Engineering Science, Vol. 1, No. 1, 2024, pp. 49-56.

[24] M. Alqarni, A. Cherif, E. Alkayyal, "ODM-BCSA: An Offloading Decision-Making Framework based on Binary Cuckoo Search Algorithm for Mobile Edge Computing", Computer Networks, Vol. 226, 2023, p. 109647.

[25] M. Babar, A. Din, O. Alzamzami, H. Karamti, A. Khan, M. Nawaz, "A Bacterial Foraging Based Smart Offloading for IoT Sensors in Edge Computing", Computers and Electrical Engineering, Vol. 102, 2022, p. 108123.

[26] S. Sridhara, A. Chakravarthy, U. Nandini, "Asiatic Lions: Behaviour in the Wild and Captivity", Animal Behavior in the Tropics, Springer, 2025, pp. 491-502.

[27] X. Tan, M. Emu, S. Choudhury, "Task Offloading in Mobile Edge Computing: Intractability and Proposed Approaches", Proceedings of the IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology, Niagara Falls, ON, Canada, 17-20 November 2022, pp. 775-778.

[28] A. ALRIDHA, A. Salman, A. Al-Jilawi, "The Applications of NP-hardness optimizations problem", Journal of Physics: Conference Series, Vol. 1818, 2021, p. 012179.

[29] B. Vijayaram, V. Vasudevan, "Wireless edge device intelligent task offloading in mobile edge computing using hyper-heuristics", EURASIP Journal on Advances in Signal Processing, Vol. 2022, 2022, p. 126.

[30] Z. Li, Q. Zhu, "Genetic Algorithm-Based Optimization of Offloading and Resource Allocation in Mobile-Edge Computing", Information. Vol. 11, No. 83, 2020.

[31] H. Yektamoghadam, R. Haghighi, M. Dehghani, A. Nikoofard, "Genetic Algorithm and Its Applications in Power Systems", Frontiers in Genetics Algorithm Theory and Applications, Tracts in Nature-Inspired Computing, Springer, 2024, pp. 83-97.

[32] M. Younis, S. Yang, "A Genetic Algorithm for Independent Job Scheduling in Grid Computing", MENDEL Soft Computing, Vol. 23, No. 1, 2017.

Attribute-Based Authentication Based on Biometrics and RSA-Hyperchaotic Systems

Original Scientific Paper

Safa Ameen Ahmed*
Information Technology & Communications University,
Informatics Institute for Postgraduate Studies, Baghdad, Iraq
University of Technology,
College of Chemical Engineering,
Department of Chemical Engineering and Petroleum Pollution, Baghdad, Iraq
safa.a.ahmed@uotechnology.edu.iq

Ali Maki Sagheer
University of Anbar, College of Computer Science and Information Technology,
Department of Computer Networks Systems, Ramadi, Iraq
Ali_makki@uoanbar.edu.iq

*Corresponding author

Abstract *– Attribute-based authentication is a security technique that allows access to resources according to characteristics of human biometrics. It confirms the security and increases the efficiency of applications, devices, and resources by controlling them and avoiding cyberattacks. Standard feature extraction approaches can also give erroneous results and limit processing efficiency when processing complex biometric data. This paper proposes a hybrid method combining the Rivest-Shamir-Adleman (RSA) method with 6D hyperchaotic systems to generate dynamic and sophisticated keys for key exchange in a more secure and effective authentication system. User attributes like ID and fingerprint are used for attribute-based authentication by extracting fingerprint features (Minutiae extraction) for biometric matching. The 6D-hyperchaotic systems generate dynamic values over time, influenced by the initial values and input constants. Three of the generated sequences were used on the sender side, and the other three sequences were used on the receiver side after processing them to satisfy the RSA condition. The time consumed for generation numbers is about 0.0717 msec. The results indicated that the system that produces the keys has robust resistance to statistical attacks. The average time for authentication is 0.307867 sec. Thus, the research presents an integrated solution that improves authentication system security and dependability, especially in critical areas that demand sophisticated data and user protection. The user identity verification reached 95.14% accuracy, and the proposed method could be integrated with the extended system by adding a node.*

Keywords: *Authentication, Fingerprint, Minutiae extraction, Hyperchaotic Systems, RSA, Public key, and Private key*

Received: March 13, 2025; Received in revised form: June 20, 2025; Accepted: July 18, 2025

1. INTRODUCTION

Due to the rapid development of digital systems and the growing demand for secure and efficient authentication methods, Attribute-Based Authentication (ABA) has emerged as a potential approach for improving cybersecurity. User attributes, including username, ID, phone number, email, gender, address, age, birthday, and biometric information are used to generate public and private keys for policy access, encryption, and decryption to protect user data [1, 2]. Recent ABE methods use elliptic curves and pairing-based encryption, including the integer factorization problem with RSA. Reduced ciphertext sizes, faster key and signature generation, and improved security [3]. RSA uses public and private keys for secure key exchange and secret transmission, producing digital signatures for data integrity and sender identity [4]. Advancements in technology and hardware enable RSA to offer robust security with adequate key lengths [5]. However, the benefit is that a substantial key size renders decryption using established algorithms challenging [6]. Multiple attacks have made token- or password-only authentication solutions less easy and reliable [7]. Biometrics are one of the most important features in authentication systems to prevent access control, impersonation, and fraud [8]. Due to biometric unique ability to verify the identity of individuals based on their physiological or behavioral characteris-

tics, such as fingerprints, iris, or voice patterns [9]. Fingerprints are identified by their lines and curves. Their lifelong reliability makes them reliable [10]. Fingerprint classification relies on global and local characteristics, with global features like Singular Point (SP) enhancing matching and alignment, and local features like minutiae and ridge features that most identification algorithms use in matching [11, 12]. Fingerprint recognition technology provides safe biometric authentication, with current developments in picture capture, preprocessing, feature extraction, and matching for dependable identification [13, 14]. As online services become more prevalent, protecting sensitive data is crucial due to the rise of sophisticated cyber threats like tampering, spying, and phishing [15]. In this environment, dependence exclusively on conventional security measures like passwords or firewalls is inadequate [16]. However, strong encryption techniques must be incorporated to ensure high security and prevent potential attacks, such as impersonation [17]. Integrating biometric authentication and utilizing personal data such as fingerprints or iris patterns with the RSA algorithm provides enhanced security. Biometric authentication enhances privacy and security, increases system complexity, reduces hacking risk, and replaces traditional methods such as those using passwords and ID numbers [18].

This paper aims to introduce a feature-based authentication system that integrates biometric feature extraction (fingerprint), the RSA algorithm, and a six-dimensional chaotic system to produce dynamic and extremely unpredictable keys. The study assesses the system's efficacy regarding authentication precision, cryptographic robustness, key generation unpredictability, and response latency, in comparison to conventional techniques.

The following section includes the structured rest of the article. Section 2 presents a review of the relevant literature. The 6D Hyperchaotic system is described in section 3. The proposed method is detailed in Section 4. The experimental results are discussed in Section 5. Section 6 offers the conclusion of this work.

2. RELATED WORK

The proposed method can be divided into three sections: Biometric identification using fingerprint matching, key generation using a 6D hyperchaotic system, and authentication using RSA and ABE. Therefore, the related work will be divided into three sections.

In the first section of fingerprint matching, S. Socheat and T. Wang (2020) proposed a fingerprint enhancement, extraction, and matching architecture. Three steps made up this system. Preprocessing the fingerprint surface with brightness and Gabor filters darkens the ridgelines to improve image quality.

Next, extract features via binarizing, thinning, localizing minutiae, removing erroneous minutiae, and classifying to locate and count fingerprint minutiae.

The final step is minutiae matching to validate an individual's identity by comparing findings to the database. Previously registered individuals' results will be validated [19]. Situmorang and Herdianto (2022) developed a biometric identification system using fingerprint images from mobile phone cameras. The methodology involved pre-processing steps, including cropping, grayscale conversion, binarization, and thinning of fingerprint images, then fine feature extraction using the Crossing Number (CN) and minutiae-based matching methods. The accuracy of fingerprint fine-grained verification was up to 92.8% using a 5MP camera and 95.11% using an 8MP camera in fingerprint recognition. The average accuracy of fingerprint fine-grained verification was found to be 63% [20].

In the second section, the key generation using a hyperchaotic system, Akif *et al.* (2021) presented a Pseudorandom Bit Generator (PRBG) that utilizes a new two-dimensional chaotic logistic map, including mouse movement data and chaotic systems to generate unpredictability. The system employs algorithms to produce pseudorandom bits derived from mouse movement coordinates, augmenting nonlinearity and randomness. The control parameters a and b are established at a=0.8 and b within the range of 2 to 5, thereby guaranteeing non-redundancy in the chaotic system [21].

In the third section, the authentication section, Tahat *et al.* (2020) produced a novel secure cryptosystem by integrating the Dependent-RSA (DRSA) with chaotic maps (CM) that rely on both integer factorization and Chaotic Maps Discrete Logarithm (CMDL) [22]. Zhang *et al*. (2024) presented the Chebyshev–RSA Public Key Cryptosystem with Key Identification (CRPKC-Ki) public key cryptography algorithm, which improves security and efficiency by integrating Chebyshev polynomials with RSA. The procedure is achieved by selecting two large-prime integers and computing N and the totient, and presents alternative multiplication coefficients K_i, generated by Chebyshev chaotic mapping, established by the shared secret selection criteria among participants [23].

3. 6D HYPERCHAOTIC SYSTEM

The 6D hyperchaotic Systems exhibiting complex and implicit extreme multi-stability demonstrate the following dynamic phenomena on a line or equilibrium plane: hidden extreme multi-stability, transient chaos, bursting, and offset boosting. This chaotic system is the inaugural high-order system to manifest all these intricate dynamic behaviors [24].

$$\dot{x}_1=\alpha(1-\beta|x_6|)\,x_2-ax_1 \quad (1)$$

$$\dot{x}_2=cx_1+dx_2-x_1x_3+x_5 \quad (2)$$

$$\dot{x}_3=-bx_1+x_1^2 \quad (3)$$

$$\dot{x}_4=ex_2+fx_4 \quad (4)$$

$$\dot{x}_5=-rx_1-kx_5 \quad (5)$$

$$\dot{x}_6=-x_2 \quad (6)$$

The mentioned chaotic system comprises six equations, Eq. (1), Eq. (2), Eq. (3), Eq. (4), Eq. (5), and Eq. (6), that include variables and parameters. $x1$, $x2$, $x3$, $x4$, $x5$, and $x6$ represent the state variables, while $\alpha, \beta, a, b, c, d, e, f, r$, and k denote the system parameters. [25]. These equations constitute a highly intricate model that illustrates the interrelated dynamic interactions inside a chaotic system, showcasing the system's capacity to produce various states of dynamic stability and unstable motion, oscillating between phases of chaos and complicated dynamic equilibria. This system comprises a set of differential equations that delineate the temporal evolution of each of the six dynamic variables, contingent upon the present values of the other variables.

4. PROPOSED METHOD

The proposed method is utilized to verify the individuals who are using the systems that are part of the network. This process is carried out more simply by employing biometric characteristics to verify the individual's identity. Additionally, this process is carried out by creating a set of keys to prevent unauthorized access to the systems. Fig. 1 shows the structure of the general proposed model.

Fig. 1. General proposed model

First, the system authenticates users by scanning their fingerprints, performing feature extraction, and executing matching actions to train the system and give the user his unique ID during the registration phase which is conducted by an offline mechanism to guarantee user attribute information such as the username, identification number, phone number, email, gender, address, age, birthday, and biometric information must be secret. The server also generates the private key (E_c, N_c) and the public key (D_c, N_c) for the user; the private key for the server is (E_s, N_s), and the public key is (D_s, N_s), which are used for authentication. Upon attempting to log in to the system, the user's fingerprint is scanned then the identical preprocessing, feature extraction, and matching protocols employed during the training phase are performed to identify the user. Authentication takes place in the initial step following user identification. However, he will be denied access to the system if the user is not identified. During the second step, the user digests his ID using his private key (E_c, N_c) in agreement with the server.

The result is encrypted with the server's public key (D_s, N_s) and sent to the server. The server decrypts the data using its private key (E_s, N_s), the result is decrypted using the user's public key (D_c, N_c) to obtain the ID, and correlates it with the user's stored ID. This procedure confirms the user's authorization to use the system. Fig. 2 shows the proposed Attribute-Based Authentication method.

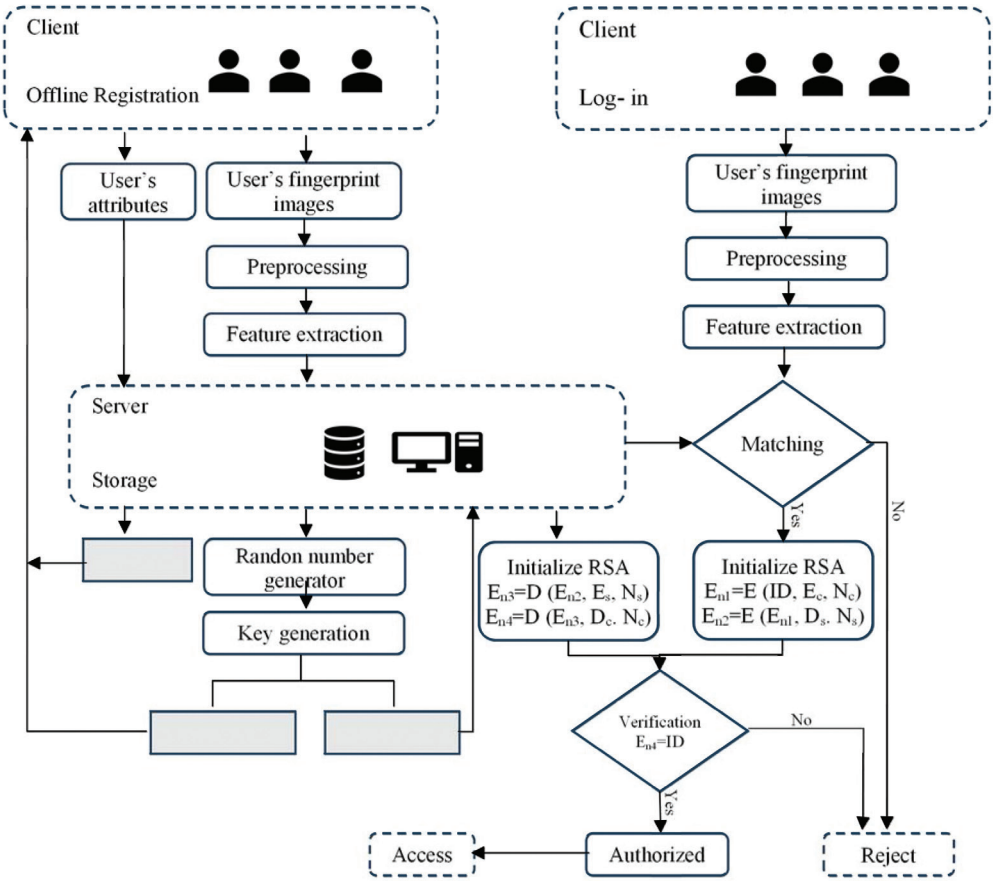


Fig. 2. The proposed Attribute-Based Authentication Method

4.1. REGISTRATION PHASE

The registration phase is conducted by an offline mechanism to guarantee that user attribute information, such as the username, identification number, phone number, email, gender, address, age, birthday, and biometric information, is kept secret. All user attributes are stored on the server. Then, the server gives the user their unique ID during the registration phase. The proposed system starts by scanning the user's fingerprint, preprocessing it, and extracting features from the fingerprint image used to train the system.

4.1.1. Preprocessing

Several processes have been performed on row fingerprint images in the dataset to generate standardized images. This ensures that the system receives consistent inputs. The preprocessing steps are noise removal, histogram equalization, resizing, binarization, and thinning. Fig. 3 shows the block diagram for the preprocessing step.

Fig. 3. The preprocessing stages

Noise removal is essential for reducing noise during fingerprint acquisition. As a result, a fine fingerprint image is obtained, and valleys and ridges are detected, as shown in Fig. 4.

(a) (b)

Fig. 4. Noise removal: (**a**) Original fingerprint image and (**b**) Noise removal fingerprint image

Histogram equalization improves grayscale image contrast by redistributing pixel values, reducing tonal density, creating a more homogeneous distribution of pixels, and enhancing detail clarity. This strategy greatly improves low-contrast images and improves feature extraction by highlighting fingerprint ridges and valleys, as shown in Fig. 5.

(a) (b)

Fig. 5. Histogram equalization (**a**) noise removed image, and (**b**) histogram equalized image

In the resizing step, the fingerprint image dimensions and orientation are changed to a standard size, and feature matching depends on this step, as shown in Fig. 6.

(a) (b)

Fig. 6. Image Resizing (**a**) equalized image, and (**b**) resized image

Finally, the fingerprint image is thinned to one-pixel lines without losing the structure. Biometric identification systems reduce changes via fingerprint analysis and minutiae extraction. This method enhances binary image feature extraction, recognition, and analysis, as shown in Fig. 7.

(a) (b)

Fig. 7. Edge Detection (**a**) resized image, and (**b**) thinning edge image

The preprocessing steps are crucial for enhancing the fingerprint image for the following processes, including feature extraction and feature matching, which improves the efficacy of the fingerprint verification system.

4.1.2. Feature extraction

The CN method is used to extract features from fingerprint images to identify the individual carrying the fingerprint. This process involves identifying minutiae, such as ridge ends and bifurcation points, and classifying them using coordinates, angles, and types. The crossing number method is based on studying the pixel arrangement around the fingerprint, which helps classify minutiae, as shown in Eq. (7).

$$CN = \frac{1}{2}\sum_{i=1}^{8}|P_i - P_{i-1}| \tag{7}$$

Where P_i is currently pixel while the $P_{i\text{-}1}$ Is the neighbor pixel. The extracted features are used for future matching. This method enhances fingerprint verification reliability by matching only relevant data. Fig. 8 shows feature extraction of the biometric image.

(a) (b)

Fig. 8. The feature extraction (**a**) thinning edge image, and (**b**) feature extraction of fingerprint image

4.2. LOG-IN PHASE

In this phase, the fingerprint image of the user is processed in the same way that the user's fingerprint is processed in the registration phase of the system, including preprocessing and feature extraction.

4.3. IDENTIFICATION PHASE

In this phase, the feature extraction obtained from the registration phase is used to train and test the system in order to identify the user in the log-in phase.

4.3.1. FEATURE MATCHING

In this phase, the extracted features of the two fingerprints are compared using a similarity score to verify if the two fingerprints match and belong to the same user. The Orientation-based Matching (OM) technique was employed to address the issue of misorientation fingerprint images resulting from inadequate image acquisition, which leads to increased matching accuracy. The feature matching procedure starts by computing the similarity of the features extracted from the two fingerprint images by calculating the Euclidean distance and the absolute orientation difference between the features extracted from the two fingerprint images to confirm the feature identity. Eq. (8) and Eq. (9) show the Euclidean distance and the absolute orientation difference.

$$distance = \sqrt{x^2 + y^2} \tag{8}$$

$$\theta = |\theta_1 - \theta_2| * \frac{180}{\pi} \tag{9}$$

Finally, in the feature-matching procedure, the ultimate similarity score is computed based on the number of matched features and their respective distances, employing Eq. (10).

$$score = \sqrt{\frac{2n^2}{f_1^2 + f_2^2}} \tag{10}$$

Where n is the number of matched features, $f1$ and $f2$ are the total number of features in the two fingerprints being compared.

4.4. AUTHENTICATION PHASE

This phase includes generating random numbers using a Six-Dimensional chaotic system, processing the number generator to check if the numbers satisfy the RAS condition, generating keys (private key and public key) for both the server and client, storing them in the server, and using RSA to authenticate the user.

4.4.1. RANDOM NUMBER GENERATOR

The 6D hyperchaotic system mentioned above consists of six equations, Eq. (1), Eq. (2), Eq. (3), Eq. (4), Eq. (5), and Eq. (6), containing variables and parameters. The state variables are $x1, x2, x3, x4, x5, x6$, and $\alpha, \beta, a, b, c, d, e, r$, and k are system parameters. When initial points ($x1(0)$, $x2(0)$, $x3(0)$, $x4(0)$, $x5(0)$, and $x6(0)$) = (0.0512, 0.0011, 0.2023, 1.0054, 0.0075, and 0.5001) Consequently, and under α = 15.4778, β = 0.1291, a = 4.8234, b =1.8789, c = 8.5012, d = -0.0154, e= 1.0001, f = -0.0137, r = 0.1723, and k= 0.0019. The number of times in this system is employed is dictated by the need to determine the values of the state variables. This approach enables the server to generate a collection of random numbers for each user, which is utilized to create private and public keys and the session key for each user.

4.4.2. KEY GENERATION

The RSA algorithm utilizes integers from the preceding stage to fulfill the specifications. The numbers are multiplied by 104 to obtain integer components that meet the RSA requirements. The numbers are assessed for primality and incremented if they are not prime. In every cycle, six integers are generated from a 6D hyperchaotic map, including three for the server and three for the client. The initial prime number is P, the subsequent is Q, and the third is the private key (E). The public key (D) has been established. If a number does not satisfy the criteria, the six numbers are eliminated, and the identical method is applied to the server and client numbers.

The following steps briefly describe the key generation

Step1: use 6DHyper Chaotic System to generate (X_1, X_2, X_3, X_4, X_5, X_6)

Step 2: Multiply (X_1, X_2, X_3, X_4, X_5, X_6) by a power of ten, such as (104), to get an integer value

Step 3: All numbers are tested to be prime; if not prime, increment them by 1 until they reach the nearest prime.

Step 4: Set the result from the (X_1, X_2, X_3) as (P_s, Q_s, E_s) for the server. (P_s, Q_s are the two primes for the server, E_s is the private key for the server)

Set the result from the (X_5, X_6, X_4) as (P_c, Q_c, E_c) for the client. (P_c, Q_c are the two primes for the client, Ec is the private key for the client)

Step 5: $N_s = P_s \times Q_s$ (N_s: N for server)

$N_c = P_c \times Q_c$ (N_c: N for client)

$\varphi(N_s)=(P_s\text{-}1)\,(Q_s\text{-}1)$

$\varphi(N_c)=(P_c\text{-}1)\,(Q_c\text{-}1)$

Step 6: Find the P_s (public key for server) and P_c (public key for client)

$(E_s * D_s) \bmod \varphi(N_s)=1$

$(E_c * D_c) \bmod \varphi(N_c)=1$

4.4.3. INITIALIZE RSA

The private and public keys produced in the preceding stage are employed to initiate the RSA algorithm, where (E_s, N_s) represents the server's private key, and (D_s, N_s) denotes the server's public key. (E_c, N_c) constitutes the client's private key, while (D_c, N_c) represents the client's public key. The process occurs on the server to verify the client's legitimacy by recognizing and verifying the client's ID. Thus, both endpoints of the connection are authenticated, ensuring that any variations in these keys will render the connection between source and destination insecure. The efficacy of the encryption and decryption process is evaluated to confirm the authenticity and integrity of the resultant keys, hence ensuring the dependability of the outcomes in practical applications such as security systems and encrypted communications.

5. EXPERIMENTAL RESULTS

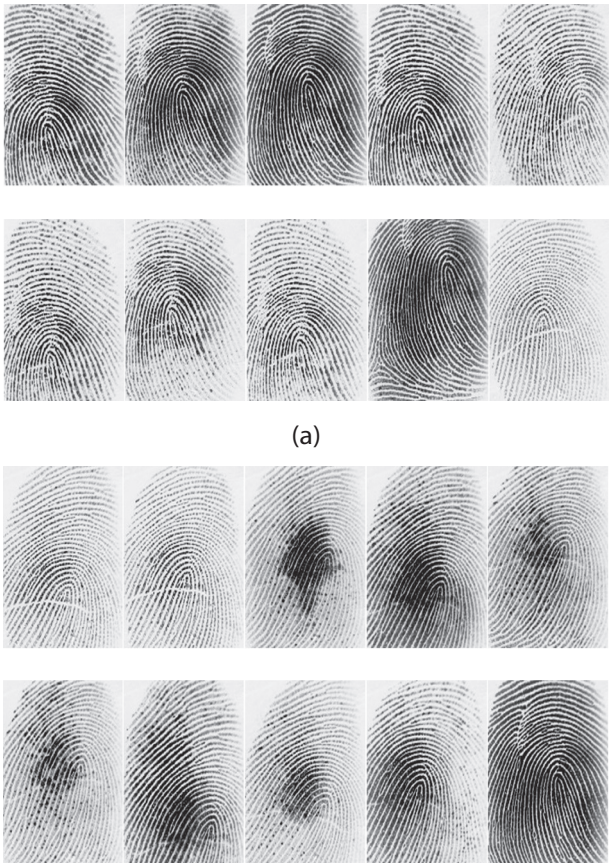
The experimental result for the proposed method is implemented and analyzed in MATLAB 2023b. This method utilized an Intel(R) Core (TM) i7-10510U CPU @ 1.80 GHz, 2.30 GHz processor, 16.0 GB (15.7 GB usable), and Windows 11 Home, 64-bit operating system, x64-based processor.

5.1. BIOMETRIC DATASET

Two datasets of fingerprint images are used to train and test the proposed method. The first one is Neurotechnology CrossMatch, which contains 408 fingerprint images for 51 fingers, with 8 impressions taken per finger. These images were captured using an optical scanner at 500 dpi with dimensions of 504×480 pixels. Those images are saved in TIFF format. Fig. 9 shows the fingerprint image of the Neurotechnology CrossMatch dataset.

Fig. 9. The Fingerprint image of the Neurotechnology CrossMatch dataset.

The second dataset is the local dataset, which consists of biometric images of the users' fingerprints. These biometric images are employed to authenticate the client. The fingerprint images are collected using a ZK9500 USB Fingerprint Scanner with a 280 MHz CPU, the fingerprint image quality is 2 megapixels and 500 dpi resolution. This dataset includes 1200 images of thumb fingerprints for 60 users; each user has 10 fingerprint images for the right hand and 10 fingerprints for the left hand. Fig. 10 shows the fingerprint image of the local dataset.



(a)

(b)

Fig. 10. Fingerprint images of the local dataset: (**a**) for the left thumb, and (**b**) for the right thumb

5.2. FINGERPRINT MATCHING RESULT

The Neurotechnology CrossMatch and local datasets improved as the number of images used in the experiment increased, but the local dataset had much superior accuracy in any scenario. The Neurotechnology CrossMatch dataset had 90.82% accuracy with 10 images, whereas the local dataset had 91.58%, demonstrating a minor advantage at low image counts. With 40 images, the Neurotechnology CrossMatch dataset's accuracy increased to 92.90%, while the local dataset achieved 93.08%, demonstrating its stability and capacity to gain from more data. The local dataset performed better at 50 images, scoring 95.14% against 92.24% for the Neurotechnology CrossMatch dataset, suggesting data adaptability as shown in Table 1.

Table 1. Accuracy results of fingerprint matching

Number of images	10	20	30	40	50
Neurotechnology CrossMatch	90.82	91.77	91.80	92.90	92.24
Local dataset	91.58	92.79	93.05	93.08	95.14

Fig. 11 shows that the local dataset surpasses the Neurotechnology CrossMatch dataset. When the number of images is limited, the disparity between the two datasets is negligible; however, when the number of images increases, the Local dataset acquires a distinct accuracy. This may be due to the local dataset aligning more closely with the test or the superior image analysis techniques.

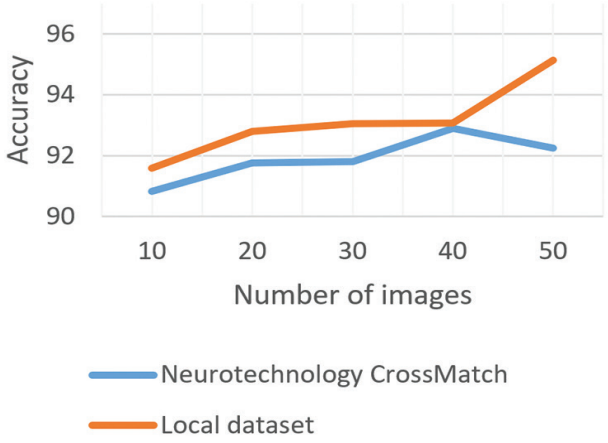
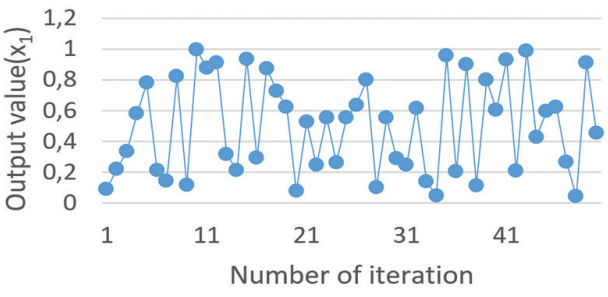


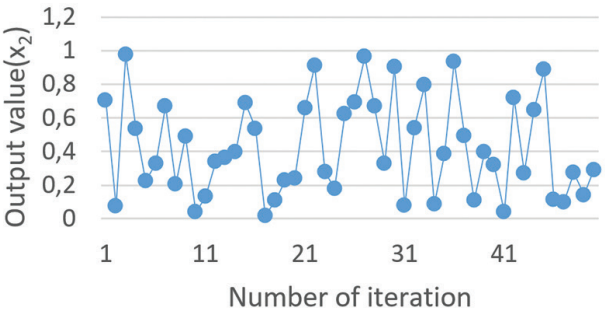
Fig. 11. Accuracy result of fingerprint matching

5.3. RANDOM NUMBER GENERATOR RESULT

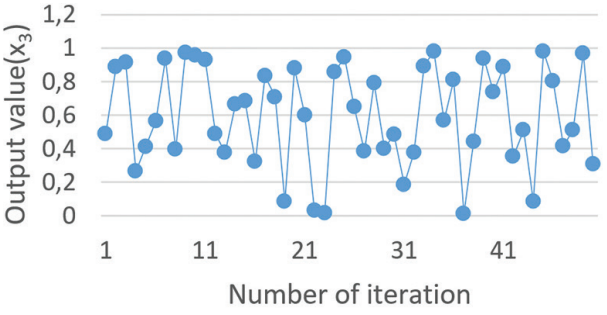
The 6D hyperchaotic system was employed to model the dynamics of a nonlinear system through mathematical equations reliant on interdependent and iterative variables. The system underwent testing through 12 computation cycles, during which each of the six primary variables ($x1$ to $x6$) exhibited complex periodic variations, indicative of the nonlinear interactions among the mathematical equations. The modulo function was utilized to normalize values inside the interval [0,1]. Fig. 12 explains the plotting of the six-dimensional sequences to visualize the randomness distribution.



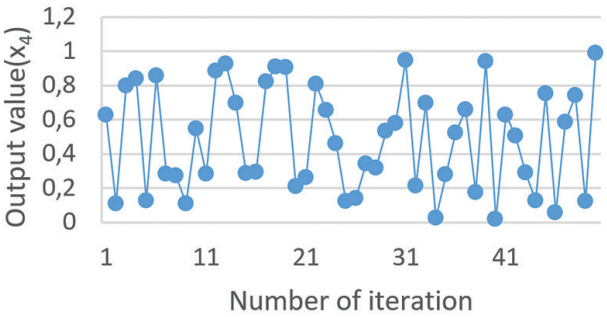
(a)



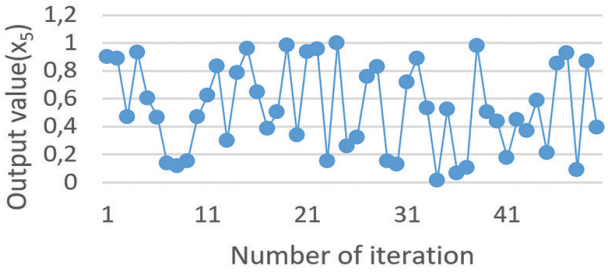
(b)



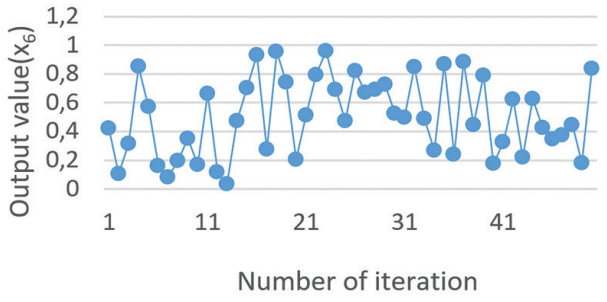
(c)



(d)



(e)



(f)

Fig. 12. The plotting of Six-Dimensional sequences (**a**) the first sequence (**b**) the second sequence (**c**) the third sequence (**d**) the fourth sequence (**e**) the fifth sequence, and (**f**) the sixth sequence

Table 2 shows that the values vary dynamically across twelve cycles, illustrating the system's nonlinear characteristics. Recursive functions and nonlinear equations significantly influence the ultimate values of each variable, and potential disorder in the values can be detected, rendering the application appropriate for the analysis of chaotic or cryptographic systems. This system demonstrated computation efficiency, completing the iterative process in a minimal duration, hence suggesting its appropriateness for time-sensitive applications. The results indicated that the system produces dynamic values that progress over time, influenced by the initial values and input constants. The mentioned values of the variables ($x1$ to $x6$) were processed and examined to transform them into prime numbers.

The values are multiplied by 10^4 to transform them into integers while preserving their precision. Subsequently, the primacy of each element is assessed. If the number is not prime, an iterative procedure is employed that increments the value by 1 until the first prime number is attained. The resultant prime numbers are retained.

Table 2. Random number generator results

iteration	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$
1	0.2310	0.4324	0.0936	0.0127	0.0088	0.0011
2	0.8058	0.0012	0.4875	0.4326	0.0398	0.4324
3	0.9037	0.4973	0.8647	0.0071	0.1389	0.0012
4	0.0554	0.6104	0.5146	0.4975	0.1560	0.4973
5	0.1079	0.6078	0.1010	0.6172	0.0098	0.6104
6	0.1864	0.9470	0.2143	0.6163	0.0186	0.6078
7	0.4061	0.5782	0.3155	0.9555	0.0322	0.9470
8	0.8142	0.6213	0.9279	0.5914	0.0700	0.5782
9	0.8252	0.2455	0.8669	0.6294	0.1404	0.6213
10	0.4751	0.8743	0.2313	0.2541	0.1424	0.2455
11	0.3955	0.0852	0.6670	0.8779	0.0821	0.8743
12	0.0768	0.7093	0.8995	0.0972	0.0683	0.0852

The results presented in Table 3 demonstrate exceptional efficiency in the rapid and precise conversion of values to prime numbers, rendering them appropriate for applications necessitating numerical data analysis or the utilization of prime numbers, such as cryptography and the examination of dynamic systems.

Table 3. Prime number result

Iteration	X1	X2	X3	X4	X5	X6
1	2310	4324	936	127	88	11
2	8058	12	4875	4326	398	4324
3	9037	4973	8647	71	1389	12
4	554	6104	5146	4974	1560	4973
5	1079	6078	1010	6172	98	6104
6	1864	9470	2143	6163	186	6078
7	4061	5782	3155	9555	322	9470
8	8142	6213	9279	5914	700	5782
9	8252	2455	8669	6294	1404	6213
10	4751	8743	2313	2541	1424	2455
11	3955	852	6670	8779	821	8743
12	768	7093	8995	972	683	852

The National Institute of Standards and Technology (NIST) has 16 statistical tests that are widely used to analyze bit sequence randomness. The statistical significance level for each NIST test has been established at 0.01. Hence, Sequences pass a test if the P value is greater than 0.01 and less than 1. In the numerical experiment of the proposed method, 1000 sets of 106-bit sequences are created and randomly selected for NIST testing. The results in Table 4 showed high randomness, with all tests passing and P values exceeding all thresholds. The recommended method improved randomization statistical performance, with higher P-value averages in some tests than in reference [21]. These findings make random sequence generation suitable for encryption, cybersecurity, and statistical modeling that require high-quality random numbers.

Table 4. The P-value of the frequency test

Test type	Average of the p-value of the proposed method	Average of the p-value of the reference [21]	Status
Frequency test	0.488	0.663	success
Frequency within a block	0.290	0.528	success
Run Test	0.778	0.654	success
The longest run of ones in a block test	0.752	0.744	success
Fast Fourier Transform test	0.453267	0.666	success
Overlapping template matching test	0.200	0.412	success
approximate entropy test	0.513506	0.954	success
Cumulative sum test	0.489497	0.450	success

5.4. KEY GENERATION RESULT

The fundamental coefficients associated with prime numbers and their features are computed for utilization in cryptographic applications. The variable value from Table 3 is used to determine Ns and Nc values, where X_1 and X_2 represent the prime numbers for the server (P_s and Q_s), and X_3 represents the private key for the server (E_s). X_5 and X_6 represent the prime numbers for the client (P_c and Q_c), and X_4 represents the private key for the client (E_c). $N_s = P_s \times Q_s$ and $N_c = P_c \times Q_c$. The Euler function (φ) corresponding to each of N_s and N_c is computed using the formulas $\varphi(N_s)=(P_s\text{-}1)\ (Q_s\text{-}1)$ and $\varphi(N_c)=(P_c\text{-}1)\ (Q_c\text{-}1)$, which emphasizes the need to develop mathematical models utilized in cryptography. The values of E_s and E_c are evaluated for compatibility with the Euler function using the Greatest Common Divisor (GCD), while the equivalent numbers D_s and D_c are ascertained through iterative calculations to provide the encryption keys. The conclusive results are recorded as fundamental coefficients for the server P_s, Q_s, E_s, and D_s. In contrast, the coefficients for the client are P_c, Q_c, E_c, and D_c for five users, as shown in Table 5.

Table 5. Public and private key results

user	Ps	Qs	Es	Ds	Pc	Qc	Ec	Dc
1	2311	4327	937	3775393	89	11	127	783
2	8059	11	4877	19133	401	4327	4327	592663
3	9041	4973	8647	40372663	1399	11	71	12011
4	557	6113	5147	3373843	1559	4973	4987	5385339
5	1867	9473	2143	7571359	191	6079	6163	455707
6	4073	5783	3163	11485571	331	9473	9587	714683
7	8147	6217	9281	40476785	701	5783	5923	890387
8	8263	2459	8669	5973629	1409	6217	6299	947603
9	4751	8747	2333	1798497	1427	2459	2543	3320411
10	3967	853	6673	2017393	821	8747	8779	3620579

The experiment demonstrated that the model delivers great precision in calculating these fundamental coefficients, rendering it an efficient instrument for creating public and private keys utilized in safe encryption schemes, such as RSA. The execution time is justifiable given the many procedures involved, which improve the program's efficiency and applicability in the real world, as shown in Table 6.

Table 6. Number generation and checking the RSA parameter time

Iteration	number generation time (msec)	Check the RSA parameter time (msec)
1	0.1183	0.0122
2	0.0383	0.0094
3	0.0549	0.0104
4	0.0911	0.0131
5	0.0630	0.0141
6	0.0894	0.0089
7	0.0647	0.0178
9	0.0645	0.0155
10	0.0919	0.0098
11	0.0262	0.0109
12	0.0868	0.0145
Average	**0.0717**	**0.01138**

From the previous table, the Average executing time for the generation of random numbers through 6D Hyperchaotic Systems, followed by testing and parameter generation to satisfy RSA criteria, is 0.08308 msec. The Shannon entropy value measures bit randomness and determines the authentication protocol's statistical attack resistance. A key with a greater entropy value is more difficult to analyze and crack. Table 7 shows the entropy of the key's size.

Table 7 shows similar and high Shannon entropy values, approaching 8, which indicates that they all provide keys with robust resistance to statistical attacks.

Table 7. The entropy of the size of the key

Size of keys (bits)	Entropy
800	7.301877
1024	7.350909
1600	7.394059
2048	7.45548
3200	7.609392
4096	7.417282
5000	7.665345
6000	7.539276
7000	7.649151
8192	7.749044

5.5. AUTHENTICATION USING RSA RESULT

RSA is one of the earliest public-key cryptosystems founded on the so-called "factoring problem". It remains the most extensively utilized system in practice. The verification method involves comparing the original user input ID with the final values obtained after completing all encryption and decryption phases. The procedure commences with the encryption of the original user input ID utilizing the client's private key (E_c, N_c) to generate the value En_1. Subsequently, En_1 is encrypted with the server's public key (D_s, N_s) to yield the value En_2.

$$En_1 = ID^{Ec} \mod N_c$$
$$En_2 = En_1^{Ds} \mod N_s$$

The server decrypts En_2 using its private key (E_s, N_s) to obtain En_3, which is decrypted by using the public key of the client (D_c, N_c) to yield the value En_4.

$$En_3 = En_2^{Es} \mod N_s$$
$$En_4 = En_2^{Dc} \mod N_c$$

After this series, the procedure's accuracy is confirmed by comparing the retrieved value En4 with the original value of the user ID, as shown in Table 8 for six users.

Table 8. Authentication using RSA result

User	ID	En_1	En_2	En_3	En_4	P_s	Q_s	E_s	D_s	P_c	Q_c	E_c	D_c	Time (sec)
1	141	190	276643	190	141	2311	4327	937	3775393	89	11	127	783	0.0708
2	106	14813	26434939	14813	106	9041	4973	8647	40372663	1399	11	71	12011	0.6483
3	132	944659	15799198	944659	132	1867	9473	2143	7571359	191	6079	6163	455707	0.1363
4	104	816440	20802234	816440	104	4073	5783	3163	11485571	331	9473	9587	714683	0.2067
5	114	2037635	19253650	2037635	114	8147	6217	9281	40476785	701	5783	5923	890387	0.6653
6	127	1143780	6118202	1143780	127	8263	2459	8669	5973629	1409	6217	6299	947603	0.1198

If the two values (ID and En_4) are the same, the process is deemed successful, and the user is authorized to proceed; however, if they are not identical, the user is unauthorized to proceed. The last column in Table 8 shows the authentication time for each user. The results demonstrated significant efficacy in maintaining verification accuracy and data integrity, establishing it as a dependable instrument for encryption applications necessitating elevated security and verification standards, including key management in banking and cybersecurity systems.

Table 9. Proposed RSA time

Reference	Method	Length of prime	key Generation Times (msec)
[23]	RSA	64	20.27
		128	25.47
		256	40.50
		512	76.55
		1024	820.14
[23]	CRPKC-Ki	64	35.02
		128	54.19
		256	105.99
		512	177.18
		1024	1250.09
	Proposed	64	16.96
		128	20.53
		256	36.38
		512	64.56
		1024	680.88

Table 9 presents a comparison of key generation times across several key sizes, including 64, 128, 256, 512, and 1024, as well as the encryption and decryption times for the proposed method, RSA [23], and CRPKC-Ki [23]. RSA [23] and CRPKC-Ki [23] are implemented and analyzed in SageMath Jupyter Notebook. SageMath is a Python-based software that offers mathematical computing, data analysis, graph drawing, and programming capabilities on several operating systems. The proposed method is implemented and analyzed using an 11th Gen Intel® Core™ i5-11320H @ 3.20 GHz processor, 16.0 GB RAM (15.8 GB accessible), and Windows 10, 64-bit operating system. By observing the key generation time in Table 9 and Fig. 13, it is obvious that CRPKC-ki has the longest key generation time, while the proposed method has a little different key generation time from RSA. The time is acceptable considering enhanced security.

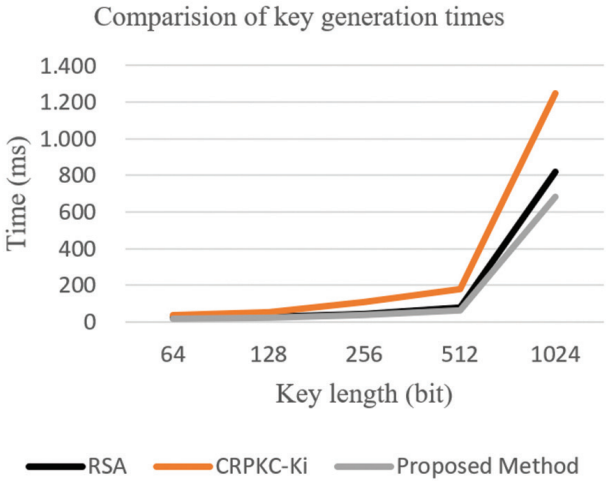


Fig. 13. Key generation times comparison of various key sizes

6. CONCLUSION

In this paper, an attribute-based authentication (ABA) system is presented that integrates biometric fingerprint recognition, the RSA algorithm, and a 6D hyperchaotic system. The proposed method exhibited robust outcomes for security and efficiency, achieving a fingerprint matching accuracy of 95.14% and an average verification time of 0.3078 seconds, so illustrating the system's efficacy in settings necessitating rapid and secure performance Random encryption keys were created using a chaotic system, giving a Shannon entropy value of 7.7490, indicating strong resilience to statistical and brute force attacks. The result using the NIST test showed high randomness, with all tests passing and P values exceeding all thresholds. In comparison to other RSA-based encryption systems like CRPKC-Ki, the proposed system exhibited a reduced key generation

time while preserving a high security level, illustrating that the integration of chaotic systems with conventional encryption algorithms improves performance and robustness against threats.

The recommended method improved randomization statistical performance, and then the client attribute (user ID) is encrypted and decrypted using the RSA algorithm to ensure authentication and protection against various attacks. This approach aims to achieve strong and effective authentication in terms of security and computational efficiency, making it suitable for applications that require high levels of protection, such as sensitive banking and information systems. The security analysis indicates that the algorithm benefits from the security advantages of both 6D hyperchaotic systems and RSA, making it robust against typical RSA cipher attacks.

7. REFERENCES

[1] Z. Leyou, W. Jun, M. Yi, "Secure and Privacy-Preserving Attribute-Based Sharing Framework in Vehicles Ad Hoc Networks", IEEE Access, Vol. 8, 2022, pp. 116781-116795.

[2] M. M. Nayyef, A. M. Sagheer, S. S. Hamad, "Attribute Based Authentication System using Homomorphic Encryption", Journal of Engineering and Applied Sciences, Vol. 13, No. 10, 2018, pp. 352-355.

[3] S. F. Tan, G. C. Chung, "Attribute-Based Encryption in Securing Big Data from Post-Quantum Perspective: A Survey", Cryptography, Vol. 6, No. 3, 2022, pp. 40-55.

[4] X. Liu, W.-B. Lee, Q.-A. Bui, C.-C. Lin, H.-L. Wu, "Biometrics-based RSA cryptosystem for securing real-time communication", Sustainability, Vol. 10, No. 10, 2018, pp. 3588-3603.

[5] A. T. Lo'ai, G. Saldamli, "Reconsidering big data security and privacy in cloud and mobile cloud systems", Journal of King Saud University Computer and Information Sciences, Vol. 33, No. 7, 2021, pp. 810-819.

[6] P. Matta, M. Arora, D. Sharma, "A comparative survey on data encryption Techniques: Big data perspective", Materials today: Proceedings, Vol. 46, 2021, pp. 11035-11039.

[7] R. Ryu, S. Yeom, D. Herbert, J. Dermoudy, "The design and evaluation of adaptive biometric authentication systems: Current status, challenges and future direction", ICT Express, Vol. 9, No. 6, 2023, pp. 1183-1197.

[8] V. Vekariya, M. Joshi, S. Dikshit, "Multi-biometric fusion for enhanced human authentication in information security", Measurement Sensors, Vol. 31, 2024, pp. 100973-100981.

[9] D. Osorio-Roig, L. J. González-Soler, C. Rathgeb, C. Busch, "Privacy-preserving multi-biometric indexing based on frequent binary patterns", IEEE Transaction on Information Forensics and Secuity, Vol. 19, 2024, pp. 4835-4850.

[10] H. Djebli, S. Ait-Aoudia, D. Michelucci, "Quantized random projections of SIFT features for cancelable fingerprints", Multimedia Tools and Applications, Vol. 82, No. 5, 2023, pp. 7917-7937.

[11] S. Bakheet, S. Alsubai, A. Alqahtani, A. Binbusayyis, "Robust fingerprint minutiae extraction and matching based on improved SIFT features", Applied Sciences, Vol. 12, No. 12, 2022, p. 6122.

[12] Z. Zhang, S. Liu, M. Liu, "A multi-task fully deep convolutional neural network for contactless fingerprint minutiae extraction", Pattern Recognition, Vol. 120, 2021, pp. 108189-108201.

[13] S. O. Husain, R. A. Reddy, P. K. Pareek, S. Jagannathan, M. Jyothi, "Robust Fingerprint Minutiae Extraction and Matching Using Fully Connected Deep Convolutional Neural Network and Improved SIFT", Proceeding of the International Conference on Data Science and Network Security, Tiptur, India, 26-27 July 2024, pp. 1-5.

[14] M. Siddiqui, S. Iqbal, B. AlHaqbani, B. AlShammari, T. Khan, I. Razzak, "A Robust Algorithm for Contactless Fingerprint Enhancement and Matching", Proceeding of the International Conference on Digital Image Computing: Techniques and Applications, Perth, Australia, 27-29 November 2024, pp. 214-220.

[15] K. Al-Mannai, E. Bentafat, S. Bakiras, J. Schneider, "Secure Biometric Verification in the Presence of Malicious Adversaries", IEEE Access, Vol. 13, 2024, pp. 5284-5295.

[16] A. H. Y. Mohammed, R. A. Dziyauddin, L. A. Latiff, "Current multi-factor of authentication: Approaches, requirements, attacks and challenges",

International Journal of Advanced Computer Science and Applications, Vol. 14, 2023, No. 1, pp. 166-178.

[17] O. F. Boyraz, E. Guleryuz, A. Akgul, M. Z. Yildiz, H. E. Kiran, J. Ahmad, "A novel security and authentication method for infrared medical image with discrete time chaotic systems", Optik, Vol. 267, 2022, pp. 169717-169735.

[18] Y.-Y. Fanjiang, C.-C. Lee, Y.-T. Du, S.-J. Horng, "Palm vein recognition based on convolutional neural network", Informatica, Vol. 32, No. 4, 2021, pp. 687-708.

[19] I. K. V S. Socheat, T. Wang, "Fingerprint enhancement, minutiae extraction and matching techniques", Journal of Computer and Communications, Vol. 8, No. 5, 2020, pp. 55-74.

[20] B. H. Situmorang, "Identification of Biometrics Using Fingerprint Minutiae Extraction Based on Crossing Number Method", Komputasi Journal Ilmiah Ilmu Komputer dan Mathmatika, Vol. 20, No. 1, 2022, pp. 71-80.

[21] O. Z. Akif, S. Ali, R. S. Ali, A. K. Farhan, "A new pseudorandom bits generator based on a 2D-chaotic system and diffusion property", Bulletin Electrical Engineering and Informatics, Vol. 10, No. 3,2021, pp. 1580-1588.

[22] N. Tahat, A. A. Tahat, M. Abu-Dalu, R. B. Albadarneh, A. E. Abdallah, O. M. Al-Hazaimeh, "A new RSA public key encryption scheme with chaotic maps", International Journal Electrical and Computer Engineering, Vol. 10, No. 2, 2020, pp. 1430-1437.

[23] C. Zhang, Y. Liang, A. Tavares, L. Wang, T. Gomes, S. Pinto, "An Improved Public Key Cryptographic Algorithm Based on Chebyshev Polynomials and RSA", Symmetry, Vol. 16, No. 3, 2024, pp. 263-278.

[24] F. Yu et al. "Chaos-Based Engineering Applications with a 6D Memristive Multistable Hyperchaotic System and a 2D SF-SIMM Hyperchaotic Map", Complexity, Vol. 2021, No. 1, 2021, pp. 6683284-6683305.

[25] B. A. Mezatio, M. T. Motchongom, B. R. W. Tekam, R. Kengne, R. Tchitnga, A. Fomethe, "A novel memristive 6D hyperchaotic autonomous system with hidden extreme multistability", Chaos, Solitons & Fractals, Vol. 120, 2019, pp. 100-115.

Automation of Surgical Instruments for Robotic Surgery: Automated Suturing Using the EndoStitch Forceps

Original Scientific Paper

Omaira Tapias*
Popular University of Cesar, Faculty of Engineering,
Valledupar, Cesar, 200003, Colombia,
omairatapias@unicesar.edu.co

Andrés Vivas
University of Cauca,
Popáyan, Cauca, 190003, Colombia,
avivas@unicauca.edu.co

Juan Carlos Fraile
University of Valladolid,
Valladolid, 47011, Spain,
jcfraile@uva.es

*Corresponding author

Abstract *– Suturing in minimally invasive robotic surgery poses significant challenges in terms of technique, du- ration, and tool usage. This study presents an innovative semi-autonomous robotic system for suture automation in minimally invasive surgery. The system integrates motorized EndoStitch forceps with a UR3 collaborative robot, combining robotic dexterity with the functionality of a proven surgical tool. The de- sign of the motorized gripper coupling, the development of a modular ROS-based control architecture, and the implementation of a library of parameterized movements optimized for suturing are described. Exper- imental results, obtained in tissue simulations, demonstrate micrometer accuracy in needle positioning. The variability of the positioning error and its relation to the characteristics of the surgical environment are analyzed. This system represents a representative advance towards reducing the surgeon's cognitive load, improving accuracy and efficiency in robotic suturing, and opening new avenues for safer and more consistent surgical procedures. The semi-autonomous robotic suturing system demonstrated micrometer- level precision in tests. Mean positioning errors ranged from* 0.5×10^{-5} *m to* 2.0×10^{-5} *m, with low standard deviations (highest at* 0.70×10^{-5} *m) indicating high repeatability.*

Keywords: *Robotic Suturing, Surgical Automation, EndoStitch Forceps, UR3 Collaborative Robot, Minimally Invasive Surgery, Semi-Autonomous Suturing*

Received: Received: March 21, 2025; Received in revised form: June 24, 2025; Accepted: June 25, 2025

1. INTRODUCTION

Minimally invasive surgery (MIS) has revolu- tionized modern medicine, offering patients faster recovery times and reduced scarring compared to traditional open procedures. Despite these ad- vances, certain maneuvers, such as suturing, re- main particularly challenging due to the dexter- ity, precision, and time they demand, especially in the context of robotic surgery where, although er- gonomics and vision are improved, teleoperation of complex sutures remains a demanding task. To address these limitations, the automation of surgi- cal instruments is emerging as a promising field, seeking to improve efficiency, consistency, and safety in the operating room. In this context, the EndoStitch forceps, a tool widely used and proven in laparoscopic surgery for its ingenious needle- passing mechanism, presents an excellent oppor- tunity for automation.

While robotic surgery has enabled many ad- vances, studies such as [1] indicate that robotic consoles can be mentally demanding for novice surgeons, requiring formal training. Challenges persist, such as ergonomics, training optimization, and cost-benefit evaluation. Therefore, research is justified in adapting and automat-

ing specialized forceps that can improve not only the procedure itself but also the surgeon's work [2],[3][4].

This work presents an innovative semiautonomous robotic system for suture automation in MIS. The system integrates a motorized EndoS- titch forceps, whose needle manipulation mecha- nism has been motorized, with a UR3 collabora- tive robot. The main objective is to develop a sys- tem capable of performing sutures with high pre- cision and repeatability, reducing human variabil- ity and improving efficiency in minimally invasive suturing procedures. The design of the motorized forceps coupling, the development of a modu- lar ROS-based control architecture, and the imple- mentation of a library of parameterized movements optimized for suturing are described.

The **main contributions** of this work include:

- A semi-autonomous robotic system for automated suturing, integrating an EndoStitch forceps with a motorized actuation mechanism and a collaborative robot, enabling automatic needle passing between its jaws.
- The potential improvement in suturing precision and consistency, which could translate to reduced human variability and optimized surgical time, thereby decreasing the surgeon's cognitive load by allowing them to focus on higher-level aspects of the intervention.
- A modular and adaptable ROS-based control platform with a library of parameterized movements, demonstrating the concept's viability with micrometer-level precision in simulators and facilitating future research and the integration of advanced functionalities such as vision guidance.

This paper is organized as follows: Section 2 (Methodology) details the methodologies employed, including the finite element simulation of the needle-tissue interaction, the description of the robotic platform (UR3 robot and motorized EndoStitch forceps), and the ROS-based software control architecture. Section 3 (Results) presents the experimental results of the system's evaluation in a simulated environment, analyzing positioning accuracy. Section 4 (Discussion) discusses the findings, study limitations, and implications of the work in the field of surgical automation. Finally, Section 5 (Conclusion) concludes the paper by summarizing the main contributions and outlin- ing future research directions.

1.1. INTERACTION MODEL BETWEEN STRAIGHT NEEDLE AND TISSUE IN SUTURING

Modeling the interaction between a straight needle and biological tissue is a critical aspect in the design and optimization of suturing procedures, particularly within the context of automated suturing performed by robotic systems employing EndoStitch-type forceps tools.

The intrinsic properties of the tissue, such as its elasticity, viscoelasticity, and stress-strain behavior, are determinant factors in the biological material's response to straight needle insertion. These properties, which can vary considerably among different tissue types and even within the same tissue (exhibiting heterogeneity and anisotropy), dictate how the tissue deforms, resists penetration, and recovers its original shape, thereby directly influencing the interaction forces.

1.1.1. Force analysis in linear needle penetration

For the specific case of linear penetration, as executed by a straight needle, the force analysis focuses on identifying and quantifying the components that oppose the needle's advancement. Cheng et al. [5] propose a model where the primary force restricting needle penetration is damped viscous friction. This force is a function of penetration velocity and is influenced by both static and kinetic friction.

The proposed friction force model is initially decomposed into static and kinetic friction components, as shown in Equation (1):

$$f_{friccion}(v, z) = f_s(z) + f_k(v, z) = \mu_s N + \mu_k N \quad (1)$$

Where:

- $f_{friccion}(v, z)$ is the total friction force, dependent on the penetration velocity v and insertion depth z.
- $f_s(z)$ is the static friction component, which must be overcome to initiate or continue motion at low velocities.
- $f_k(v, z)$ is the kinetic friction component, active once the needle is in motion.
- μ_s and μ_k are the static and kinetic friction coefficients, respectively, which depend on the properties of the needle-tissue interface.
- N is the normal force acting on the needle shaft, perpendicular to the penetration direction. This normal force is generated by the compression and displacement of the tissue surrounding the needle.

To more explicitly incorporate the dependence on penetration velocity, the model is refined as shown in Equation (2):

$$f_{friccion}(v, z) = (\mu_s + \eta_v)N \quad (2)$$

In this formulation, η represents a viscous damping constant, which models how frictional resistance increases with penetration velocity v.

While the cutting force ($F_{cutting}$) exerted by the sharp needle tip to initiate tissue separation is crucial, especially at the onset of penetration, the model by Cheng et al. [5] emphasizes the dominant contribution of friction forces along the needle shaft once it has penetrated to a certain depth. Unlike curved needles, where a sweeping force (F_{sweep}) can arise due to internal tissue compres- sion from the circular motion, this effect is

less pronounced in the linear penetration of a straight needle, or it manifests primarily as the normal force N that contributes to friction. Developing sensors capable of detecting the contact force of a needle presents a significant technical challenge; however, recent research has explored the use of piezoelectric materials in the surgical field for this purpose [6], [7], [4] .

Currently, commercially available sensors suit- able for placement on surgical needles are lacking due to constraints related to size and integration with surgical tools. Therefore, this work presents a simulation developed in Ansys to analyze this aspect.

1.2. RELATED WORK

The automation of surgical instruments in robotic surgery has experienced a significant surge in recent years, driven by the pursuit of more precise, efficient, and less invasive procedures. Various studies have addressed the motorization of existing surgical instruments and their integration with robotic platforms, with the aim of reducing human variability and improving clinical outcomes. In the specific context of suturing, a fundamental task in numerous surgical interventions, different approaches have been explored to automate the process, from fully autonomous systems to semiautomatic solutions that assist the surgeon in specific tasks.

The following describes some relevant works that address similar aspects:

The study presented in [8] investigates the automation of a suturing device for minimally invasive surgery, based on the EndoStitch, optimizing it with a DC motor controlled by a button to allow for single-handed suturing (see Fig. 1 (c)). The conventional EndoStitch was compared to the automated version in an experiment with 20 surgeons of varying levels of laparoscopic experience, measuring suturing time and accuracy.

While in [8] they also automated the EndoStitch forceps, their approach was limited to a motorized drive with manual control (button) to facilitate single-handed suturing. Unlike their design, our work extends the automation of the EndoStitch by integrating it with a collaborative robot, opening the door to more precise and programmable control of the suturing process.

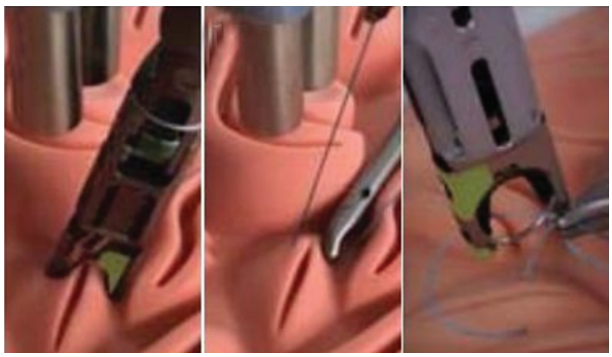
In the work by [9], a suturing robot for singleport microsurgical endoscopic surgery is presented, focused on a flexible instrument with a suturing probe and a custom joint, along with a novel mechanism for needle driving and locking. For its teleoperated control, the robot's kinematics are provided. Continuous stitch and knotting experiments prove the robot's feasibility in confined spaces; however, the limited ability to suture certain wound orientations, the needle size, the ma- neuverability in confined spaces, and the thread gripping properties impose limitations. The instrument is based on the EndoStitch forceps but adds three degrees of freedom to the end effector. Unlike the present work, the instrument they use is not yet commercially available (see Fig. 1 (d)).

In the work presented in [10], a platform was developed that combines the real handle of the instrument with a three-dimensional simulation of its tip, allowing the user to practice needle transfer and thread handling. The simulation demonstrated that the current EndoStitch interface can be difficult to master and that its automation could facilitate learning and use in minimally invasive surgery, reducing errors and training times. Unlike our work, they present a completely automated procedure.

The work by [11] presents an autonomous la- paroscopic robotic system for suturing (STAR), with the motorization of the Endo 360 tool (see Fig. 1 (a)). A segmented suturing planning strategy is proposed based on point clouds obtained from the 3D endoscope, calculating the locations of the suture points based on the coordinates of the incision groove and the cut. The consistency of the bite size obtained by STAR matches that obtained by manual suturing; however, the execution time to complete the suturing process is longer compared to manual intervention. The limitations of the system include the limited speed of the motors, the use of static endoscopic images for suturing trajectory planning, and possible disturbances in the tissue. Unlike the present work, they created the motorized adapter for another forceps, the Endo 360.

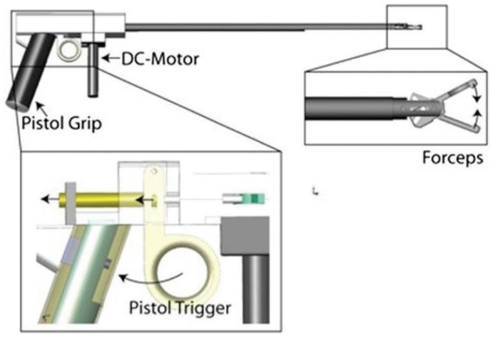
In the field of automated suturing, multiple methods have been investigated to determine the points to enter and exit the needle, including: Ink marking [12], [13], [14], the incorporation of a 3D environment with a haptic interface [15], [16], [17], the location of the desired point through teleoperated control of the instrumentation [18], wound texture analysis [19], and the analysis of wound point clouds to define the stitch position [11]. In this study, it was decided to start with direct mark- ing on the target and calibration using the robot.

Considering the above paragraphs, an alterna- tive for suture automation is proposed: a motorized coupling was developed that allows for automatic actuation of the jaws of the EndoStitch forceps. The design of this coupling integrates electronic and mechanical components that transmit movement from a motor to the jaws of the forceps, controlling the opening and closing for gripping the needle and passing the thread. The robustness and precision of this coupling are fundamental in ensuring the reliability and safety of the system. The design of the coupling was addressed in a previous publication [20] (see Fig. 1 (b)). The platform is based on a modular architecture that includes: a software block for system configura- tion, a library of adaptable movements for specific suturing tasks, a mechanism for determining the coordinates of the workspace, management of motion commands for the robot, and behavior of the forceps. The system was evaluated by performing sutures in a simulated scenario, using a silicone pad that simulates human tissue.

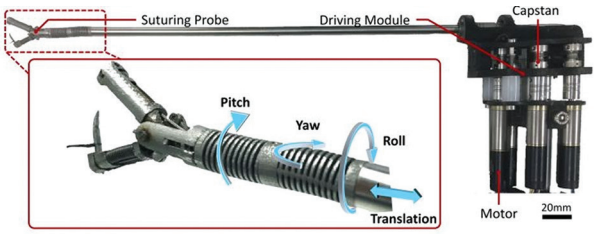


(a)

(b)



(c)



(d)

Fig. 1. Related Projects: Start automatic suturing system [11] (**a**), Endostitch adapter design [20] (**b**), EndoStitch with motorized handle [8] (**c**), Robot for single-port surgery [9] (**d**)

2. METHODS

This section outlines the methodologies em- ployed in the development and evaluation of the proposed semiautonomous robotic suturing system. To provide a foundational understanding of the mechanical interactions involved, we first present a finite element simulation analyzing the forces between the surgical needle and the tis- sue during penetration. This simulation addresses the challenge of directly measuring these forces at the tip of the needle due to sensor limitations. Following this, the subsequent subsections detail the development, implementation, and experimental evaluation of the robotic system itself.

The core of this study involves a semiautonomous robotic system designed for the automation of surgical suturing. This system integrates custom-motorized EndoStitch forceps with a UR3 robotic arm, with the goal of improving precision, repeatability, and efficiency in minimally invasive procedures. The collaborative robot UR3 (Universal Robots, Odense, Denmark) was selected as the central manipulation platform. This 6-degrees-of-freedom (DOF) robot offers a nominal payload of 3 kg, a maximum reach of 500 mm, and a declared positional repeatability of ±0.1 mm. The selection of the UR3 was based on its combination of flexibility, precision, compact size, and ease of programming, deeming it appropriate for the delicate manipulation and precise positioning required in surgical suturing tasks.

2.1. FINITE ELEMENT SIMULATION

Given the absence of sensors that can be placed at the needle tip, a simulation study of the needle-tissue interaction forces were conducted. Among the constitutive models for finite element analysis of soft tissues, a simulation was performed using ANSYS software due to its explicit dynamics capabilities, which allow for the analysis of puncture and penetration effects generated by the needle on the tissue. The environment configu- ration utilized results from Mahmud et al. [21], who analyzed skin properties, determining the following values for the Ogden model terms: μ_p = 110 Pa and α_p = 26. Based on the properties of skin components, parameters for the skin and muscle mod-els were selected as shown in Table 1. The definition of tissue characteristics, including its type and thickness, is fundamental, as the combination of these factors affect the tissue's response to needle penetration.

Table 1. Parameters for the model

Tissue	Model	Parameters
Skin	Ogden	$\mu = 110$ Pa
Muscle	Mooney-Rivlin	$\alpha = 26$ $C_{10} = 30000$ Pa $C_{01} = 30000$ Pa $d = 3.33 \times 10^{-5}$ Pa^{-1}

Fig. 2 illustrates the movement sequence developed over 10 seconds, showing the corresponding deformation and applied force indication. The simulation focuses on the corner of the simulated tissue, reflecting the common use of EndoStitch forceps for passing the needle at the tissue edge to close wounds or join tissues.

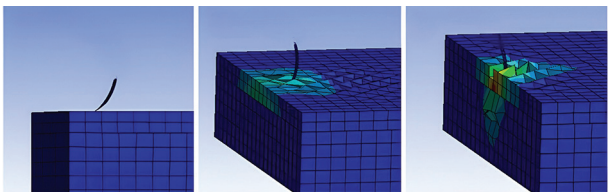


Fig. 2. Sequence of the needle penetration simulation

In the developed simulation, the equivalent stress between the bodies at the moment of contact was evaluated. Fig. 3 shows the curve representing the pressure found between the bodies.

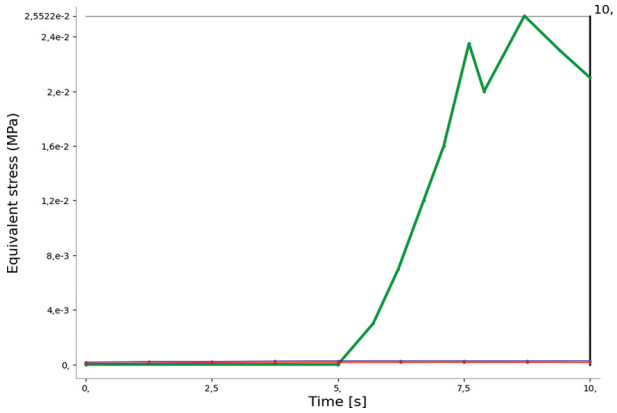


Fig. 3. Equivalent stress response

The finite element simulation provides a detailed visualization of the needle-tissue interaction during penetration, as observed in Fig. 2. This sequence illustrates the progressive stages of needle insertion: from initial contact (left), through partial penetration with visible tissue deformation and stress distribution (center), to deeper penetration where stress concentrations are more evident and extensive around the needle (right). The temporal evolution of mechanical stress in the tissue is quantified by the equivalent (von Mises) stress curve as a function of time, presented in Fig. 3. Initially, between 0 and approximately 5 seconds, the stress is minimal, corresponding to the approach or superficial contact phase of the needle. Subsequently, a gradual increase in stress is observed up to 7.5 seconds, followed by a more pronounced rise, reaching a peak of approximately 0.0255 MPa near 9.5 seconds. This peak coincides with the phase of greatest tissue penetration and deformation. Although equivalent stress is a scalar measure of the stress state intensity at a point and not the direct vectorial interaction force, its magnitude and temporal evolution are intrinsically proportional to the resistance the tissue offers to penetration. Higher equivalent stress in the tissue implies that it is exerting a greater reaction force on the needle. Therefore, this curve is fun- damental for identifying critical loading points dur- ing suturing, estimating the maximum forces the robotic system must be capable of applying, controlling, and predicting regions of high-stress concentration that could lead to tissue damage if certain thresholds are exceeded.

2.2. ROBOTIC PLATFORM: UR3 COLLABORA- TIVE ROBOT

A commercial EndoStitch forceps (Autosuture EndoStitch®, Medtronic) was modified through the design, manufacture, and integration of a specific motorized coupling. This coupling allows for automatic actuation of the jaws of the forceps, eliminating the need for direct manual intervention during the suturing process. The design of the coupling was carried out in CAD software (SolidWorks 2020, Dassault Syste`mes) and manufactured using 3D printing techniques.

For the actuation of the EndoStitch forceps, two types of servomotors were selected: an SG90, with a torque of 1.6 kg/cm (0.15 Nm), to control the needle transfer, and an MG995, with a torque of 8.5 kg/cm (0.83 Nm), to manage the opening and closing of the forceps. The choice of the SG90 for needle transfer is based on its compact size and sufficient torque to overcome the mechanical resistance associated with this process, minimizing the additional weight on the robot. The additional hardware consisted of a coupling structure manufactured using 3D printing, utilizing a rigid and lightweight plastic material to ensure structural integrity without overloading the UR3 robot. An Arduino controller was used to manage the servomotors, allowing for the programming of precise and coordinated movements, as well as communication with the robot through ROS (Robot Operating System). This configuration allowed for the integration of the forceps controls with the robot control, facilitating the execution of automatic suturing tasks and the collection of system performance data in real-time.

The design of the actuation system is based on the conversion of the rotational movement of the servomotors into the linear movement necessary to control the EndoStitch forceps. For opening and closing the forceps, the MG995 servomotor actuates a system of connecting rods connected to the handles of the forceps. This mechanism transforms the rotation of the motor into a linear displacement of the handles, allowing the forcep jaws to be opened or closed as needed to grasp the tissue. For needle transfer, the SG90 servo- motor actuates a lever that activates the internal mechanism of the forceps responsible for releas- ing the needle from one jaw and positioning it in the other. The precision of the movement in both mechanisms is controlled by the detailed programming of the servomotor angles, allowing for controlled and efficient manipulation of the forceps to perform suturing tasks with precision [20].

To ensure compatibility and optimal function- ality with the EndoStitch™ forceps, V-Loc™ 180 non-absorbable reloads were used. These reloads consist of a barbed monofilament thread, designed to provide continuous and secure tissue approxi- mation without the need for surgical knots. The needle at one end of the V-Loc™ 180 thread facilitates penetration and pas-

sage through the tissue, while the loop at the other end allows for a secure initial anchorage, simplifying and speeding up the suturing process. This choice not only guarantees the correct functionality of the forceps but also takes advantage of the benefits of a self-retaining thread, potentially reducing operating time and improving clinical outcomes.

2.3. SOFTWARE CONTROL ARCHITECTURE

The control software for the robotic system was developed using the Robot Operating System (ROS) framework, specifically the Melodic Morenia distribution, running on an Ubuntu 18.04 LTS environment. The software architecture is based on a modular structure, which facilitates the develop- ment, debugging, and maintenance of the system. This architecture is depicted in Fig. 4.

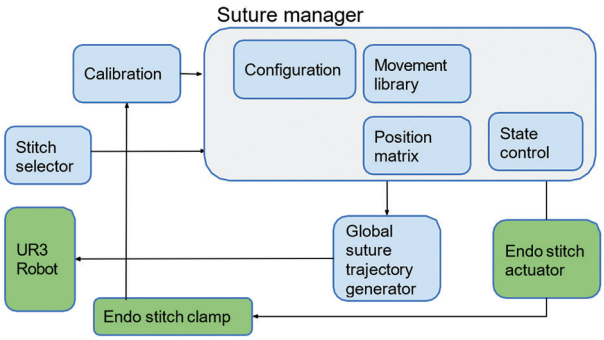


Fig. 4. System architecture for semi-automatic suturing

Fig. 4 presents a schematic diagram of the software control architecture developed for the robotic suturing system. This architecture, implemented on ROS (Robot Operating System), stands out for its modular design. The system's hardware components are identified in green: the collaborative "UR3 Robot," the motorized surgical "Endo stitch clamp," and the "Endo stitch actuator" responsible for actuating said clamp. These physical elements are integrated and managed by the software modules (represented in blue and light gray). The "Suture manager" acts as the cen-tral control core, orchestrating the operation. It receives inputs from the "Calibration" module (for spatial referencing) and the "Stitch selector" (to define the suturing task). Internally, the "Suture manager" contains the system "Configuration," the "Movement library" with movement primitives, the "Position matrix" which is used by the "Global suture trajectory generator" to plan the UR3 robot's trajectories and the "State control" which manages the operation sequence and commands the "Endo stitch actuator."

The main software modules include:

2.3.1. Configuration Module

This module allows configuring the system pa- rameters, including the suture characteristics: jaw closing speed, jaw opening percentage, distance between stitches, approach angle, the UR3 robot control parameters (e.g., maximum speeds and accelerations), and the type of stitch, which in the experiment is a continuous simple stitch.

2.3.2. Movement Library for Automated Suturing

This module implements a library of parameterized movements designed specifically for suturing tasks, ranging from the initial approach to thread tensioning. Each movement, defined in the joint space of the UR3 robot, is executed sequentially to perform a complete stitch. Fig. 5 graphically illustrates the functions that make up this library.

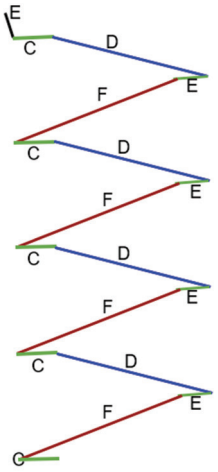


Fig. 5. Movement library for suturing

Trajectory (A) (in black) corresponds to the ini- tial approach to the suture area. Trajectory (B) (in green) represents the jaw change for the needle; before its execution, the jaws must be closed, ensuring secure needle holding through an inte- grated safety system. Trajectory (C) (in blue) is the thread extension movement, adjustable to dis- pense more thread in the first stitches and less in the last. Trajectory (D) (in black) represents the passage of the needle between the jaws, necessary depending on the type of suture used. Trajectory (E) (in red) represents the approach to the next stitch, a point from which the cycle repeats until the suture is complete. The arrangement of trajectories (B) in the graph symbolizes the different stitches needed to close the wound. It is essential to highlight that this pattern must be adapted to the specific morphology of the wound.

The implementation of this library of parameterized movements offers a modular and adaptable solution for different surgical scenarios. Breaking down the suturing process into a sequence of elementary and parameterizable movements facilitates the programming and adaptation of the robot to the specific geometry of the wound and the needs of the surgeon. This flexibility, combined with the integrated safety system to ensure secure needle holding, allows for optimizing the

precision and efficiency of the suturing process, minimizing inter-operator variability and potentially reducing surgical time, thereby improving clinical outcomes. Furthermore, the parameterization of the movements facilitates the incorporation of machine learning and adaptive control techniques, opening the door to the creation of autonomous suturing systems with the ability to adjust in real time to the changing conditions of the surgical environment.

2.3.3. Trajectory Planner

This module implements a trajectory planning algorithm based on the joint space of the UR3 robot, using cubic splines to generate smooth and continuous trajectories that avoid collisions and minimize execution time. The trajectory planner considers the speed and acceleration limitations of the UR3 robot and the EndoStitch forceps to ensure the safe and efficient operation of the system.

2.3.4. State Controller

This module manages the current state of the robotic system, including the position and orienta- tion of the UR3 robot, the state of the EndoStitch forceps (e.g., open, closed, passing the needle), and the progress of the suture. The state controller implements a state machine-based control logic that allows coordinating the execution of the different movements and tasks of the robotic system.

2.3.5. Calibration Module

This module implements a calibration proce- dure to align the coordinate system of the UR3 robot with the real workspace. This procedure is based on the identification of a set of reference points in the workspace, defined as the base of the robot. To do this, the robotic arm, equipped with the forceps, is sequentially moved to each calibration point. The corresponding Cartesian positions are recorded and then correlated with the desired positions of these points in the coordinate system of the UR3 robot. This correlation process allows for establishing a precise transformation be- tween the robot model and the physical environment, minimizing positioning errors and ensuring the accuracy of suturing tasks.

2.3.6. Robotic Suturing Protocol

The protocol for automated suturing, illustrated in Fig. 6, defines the integral workflow of the system, encompassing everything from the initial preparation of the surgical environment to the pre- cise execution of the stitches and the controlled completion of the process.

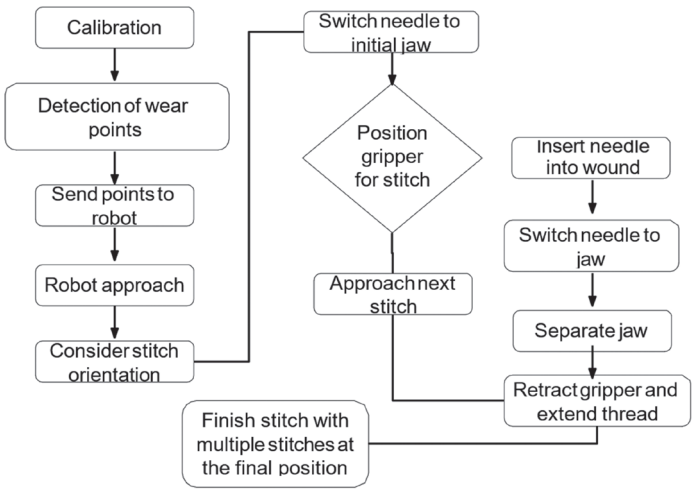


Fig. 6. Sequential diagram of the robotic suturing protocol

The protocol, designed to perform precise and reproducible sutures in various types of wounds, minimizing inter-operator variability and optimizing surgical time, is composed of the following stages:

Calibration and Wound Definition: This initial stage comprises system calibration, essential to align the robot's workspace with the physical reality. Next, key points that delimit the wound or the area to be sutured are detected, and defined by its shape and extension. Finally, the detected points are sent to the robot's control system, establishing the frame of reference for suturing.

Approach and Initial Preparation: In this phase, the robot makes an approach to the suture area, bringing the EndoStitch forceps to the starting point of the wound. At the same time, the desired orientation of the stitch is considered, adjusting the position of the forceps to ensure the correct penetration and direction of the needle in the tissue. In addition, the initial transfer of the needle is carried out, ensuring that it is correctly positioned in the initial jaw of the forceps.

Automated Suturing Cycle: This iterative stage comprises the repeated execution of the in- dividual stitch:

Positioning for the Stitch: It is evaluated whether it is necessary to position the gripper for the stitch.

Needle Insertion and Passage: The needle is inserted into the wound, passing through the tissue and completing the passage to the opposite jaw of the forceps (Needle Transfer).

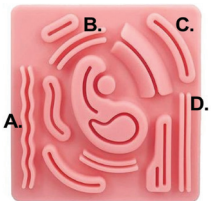
Thread Extension and Tensioning: The forceps jaws are separated, extending the thread, while the robotic system retracts, extending and tensioning the thread, ensuring proper tissue closure.

Approach to the Next Stitch: The robot approaches the next stitch, preparing for needle insertion in the next position along the wound.

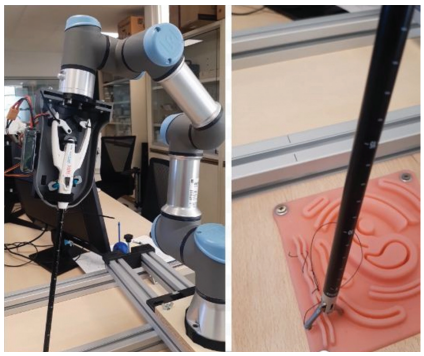
Suture Completion: After completing each cycle, it is decided whether it is the end of the process (Suture Completion). The repetition of the automated suturing cycle ceases when the desired number of stitches is reached, ensuring complete and effective closure of the wound with multiple stitches in the final position. This final stage could be considered as a knot to secure the suture.

2.4. EXPERIMENTAL EVALUATION ENVIRONMENT

The robotic suturing system was evaluated in a simulated environment using laparoscopic suturing training pads, model PH03-163, manufactured with soft silicone to simulate soft tissue. These pads, with dimensions of 14 x 14 x 1 cm and a weight of 181 g, were specifically designed for practicing basic laparoscopy skills. The silicone pad was fixed to a rigid base to simulate surgical tissue. The workstation was organized to ensure that all components of the system (UR3 robot, motorized EndoStitch forceps, control electronics, simulation pad) worked together efficiently, see 7 (b), (c). Constant and controlled lighting was main- tained to ensure correct visibility during the eval- uation. Fig. 7(a) presents the wounds selected for testing with the robot and forceps.



(a)



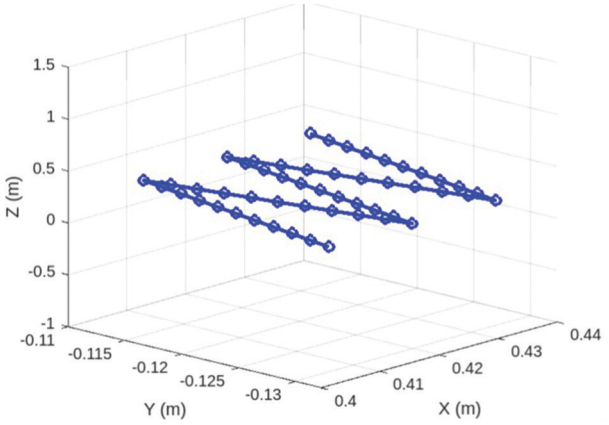
(b) (c)

Fig. 7. Experimental setup. (**a**). Suture pad used and wounds selected. (**b**) Robot forceps assembly. (**c**) Forceps on pad

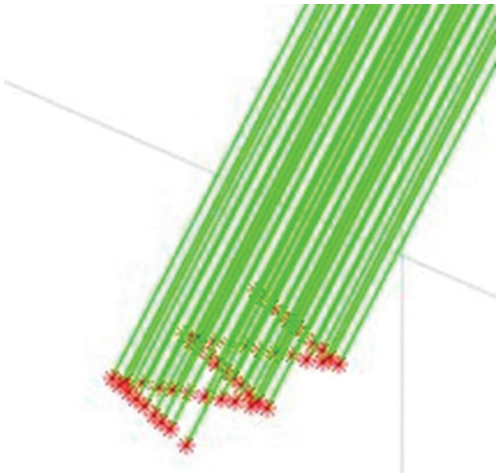
The workflow is governed by a state machine whose activations and state changes allow the execution of stitch by stitch. Fig. 8 shows the evolution of a suture performed on wound type (a). The sequence can be observed through which the robot places the stitches on the wound until the suture is complete.

Fig. 8. Sequences of stitches on a wound

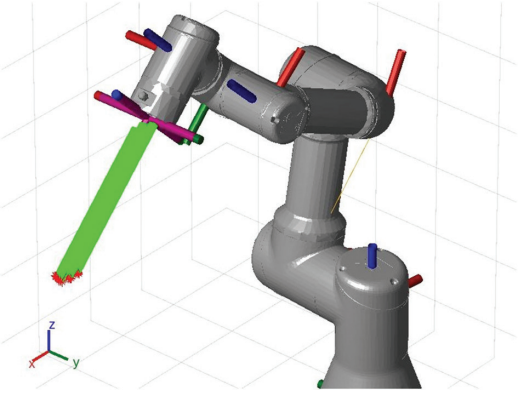
To have a margin for comparison of the experiments, a simulation was performed in Matlab where the robot develops the desired movement. In this case, it is observed how the robot is moving; the pattern shown in Fig. 9(a) represents the movement of separation of the wound to extend the thread. In Figs. 9(b), (c), the robot is shown developing the desired trajectory.



(a)



(b)



(c)

Fig. 9. Simulation. (**a**) Desired trajectory, (**b**) Achieved trajectory, (**c**) Robot movement

The results of the simulated desired trajectory are shown in Fig. 10. As expected, the simulated error was significantly smaller—on the order of nanometers.

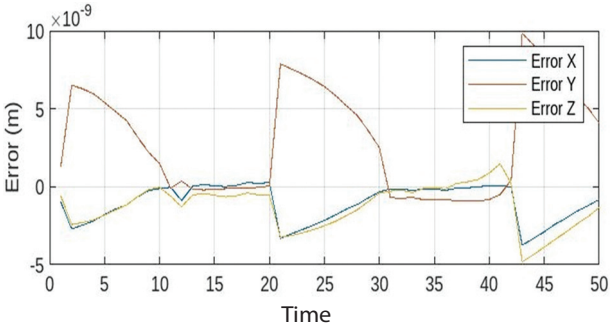


Fig. 10. Error obtained in the simulation

It can be observed that the errors in the X and Y axes are nearly identical, while the error in the Z-axis exhibits some noticeable peaks, indicating a different behavior.

3. RESULTS

This section presents the results of an experiment designed to evaluate the accuracy of a semiautonomous robotic system in performing surgical sutures. The experiment consisted of performing four independent suturing tests on different wounds present on the test pad using a motorized EndoStitch forceps integrated with the UR3 robotic arm. The main objective was to quantify the positioning error of the system in the three dimensions during the suturing process. In each test, the robot performed a series of stitches in the simulated tissue, and the deviation between the desired position and the actual position of the needle was measured in each stitch. (The table shows the experimental data 2)

Table 2. Statistical Summary of Positioning Error by Test Configuration ($N = 6$ per test)

Test	Axis	Mean (x^-) ($\times 10^{-5}$ m)	Std. Dev. (s) ($\times 10^{-5}$ m)	95% CI ($\times 10^{-5}$ m)	CV (%)
*A	X	0.9	0.25	[0.64, 1.16]	27.8
	Y	0.8	0.20	[0.59, 1.01]	25.0
	Z	1.6	0.55	[1.02, 2.18]	34.4
*B	X	1.1	0.40	[0.68, 1.52]	36.4
	Y	1.0	0.35	[0.63, 1.37]	35.0
	Z	1.2	0.45	[0.73, 1.67]	37.5
*C	X	2.0	0.70	[1.27, 2.73]	35.0
	Y	0.9	0.25	[0.64, 1.16]	27.8
	Z	1.7	0.60	[1.08, 2.32]	35.3
*D	X	0.6	0.15	[0.44, 0.76]	25.0
	Y	0.5	0.10	[0.39, 0.61]	20.0
	Z	0.7	0.20	[0.49, 0.91]	28.6

CI: Confidence Interval; CV: Coefficient of Variation; $t_{0.025,5} = 2.571$

3.1. STATISTICAL SUMMARY OF POSITIONING ACCURACY

The experimental design employed $N = 6$ independent repetitions for each test configuration, totaling 24 trials across four wound geometries.

This sample size was determined based on statistical power analysis to detect meaningful differences in positioning accuracy while maintaining practical experimental constraints.

For precision measurements in robotic systems, $N = 6$ provides sufficient statistical power ($\beta > 0.80$) to detect effect sizes of practical signifi- cance ($\delta \geq 0.5 \times 10^{-5}$ m) with $\alpha = 0.05$. The sample size ensures reliable estimation of population parameters while accounting for the controlled experimental environment and high precision instrumentation that minimizes measurement variability.

The statistical analysis demonstrates high system reliability, with positioning accuracy consis- tently in the 10^{-5} meter range across all configurations. The mean positioning errors range from 0.5×10^{-5} m (Test D, Y-axis) to 2.0×10^{-5} m (Test C, X-axis), representing a 200-fold improvement over manual suturing precision ($\sim 2 \times 10^{-3}$ m). Standard deviations remain below 0.70×10^{-5} m in all cases, indicating consistent performance across repetitions.

The 95% confidence intervals provide statisti- cal assurance of system performance bounds un- der operational conditions. Coefficients of vari- ation (CV) rang-

ing from 20.0 % to 37.5 % reflect the inherent variability in complex robotic systems while maintaining clinically acceptable precision levels. The narrow confidence intervals, particularly for Test D (straight geometry), confirm that the proposed system achieves reproducible submillimetric accuracy suitable for precision surgical applications.

The statistical robustness of $N = 6$ repetitions is validated by the tight confidence bounds and consistent CV values, providing sufficient evidence for the system's reliability and establishing a foun- dation for clinical translation studies with larger sample sizes.

The main evaluation metric was the positioning error, defined as the vector distance between the target position of the needle and its actual position after insertion. This error was decomposed into three components: error in the X component (lat- eral), error in the Y component (depth), and error in the Z component (vertical). The results of each test are presented in Figs.11 and 12, where the error in the three components is plotted for each stitch performed. The analysis of these results allowed identifying characteristic error patterns of the system, as well as the variability of the performance between the different tests.

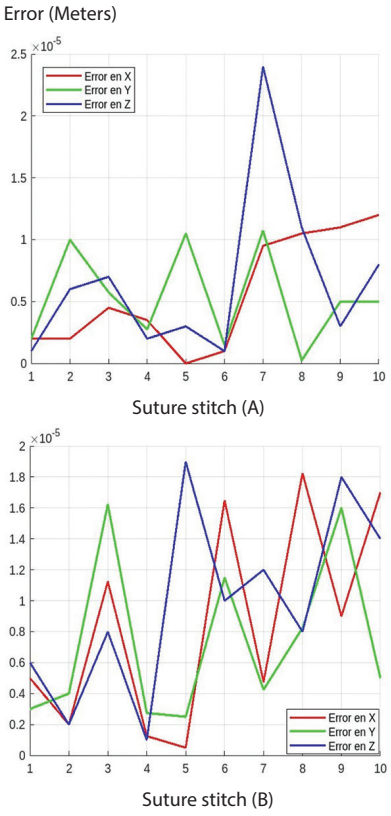


Fig. 11. Experimental results of the tests on wound types A and B, the corresponding wounds are those shown in Figure 7(a).

Figs. 11 and 12 illustrate the results of four independent robotic suturing tests (a, b, c, d) conducted on tissue simulators, displaying the positioning error (in meters) for each stitch across the X (red), Y (green), and Z (blue) components. The horizontal axis represents the stitch number, while the vertical axis indicates the error magnitude. The tests were performed using a semi-autonomous robotic system based on ROS, employing a latex training pad with simulated wounds: test (a) corresponds to a straight but sinusoidal wound, and test

(b) to a slightly curved arc-shaped wound, both featuring a groove that enhances the edges. Target positions were predefined and stored using the same robot, with no modifications to the control architecture between tests, allowing the evaluation of the influence of wound geometry and tissue orien- tation on the system's accuracy.

DETAILED ANALYSIS BY TEST

Test (A): Sinusoidal Wound

In the sinusoidal wound test, the positioning error in the X and Y directions remains generally stable, with values below 1.5×10^{-5} meters, whereas a significant peak of approximately 2.4×10^{-5} meters is observed in the Z component at stitch 7. This increase may be attributed to the sinusoidal geometry, which introduces rapid changes in tissue orientation at the groove's inflection points. In these regions, the interaction between the EndoStitch forceps and the latex surface likely complicates precise control of needle insertion depth (Z) due to variations in material resistance or deformation. The upward trend in X error from stitch 6 suggests a progressive accumulation of inaccuracies while following the curved trajectory, possibly amplified by propagated errors in the horizontal plane. In contrast, the Y error remains more contained, though it exhibits a peak at stitch 2 (with no immediately apparent cause), indicating relatively stable control in the direction of advancement, but with potential limitations in compensating for the sinusoidal curvature within the planned trajectory.

Test (B): Slightly Curved Parallel Wound

In the slightly curved wound test, errors across all three components (X, Y, Z) show greater vari- ability, ranging between 0.2×10^{-5} and 1.8×10^{-5} meters, with notable peaks at stitches 2, 5, and 8. The arc-shaped curvature, though mild, introduces additional complexity to positioning control, requiring continuous trajectory adjustments by the UR3 robot. The alternating peaks across components suggest that the system struggles to coordinate movements between axes while adapting to the curve, potentially exacerbating inherent instabilities or inter-axis interactions. Furthermore, the curvature may induce subtle variations in the ten- sion of the simulated tissue along the groove, affecting needle insertion and contributing to error dispersion. The absence of dynamic adjustments in the control architecture between tests indicates that the system was not optimized for this geometry, allowing small initial errors to propagate and amplify along the trajectory, particularly in the Z direction, where maintaining needle perpendicularity to the groove is more challenging.

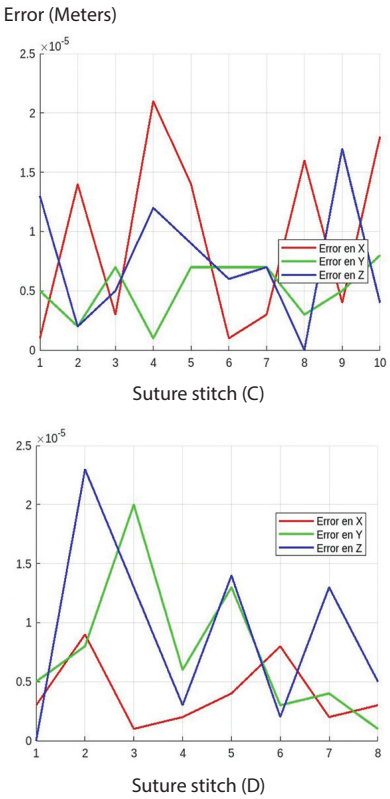


Fig. 12. Experimental results of the tests on wound types C, D. The corresponding wounds are those shown in Figure 7(a).

Test (C): Non-Parallel Curved Wound

In test (c), the error in the X component (red) stands out with a higher magnitude, reaching peaks of up to 2.5×10^{-5} meters at stitches 2, 4, and 8, while Y (green) remains stable ($< 1.5 \times 10^{-5}$ m) and Z (blue) fluctuates with a peak at stitch 2 ($\sim 2.0 \times 10^{-5}$ m). This is due to the arc geometry in the opposite direction to test (b), requiring constant adjustments in the EndoStitch forceps orientation to follow the curved groove. The nonparallel alignment with the UR3 robot's axes misaligns the X-axis with the main suturing direction, leading to cumulative lateral deviations. The stability in Y suggests effective transverse control, but Z fluctuations indicate challenges in maintaining uniform needle depth amid changing tissue orientation. A possible justification is that the ROSbased control architecture, lacking dynamic adjustments, fails to adequately compensate for the multi-axis complexity of the reversed curve, allowing small initial inaccuracies to amplify, particularly in X, where reorientation is most critical.

Test (D): Narrow Straight Wound

In test (d), initial errors in Y (green) and Z (blue) are higher, peaking at stitch 2 ($\sim 2.0 \times 10^{-5}$ m), while X (red) remains low. From stitch 4 onward, all three components converge to consistent values ($< 1.0 \times 10^{-5}$ m), indicating stabilization. The straight trajectory facilitates predictable tracking, reducing variability after the initial stitches. The early Z peak may stem from a specific interaction between the forceps and the uniform-height groove in the latex pad, such as local deformation or suboptimal motion parameterization, while the Y error suggests an initial transverse calibration issue. The likely justification is that the linear simplicity allows the system to quickly adapt after a brief settling period, and the groove's uniformity minimizes external factors, though unmeasured forces prevent confirming whether material resistance contributed to the Z peak.

The greater variability in (c) is justified by the complexity of the reversed curvature, demanding multi-axis coordination that the static control system cannot optimize, resulting in accumulated X errors. In (d), stability after initial stitches is explained by the ease of following a straight path, enabling the robot to correct early inaccuracies and leverage geometric predictability. Z peaks in both tests may be justified by latex interaction (resistance or deformation), unaddressed by the static control, while Y stability in (c) and convergence in (d) reflect the system's ability to handle directions less affected by specific geometry.

It is important to note that, despite this variability in the error pattern, the magnitude of the positioning error consistently remains in the order of micrometers. Although there are specific peaks that reach tens of micrometers, the overall error scale is in this micrometric range. This precision, even considering the variability, represents a level of accuracy on the order of magnitude of the typical accuracy of manual suturing performed by experienced surgeons (approximately 2 mm). This difference suggests that, despite the variability observed in the different test conditions, the robotic system has the potential to offer a notably improved positioning accuracy in the suturing task compared to manual execution.

4. DISCUSSION

The results of this study, although promising, reveal critical aspects for the successful development of robotic suturing systems. The variability observed in the positioning error between tests, even with a micrometric magnitude, underscores the need for robust and adaptable control that can compensate for disturbances and uncertainties present in the real surgical environment. While the error in the simulation reached nanometric levels, this difference highlights the inherent complexity of the real environment compared to the simplified models used in simulation. However, the identification of similar patterns in the error peaks between the simulation and the experiments suggests that the model captures fundamental elements of the system, validating its usefulness in optimizing the design and control strategies. This observation is crucial, as it implies that the simulation can be a valuable tool to predict and mitigate certain types of errors before conducting tests in more complex and costly environments.

The implementation of a parameterized movement library, along with the strategic choice of the UR3 robot

and the motorized EndoStitch forceps, proved to be suitable options for the research. However, it is imperative to recognize the limitations inherent in the evaluation in a simulated environment. The complexity of living tissue, with its viscoelastic properties and dynamic behavior, is not fully replicated in the simulation pads. Therefore, a more thorough statistical analysis, including a greater number of repetitions and the evaluation of the statistical significance of the observed differences, along with tests in more realistic models (such as animal tissues or biomimetic models), are essential steps for a complete and robust evaluation of the system. This more rigorous evaluation will allow determining more precisely the real potential of the robotic system in a clinical surgical environment.

In comparison with the work of [11], this study focuses on the automation of a different forceps. However, instead of considering them as mutually exclusive solutions, the possibility of a strategic coexistence could be explored. Integrating the advantages of both automated forceps could lead to a more versatile and adaptable system. For example, the EndoStitch Forceps, with its larger bite, could be more efficient for tasks such as anastomosis and tissue coaptation, where precise joining or approximation of large areas of tissue is required. In contrast, the other forceps could be more suitable for tasks that demand greater precision in confined spaces.

Unlike grasper-based approaches such as Endowrist or Endo360, where needle re-grasping poses a significant technical challenge, our solution with the motorized EndoStitch leverages its inherently and clinically validated capability for automatic needle transfer. This distinctive feature drastically simplifies robotic control during this critical phase, enhancing the reliability of needle handling and allowing the system to focus on precise positioning. Potentially, this translates into more consistent and uniform suturing, a possible reduction in the surgical time devoted to suturing, and a lower cognitive load for the surgeon, with the potential to improve therapeutic outcomes by enabling smoother and more efficient automated procedures.

The simulation analysis reveals a contact interaction. This value, although indicative of the internal load, differs from the high reaction forces (e.g., 26-30 N by [22], 0-40 N/m by [23]) reported in the literature for needle insertion. This discrepancy can be attributed to the rigid boundary conditions of the current simulation and the absence of a tissue damage model, factors that tend to artificially increase the penetration resistance in the model compared to ex vivo experimental scenarios where the tissue can deform or tear, reducing the forces and local stress.

Despite these limitations, stress analysis offers significant advantages over simple force measurement. It allows for the prediction of material failure or deformation by comparing the calculated stress with the material's intrinsic limits, the identification of critical points through the visualization of stress concentrations (crucial for the design of sharp tools), and the optimization of designs for efficiency and safety. Unlike total force, stress provides a detailed understanding of how the load is distributed internally, offering deeper insight into contact mechanics and its impact on the material. For future research, and to better align the results with clinical reality, it is essential to incorporate more complex material models that reflect tissue anisotropy and heterogeneity, as well as damage or fracture models.

5. CONCLUSION

This study successfully developed and evaluated a semi-autonomous robotic system for the automation of surgical suturing, integrating a motorized EndoStitch forceps with a UR3 robot and a ROS-based control architecture. The system demonstrated significant precision in performing suturing tasks on simulated tissue.

The quantitative evaluation of the system's positioning accuracy revealed promising results. Across four different simulated wound geometries and with N=6 repetitions per test, mean positioning errors were consistently maintained in the micrometer range. Specifically, mean errors along the X, Y, and Z axes ranged from as low as 0.5×10^{-5} m (Test D, Y-axis) to a maximum of 2.0×10^{-5} m (Test C, X-axis), as detailed in Table 2. Standard deviations remained consistently low, with the highest recorded at 0.70×10^{-5} m, indicating repeatable performance. This level of precision represents a substantial, approximately 200-fold improvement over typical manual suturing accuracy (estimated around 2×10^{-3} m).

While the overall performance achieved micrometer-level accuracy, the analysis also highlighted variability influenced by wound geometry. For instance, Test C (non-parallel curved wound) exhibited peak errors in the X-component up to 2.5×10^{-5} m. Coefficients of Variation (CV) across all tests and axes ranged from 20.0% to 37.5%, reflecting the inherent complexities of robotic interaction with deformable surfaces and the current control strategy. Despite this variability, the achieved precision, even at its observed peaks, remains orders of magnitude better than manual execution.

These findings validate the feasibility of the developed system for automated suturing and underscore its potential to significantly enhance precision and consistency compared to manual techniques. The obtained quantitative data provide a strong baseline and highlight that while micrometer accuracy is achievable, future work should focus on optimizing control algorithms to further reduce error variability, particularly in response to complex tissue interactions and wound paths. Further evaluation in more realistic biological models and eventually in clinical settings, alongside the development of adaptive control strategies, possibly incorporating machine learning, will be crucial for translating this promising technology into tangible clinical benefits, such as reduced surgeon cognitive load and improved patient outcomes.

6. REFERENCES

[1] E. Lau, N. A. Alkhamesi, C. M. Schlachta, "Impact of robotic assistance on mental workload and cognitive performance of surgical trainees performing a complex minimally invasive suturing task", Surgical Endoscopy, Vol. 34, No. 6, 2020, pp. 2551-2559.

[2] F. Ju, X. Luo, L. Ding, "Multifunctional Robotic Surgical Forceps With Tactile Sensor Array for Tissue Palpation and Clamping Status Detection", IEEE Sensors Journal, Vol. 24, No. 10, 2024, pp. 16883-16891.

[3] Y. Yamasaki et al. "Effects of a force feedback function in a surgical robot on the suturing procedure", Surgical Endoscopy, Vol. 38, No. 3, 2024, pp. 1222-1229.

[4] K. Wang, J.-N. Ma, C.-Y. Zhang, Z. Pei, W.-T. Tang, Q. Zhang, "Self-powered high-sensitivity piezoelectric sensors for end-fixture force sensing in surgical robots based on T-ZnO", Colloids and Surfaces A: Physicochemical and Engineering Aspects, Vol. 697, 2024, p. 134424.

[5] Z. Cheng, M. Chauhan, B. L. Davies, D. G. Caldwell, L. S. Mattos, "Modelling needle forces during insertion into soft tissue", Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Milan, Italy, 25-29 August 2015, pp. 4840-4844.

[6] Y. Zhang, F. Ju, X. Wei, D. Wang, Y. Wang, "A Piezoelectric Tactile Sensor for Tissue Stiffness Detection with Arbitrary Contact Angle", Sensors, Vol. 20, No. 22, 2020, p. 6607.

[7] C.-H. Yeh et al. "Application of piezoelectric actuator to simplified haptic feedback system", Sensors and Actuators A: Physical, Vol. 303, 2020, p. 111820.

[8] T. Gopel, F. Hartl, A. Schneider, M. Buss, H. Feussner, "Automation of a suturing device for minimally invasive surgery", Surgical Endoscopy, Vol. 25, No. 7, 2011, pp. 2100-2104.

[9] Y. Hu, W. Li, L. Zhang, G.-Z. Yang, "Designing, Prototyping, and Testing a Flexible Suturing Robot for Transanal Endoscopic Microsurgery", IEEE Robotics and Automation Letters, Vol. 4, No. 2, 2019, pp. 1669-1675.

[10] S. N. Kurenov, S. Punak, M. Kim, J. Peters, J. C. Cendan, "Simulation for training with the autosuture endo stitch device", Surgical Innovation, Vol. 13, No. 4, 2006, pp. 283-287.

[11] H. Saeidi, H. Le, J. D. Opfermann, S. Leonard, A. Kim, M. H. Hsieh, J. U. Kang, A. Krieger, "Autonomous Laparoscopic Robotic Suturing with a Novel Actuated Suturing Tool and 3D Endoscope", Proceedings of the International Conference on Robotics and Automation, Montreal, QC, Canada, 20-24 May 2019, pp. 1541-1547.

[12] S. A. Pedram, P. Ferguson, J. Ma, E. P. Dutson, J. Rosen, "Optimal needle diameter, shape, and path in autonomous suturing", arXiv:1901.04588, 2019.

[13] J. Colan, J. Nakanishi, T. Aoyama, Y. Hasegawa, "Optimization-based constrained trajectory generation for robot-assisted stitching in endonasal surgery", Robotics, Vol. 10, No. 1, 2021.

[14] S. A. Pedram, C. Shin, P. W. Ferguson, J. Ma, E. P. Dutson, J. Rosen, "Autonomous Suturing Framework and Quantification Using a Cable-Driven Surgical Robot", IEEE Transactions on Robotics, Vol. 37, No. 2, 2021, pp. 404-417.

[15] Y. Tian, M. Draelos, G. Tang, R. Qian, A. N. Kuo, J. A. Izatt, K. Hauser, "Toward Autonomous Robotic Micro-Suturing using Optical Coherence Tomography Calibration and Path Planning", Proceedings of the IEEE International Conference on Robotics and Automation, Paris, France, 2020, pp. 5516-5522.

[16] A. E. Abdelaal, J. Liu, N. Hong, G. D. Hager, S. E. Salcudean, "Parallelism in Autonomous Robotic Surgery", IEEE Robotics and Automation Letters, Vol. 6, No. 2, 2021, pp. 1824-1831.

[17] H. Dehghani, Y. Sun, L. Cubrich, D. Oleynikov, S. Farritor, B. Terry, "An Optimization-Based Algorithm for Trajectory Planning of an Under-actuated Robotic Arm to Perform Autonomous Suturing", IEEE Transactions on Biomedical Engineering, Vol. 68, No. 4, 2021, pp. 1262-1272.

[18] S. Gao, S. Ji, M. Feng, X. Lu, W. Tong, "A study on autonomous suturing task assignment in robot-assisted minimally invasive surgery", The International Journal of Medical Robotics + Computer Assisted Surgery: MRCAS, Vol. 17, No. 1, 2021.

[19] H. Vargas, V. Munoz, A. Vivas, "Automated suturing: sharp wound recognition and planning with surgical robot", Advanced Robotics, Vol. 37, No. 14, 2023, pp. 900-917.

[20] O. L. T. Diaz, J. L. C. Gonzalez, O. A. V. Alban, J. C. F. Marinero, "Design, simulation and first test of an automatic suturing device coupled to a robot", Reveia, Vol. 21, No. 41, 2024, pp. 1-18.

[21] G. H. Mahmud, Z. Yang, A. M. T. Hassan, "Experimental and numerical studies of size effects of Ultra High Performance Steel Fibre Reinforced Concrete (UHPFRC) beams", Construction and Building Materials, Vol. 48, 2013, pp. 1027-1034.

[22] Y. Jiang, Q. Song, X. Luo, "3D Cohesive Finite Element Minimum Invasive Surgery Simulation Based on Kelvin-Voigt Model", Chinese Journal of Mechanical Engineering, Vol. 35, No. 1, 2022.

[23] M. Terzano, D. Dini, F. R. y Baena, A. Spagnoli, M. Oldfield, "An adaptive finite element model for steerable needles", Biomechanics and Modeling in Mechanobiology, Vol. 19, No. 5, 2020, pp. 1809-1825.

The Conditional Handover Parameter Optimization for 5G Networks

Original Scientific Paper

Zolzaya Kherlenchimeg
National University of Mongolia,
School of Information Technology and Electronics
Oyutnii gudamj 14/3, Ulaanbaatar, Mongolia
zolzayakhh@num.edu.mn

Battulga Davaasambuu*
National University of Mongolia,
School of Information Technology and Electronics
Oyutnii gudamj 14/3, Ulaanbaatar, Mongolia
battulgad@num.edu.mn

Telmuun Tumnee
National University of Mongolia,
School of Information Technology and Electronics
Oyutnii gudamj 14/3, Ulaanbaatar, Mongolia
telmuun@num.edu.mn

*Corresponding author

Nanzadragchaa Dambasuren
National University of Mongolia,
School of Information Technology and Electronics
Oyutnii gudamj 14/3, Ulaanbaatar, Mongolia
nanzadragchaa@num.edu.mn

Di Zhang
School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 518107, China
zhangd263@mail.sysu.edu.cn

Ugtakhbayar Naidansuren
National University of Mongolia,
School of Information Technology and Electronics
Oyutnii gudamj 14/3, Ulaanbaatar, Mongolia
ugtakhbayar@num.edu.mn

Abstract *– Conditional Handover (CHO) is a state-of-the-art handover technique designed for 5G and beyond networks. It decouples the preparation and execution phases of the traditional handover process and aims to reduce wrong cell selection by utilizing a predefined list of target cells. Despite its advantages, the limitations of static parameter configuration compromise CHO performance. This paper proposes a self-optimization mechanism for CHO parameters in 5G networks. Our proposed mechanism is an automated method for estimating and optimizing CHO parameters, dynamically adjusting key parameters to fine-tune the conditions that trigger the execution phase of the handover process. In addition, we introduce a second handover trigger referred to as the cell outage condition. We compared the performance of our proposed mechanism with the baseline CHO, velocity-based, and cell-outage based mechanisms, using Ping-Pong Handovers (PPHO) and Radio Link Failures (RLF). The simulation results demonstrate reduction of up to 7% in RLF, a 0.15% decrease in handover errors, and an improvement of approximately 10% in handover performance at velocities of up to 200 km/h in high-mobility scenarios.*

Keywords*: mobile networks, conditional handover, handover performance, optimization*

Received: May 2, 2025; Received in revised form: July 2, 2025; Accepted: July 2, 2025

1. INTRODUCTION

Fifth generation (5G) mobile technology has significantly impacted daily life by enabling new features such as the Internet of Things (IoT), massive machine-type communications (mMTC), Ultra-Reliable Low-Latency Communications (URLLC) and vehicular networks (V2X) [1]. Next generations of mobile networks, including 5G-Advanced and Beyond 5G (B5G), aim to support even higher data rates, near-zero latency, and ubiquitous connectivity [1, 2]. In addition, B5G integrates artificial intelligence for network optimization, employs terahertz (THz) communication bands to enable extreme bandwidth, holographic and tactile Internet applications, and holistic network coverage integrating terrestrial, aerial, and satellite segments [1,2].

5G networks also use dual-connectivity and multi-connectivity techniques to improve connectivity and overall performance. Using mm-wave frequencies enables denser cellular networks to meet higher data throughput requirements. However, these challenges include moving mobile users and requiring frequent handovers between microcells. Handover processes in mobility management ensure that users maintain uninterrupted connectivity during cell transitions.

The 5G literature has introduced three primary handover techniques to minimize interruption time and reduce handover failures. The first two techniques have been developed to address dual-connection and multi-connection scenarios. In 5G and beyond networks, the slicing technique is an important virtualization [3]. It is also necessary to define the procedures for handover between slices.

The third technique, CHO, is a cutting-edge mechanism specifically designed for 5G homogeneous networks [4,5]. The CHO is developed to decouple the preparation and execution phases of the traditional handover process and aims to reduce wrong cell selection by utilizing a predefined list of target cells [6]. Also, it was initially introduced to enhance service quality and reliability for users with single connectivity within homogeneous networks. In the paper [7], the authors have dedicated their efforts to refining the CHO process, particularly to minimizing wrong target-cell selections during the preparation phase. If the wrong target cell is selected, the handover process fails in the execution phase, and the user equipment (UE) reconnects to the serving cell. The UE then initiates a new handover procedure, resulting in temporary unavailability of the user's data path and an extended service interruption time.

The adoption of CHO in 5G networks offers notable benefits, but also increases control messages, which strains network management [8]. Also, deploying and configuring multiple conditions with static parameters in CHO may adversely effect overall network performance [9]. To address the challenge of CHO, we propose a parameter self-optimization mechanism for CHO in 5G networks. Our Autotuning-based Parameters for Conditional Handover (APCHO) mechanism estimates and optimizes the trigger parameters of CHO. Our main contributions are as follows:

- We formulate an auto-tuning mechanism for the offset parameter in CHO's execution condition to reduce handover errors. Our mechanism estimates the offset parameter based on the UE's velocity, which serves as the main handover trigger.
- We propose an efficient cell outage condition that is related to cell size. Our mechanism uses this condition as an additional trigger for the handover procedure. This mechanism helps to manage late and early handovers and improve the efficiency of handover procedures in 5G HetNets.
- Furthermore, our APCHO mechanism uses an approach based on mobility management performance data to dynamically adjust the weight of parameters to mitigate the effect of handover errors.

The paper is organized as follows. Section I provide an overview of the topic, while Section II introduces the background and related works. In Section III details the proposed mechanism. Section IV presents the simulation setup and results. Finally, Section V concludes the paper with final remarks.

2. BACKGROUNDS OF CONDITIONAL HANDOVER

This section provides an overview of CHO and Handover Performance Indicator (HPI), and related works.

2.1. CONDITIONAL HANDOVER

Mobile networks use a handover mechanism comprising three phases: preparation, execution, and completion. The cell to which the UE is currently connected cell, called the serving cell, begins the preparation phase with a measurement command. In response, the UE returns the measurement results to the serving cell. Based on the measurement results, the serving cell identifies a suitable cell called the target cell for the user's transition and sets the trigger for the execution phase. In the completion phase, the serving cell releases the radio resources and other configurations based on the handover completion message received from the target cell.

The CHO involves three conditions: adding cells to the target cell list, removing cells from the target cell list and initiating the execution phase, as shown in Fig. 1. For example, if the add condition is met (case number 1 in blue on Fig. 1), the serving cell adds a target cell 1 to the target cell list. Additionally, the UE measures the signal strength of all cells in the target cell list by activating a measurement command until the target cell's signal strength meets the other two conditions, and the serving cell sends a request to reserve and configure radio resources of target cells for the UE [8]. On the other hand, if the remove condition is met (case number 3 in blue on Fig. 1), target cell 1 is removed from the target cell list. Then, the radio resources of target cell 1 are released. These three conditions are part of the preparation phase.

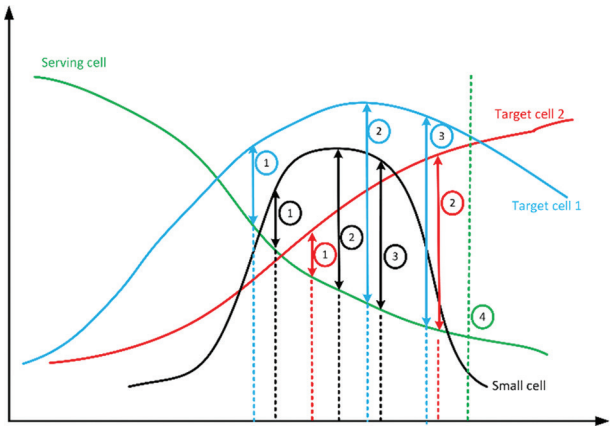


Fig. 1. CHO procedures

Figure 1 also shows the CHO procedures for target cell 1, microcell and cell outage threshold of the serving cell. If all parameters are statically configured and the serving cell's (case number 4 in green on Fig. 1) cell outage threshold is not met, the UE begins the execution phase for handover to a microcell.

The UE is transferred to a smaller cell, and a new CHO procedure starts after a certain time.

Once the execution condition (case number 2 in blue on Figure 1) is met, the UE begins transitioning to the designated target cell at the beginning of the execution phase. Once completing the execution phase, the handover process moves to the completion phase. During this stage, the newly connected cell becomes the new serving cell, and releases the radio resources associated with the previous serving cell as well as those from the target cell list. The three conditions of CHO are introduced below [10].

- Equation 1 shows the add condition of CHO.

$$RSRP_{target} \geq RSRP_{serving} + o_{add} \quad (1)$$

  where $RSRP_{target}$ and $RSRP_{serving}$ are the target and the serving cell's Reference Signal Received Power ($RSRP$), and o_{add} is the offset for all candidate cells listed in the measurement report.

- The remove condition of CHO is introduced in Equation 2.

$$RSRP_{target} \geq RSRP_{serving} - o_{remove} \quad (2)$$

  where $RSRP_{target}$ and $RSRP_{serving}$ are the $RSRP$ of the target and the serving cell, and o_{remove} is the offset for all cells in the target cell list.

- The UE initiates the execution phase by establishing a new connection with a target cell that satisfies Equation 3.

$$RSRP_{target} \geq RSRP_{serving} + o_{exec} \quad (3)$$

  where $RSRP_{target}$ and $RSRP_{serving}$ are the $RSRP$ of the target and the serving cell, and o_{exec} is the offset for all cells in the target cell list.

In summary, CHO introduces the following changes during the preparation and execution phases: a) separation of handover phases to operate independently, enhancing efficiency and reliability; b) creation of a target cell list based on the measurement results for cell selection; c) establishment of clear criteria for adding, removing, and executing handover conditions; and d) pre-configuration of radio resources for all target cells, minimizing transition delays. However, the implementation of CHO presents several constraints:

- A substantial increase in the exchange of control messages.
- Heightened intricacy in prioritizing cells within the target cell list.
- A well-refined algorithm is needed to select the appropriate target cell for execution initiation.
- The need to fine-tune parameter values for conditions.
- The need to optimize execution criteria to initiate handover at the most reasonable time.

2.2. HANDOVER PERFORMANCE INDICATORS

In the context of handover parameter optimization, researchers have introduced metrics to evaluate handover performance. In Saad et al. [11], indicators such as handover probability, ping-pong handover probability, and outage probability are introduced as measures of handover performance. Additionally, the authors analyzed the impact of these handover indicators and conducted a performance analysis of the proposed optimization mechanism in a simulation environment.

HPI is defined as a metric for monitoring the performance of handover procedures for each cell pair. The calculation of HPI involves summing three indicators: Handover Failure (HOF), Handover Ping-Pong (HPP), and Radio Link Failure (RLF). We define HPI as the sum of these indicators.

$$HPI = \omega_{HPP} * HPP + \omega_{RLF} * RLF + \omega_{HOF} * HOF \quad (4)$$

where ω_{HPP}, ω_{RLF}, and ω_{HOF} are the weights for each indicator of handover performance.

These include:

- HPP refers to a ping-pong handover, where the UE begins a new handover procedure to the serving cell after a successful handover to the target cell. HPP is the ratio of ping-pong handovers to total handovers attempted handovers (HO).

$$HPP = \frac{number\ of\ ping{-}pong}{total\ HO} \quad (5)$$

  where number of *ping-pong* is the number of ping-pong handovers, and $total\ HO$ is all attempted handovers.

- RLF can occur when the UE moves out of the coverage area of the serving cell before or during the handover process. RLF is defined as the ratio of the number of radio link failures (RLFs) to the total number of attempted HO.

$$RLF = \frac{number\ of\ RLF}{total\ HO} \quad (6)$$

  where the $number\ of\ RLF$ is the number of RLFs, and the total HO is all attempted handovers. RLF indicates the frequency of RLF.

- HOF is the ratio of the number of handover failures to the total number of attempted handovers. Additionally, HOF also includes wrong cell selection.

$$HOF = \frac{number\ of\ handover\ failures}{total\ HO} \quad (7)$$

  where the number of *handover failures* refers to all handover failure events, and the total *number of* attempted handovers.

2.3. RELATED WORKS

Many studies have focused on CHO and its improvements. The fast conditional handover (FCHO), introduced in [12, 13], enhances the handover process by retaining the target cell list after completion of a

conditional handover (CHO). This allows subsequent handovers to be performed independently, without repeating the full preparation steps, such as configuration and measurements. As a result, FCHO significantly reduces both mobility failures and signaling overhead, as demonstrated in experimental evaluations. In [14], the authors first analyzed real-world data collected from a mobile network to study its configuration and performance. They then improved Iqbal et al.'s FCHO by adapting target cell selection criteria tailored for public transit systems. While their approach effectively reduces signaling overhead in public transportation scenarios, it has limitations, including being specifically designed for a particular use case and not considering velocity in the mobility model. In other words, it focuses on demonstrating parameter selection for specific environments.

In [15], the authors investigated the reallocation of resources during the CHO preparation phase. During target cell evaluation, they utilized beam-specific measurement reports to update Contention Free Random Access (CFRA) resources. The advantage of this approach was demonstrated through experimental results showing reductions in average handover delay and handover failures. However, the results indicated that relying solely on resource allocation was insufficient to significantly improve delay and failure rates.

In [16], the authors proposed an AI-based approach to address measurement report challenges in non-terrestrial networks, where the distance between the user and the base station is significant. Their method involves predicting the handover execution point in the two-step process based on the outcome of the one-step phase. The key experimental finding was that unnecessary handovers and ping-pong handovers occurred more frequently than RLF and HOF, highlighting a significant limitation in the handover decision process. Another notable work on AI-based signal overhead reduction is presented in [17]. The authors introduce an AI-assisted conditional handover using a classifier that performs CHO preparations and manages measurement reports to reduce unnecessary signaling. In this approach, the measurement reports are classified as necessary or unnecessary based on the signal strength received by the user from the base station. According to the simulation results, this method reduced signaling overhead by 53%, and other indicators such as RLF also showed a significant decrease.

In [18], the authors presented a method for adjusting handover hysteresis and time-to-trigger (TTT) based on different UE velocities and RSRP values. The proposed approach demonstrated a reduced ratio of handover failures to total handover through simulation results. Moreover, the mechanism significantly lowered the average number of ping-pong handovers and handover failures compared to other schemes. While the study is similar to ours in showing the relationship between velocity and handover performance, it differs in that it does not take performance-based actions.

Deb et al. [19] presents an analytical evaluation of CHO performance, proposing a Markov model with offsets, time-to-preparation, and time-to-execution. Through experiments and analysis, Deb [19,20] demonstrates that channel fading plays a crucial role in reducing handover failure and latency. The impact of CHO parameters on user velocity is analyzed. Although multiple static configurations of the parameters were tested in the experiments, the results indicate that static tuning alone is insufficient.

Among many AI-based approaches, we focused on those most similar to our work in terms of optimization objectives, performance indicators, and simulation-based evaluation. In [21], Kwon et al. proposed a deep learning-based mechanism to optimize handover in 5G networks by dynamically adjusting the hysteresis value. Their method considered key performance indicators, including handover failures, ping-pong handovers, throughput, and latency. The experimental results demonstrated high throughput and low latency, particularly in high-mobility scenarios.

In Lee et al. [22], a prediction-based deep-learning approach for 5G mmWave networks is discussed. The proposed approach uses deep learning to predict the next base station, making it a compelling method that learns from previously executed handovers to support decision-making and prediction. The researchers' experiments showed that the proposed method achieved an accuracy of 98.8%. By reducing the wrong cell selection, the overall handover performance was improved. In [23], tests were conducted on a 5G emulator with user speeds ranging from 20 to 130 km/h. The proposed adaptive mechanism makes optimal handover decisions based on Received Signal Strength Indicator (RSSI), user direction, and velocity. Experimental results showed that at higher speeds, RSSI decreased, negatively affecting service quality. For example, handover latency was 8 ms at a speed of 80 km/h, and increased to 12 ms at 130 km/h.

Yin et al. propose an approach that combines the advantages of traditional handover and conditional handover (CHO) [24]. In the cell selection phase, the target cell list for CHO is defined with limitations—the selection is based not only on the strongest signal but also on the user's historical handover data. The execution phase is triggered by using the traditional handover execution condition (Event A4). Experimental results showed that, compared to CHO alone, the proposed method reduced total handover failures. However, it also led to an increase in handover attempts, which is a drawback. In [25], the researchers proposed a Q-learning-based reinforcement learning approach for handover decision optimization. Overall, this method aims to improve QoS by reducing handover failures and ping-pong handovers while enhancing overall performance. However, to achieve optimal handover decisions, the approach requires repeated actions over time, which leads to increased processing delays. Additionally, the computational complexity of the method is relatively high.

A reinforcement learning-based adaptive handover approach with optimized decision-making for 5G mmWave bands is presented in [26]. The proposed method predicts the reference signal received power (RSRP) of the target gNB, selects the best target cell from a neighboring cell list, and dynamically determines the handover trigger and hysteresis values. The study analyzes handover success rate, delay, latency, and user throughput using a high-mobility model with speeds up to 200 km/h in an urban test environment. While the handover success rate improved, and handover delay decreased, the approach faces challenges related to computational resource demands and complexity when applied in HetNet environments.

Sattar et al [27] presents a novel rectangular microstrip patch antenna designed for 28 GHz using FR4 substrate, where three feeding techniques are analyzed, showing that proximity coupling significantly enhances gain (from 5.50 dB to 6.83 dB) and bandwidth (from 0.6 GHz to 3.60 GHz), making the antenna highly suitable for 5G applications. [2,28] propose a self-optimization approach for handover hysteresis and time-to-trigger (TTT). This method incorporates 5G network KPIs such as handover probability, handover failures, ping-pong handovers, and radio link failures (RLF) into its calculations. Experiments were conducted at speeds up to 120 km/h, and the results compared three mechanisms. Among them, the velocity-based approach achieved a handover attempt rate similar to the others, while significantly reducing ping-pong handovers and handover failures.

3. AUTOTUNING-BASED PARAMETERS FOR CONDITIONAL HANDOVER

The seamless mobility between cells through efficient handover processes is essential for maintaining high-quality service in cellular networks. Service interruptions or errors during handovers can significantly degrade network performance, resulting in increased latency, reduced data throughput, and connection interruptions. We present an APCHO mechanism that automatically adjusts the parameter values to address these challenges. APCHO automatically changes crucial parameters, such as offsets and the cell outage threshold, based on different network environments and the user's velocity. A distinguishing feature of APCHO is its dual-trigger decision framework. One of the key advantages is that APCHO incorporates the threshold of cell outage as an additional condition, where serves as a second trigger for the handover execution phase.

3.1. AUTO-TUNING PARAMETERS

APCHO runs parameter optimization based on the UE velocity and HPI control. First, our proposal adjusts the offset value of the execution condition based on the UE's velocity. We adjust only the offset parameter of the execution condition introduced in Equation 3. Initially, APCHO calculates the o_{exec} parameter using Equation 8 for every target cell identified in the measurement report. This step ensures that each target cell's execution condition is tailored to its specific characteristics, paving the way for efficient handover decisions.

$$o_{exec} = log_{hysteresis} \frac{(V_{max} - V_{current} + 1)}{V_{max}} \quad (8)$$

where V_{max}, $V_{current}$ are the UE's maximum and current velocities, respectively. *Hysteresis* is a predefined value, the same as in traditional handover. For example, if the UE's velocity is high, the value of o_{exec} is low. After calculating o_{exec}, the serving cell sends configuration to the UE for monitoring the target cell using Equation 3.

The second step involves incorporating HPI into our calculation. While HPI encompasses three indicators, one stands out when handover errors occur, carrying more weight than the others. Our calculation detects this pivotal indicator and dynamically adjusts the offset to mitigate its impact. This means that the value of o_{exec} is fine-tuned based on the primary indicator's effect on HPI, either increasing or decreasing it as needed. Through this adaptive approach, the algorithm effectively mitigates the overall impact of HPI on system performance. Equation 9 shows these HPI-driven adjustments.

$$o_{exec} = o_{exec} \pm \delta \quad (9)$$

where δ is the predefined adjustment value, and o_{exec} is the current value of o_{exec}. For example, when HPP is greater than RLF, APCHO decreases the offset value by δ, but it increases the offset when RLF is greater than HPP.

The third step is to adjust the cell outage threshold based on the difference in cell sizes between the serving and target cells. Equation 10 shows the calculation of $o_{threshold}$. When a UE moves from a macrocell to a microcell, APCHO sets a lower value for $o_{threshold}$. As a result, the UE keep a connection with the macrocell. Conversely, the approach increases the $o_{threshold}$ value as the UE moves away from the microcell.

$$o_{threshold} = o_{threshold} - \frac{S_{serving}}{S_{target}} \quad (10)$$

where $S_{serving}$, S_{target} are the size of the serving and target cells, respectively. $o_{threshold}$ is the current threshold.

When the serving cell's signal strength falls below the threshold, APCHO proactively selects the best target cell, preventing delayed handovers and signal drops. On the other hand, when the cell outage condition is satisfied, APCHO initiates the handover, ensuring a seamless transition and minimizing signal loss (Equation 11).

$$RSRP_{serving} \leq o_{threshold} \quad (11)$$

where $o_{threshold}$ is the cell outage threshold calculated by Equation 10. $RSRP_{serving}$ is the $RSRP$ of the serving cell.

3.2. APCHO PROCEDURES

We implemented a network function in the core network that collects handover information from all cells

and computes the HPI for each pair after every handover procedure. APCHO enables continuous monitoring and uses HPI calculation through the following three steps:

1. HPI is calculated using parameter values for each cell combination. If the HPI is higher than the threshold, the necessary parameters need to be adjusted using Equation 9,
2. The impact of the three HPI indicators is assessed and prioritized based on their influence. The parameter with the greatest impact is identified, and its associated function is used to fine-tune its value accordingly,
3. The HPI is then recalculated.

The message flows of APCHO are presented in Fig. 2.

If the add condition is satisfied (Equation 1), candidate cells are added to the target cell list. The serving cell determines the appropriate values for the offset of the execution condition using Equation 3 and the cell outage threshold using Equation 10.

Step 1 - Measurement command and report: The preparation phase begins when the serving cell sends a measurement command. The UE replies to the serving cell with a report.

Step 2 – CHO decision and add condition: on receiving the report, the serving cell makes the CHO decision based on Equation 1.

Step 3 – HO request: if a handover is necessary, the serving cell sends handover requests to all candidate cells that have just been added to target cell list.

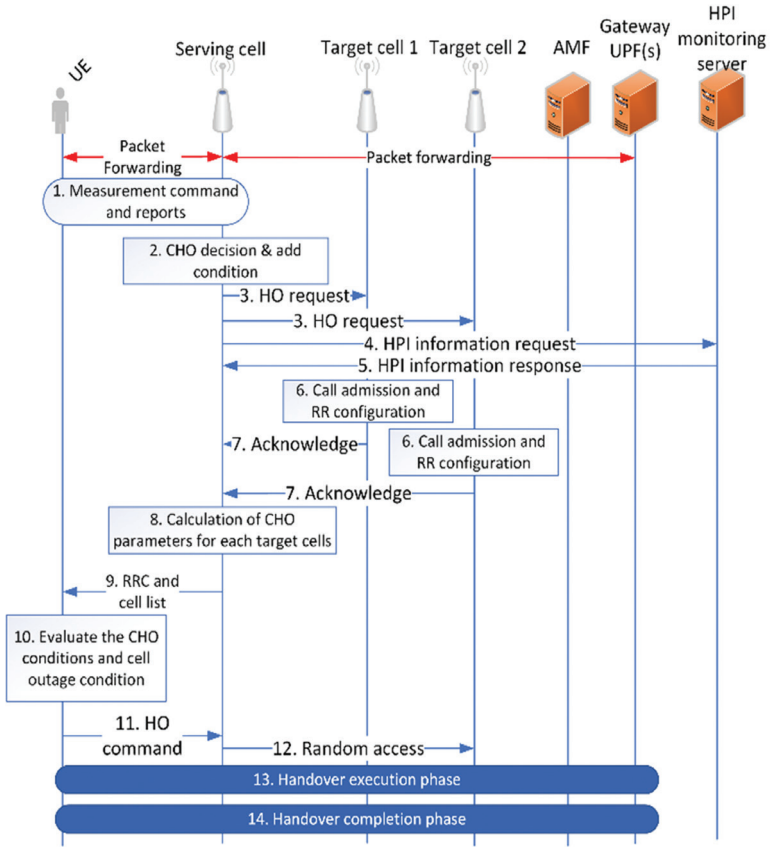


Fig. 2. The message flows of APCHO

Step 4 – HPI information request: simultaneously, the serving cell sends an HPI information request to the calculation server.

Step 5 – HPI information response: the calculation server responds with HPI information for each combination of the serving and target cells. The calculation server already calculated HPI information based on the handover historical information.

Step 6 – Call admission and RR configuration: the target cells accept the handover request and prepare radio resources (RR) for the UE's active services.

Step 7 – Acknowledge: the target cells send an acknowledgment to the serving cell. This acknowledgment includes all information needed for the next phases of the handover procedure.

Step 8 – Calculation of CHO parameters for each target cells: After receiving all acknowledgments and necessary information, the serving cell calculates the execution parameter o_{exec} based on Equation 8 and 9, and the cell outage parameter $o_{threshold}$ based on Equation 10 for each target cell.

Step 9 – RRC and cell list: the serving cell sends a configuration message that includes RRC configuration,

target cell list, and parameters of the execution and cell outage conditions.

Step 10 – Evaluate the CHO conditions and cell outage condition: the UE monitors the target cell list using the CHO conditions and cell outage condition.

Step 11 – HO command: if one of conditions is met, the UE notifies the serving cell and begins the handover execution phase with the selected target cell.

The algorithm starts by initializing the test environment. When the user approaches the edge of the serving cell, the UE sends a measurement report. The serving cell checks neighboring cells against predefined add conditions, creates a target cell list, and decides whether to initiate a handover. If needed, it sends handover requests to the target cells and simultaneously requests additional data from the Handover Performance Indicator (HPI) control server.

After receiving the responses, the serving cell calculates execution and cell outage condition parameters for each target and sends them to the UE. The UE continuously monitors these target cells and initiates handover execution if either the execution condition or the cell outage condition is satisfied. If neither condition is met, the algorithm loops back to reevaluate conditions in the next cycle. Fig. 3 illustrates this iterative process with a blue dashed line labelled "Evaluate CHO and cell outage conditions."

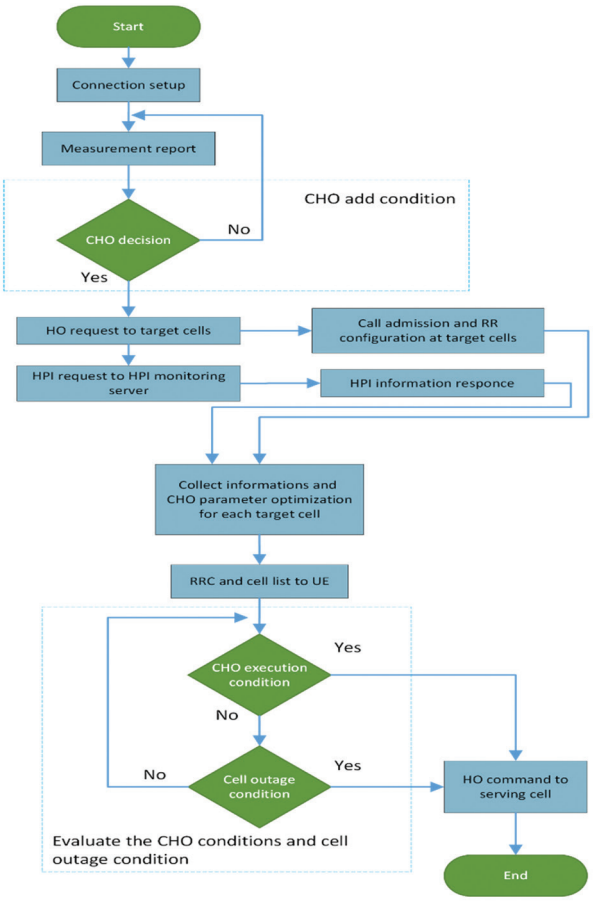


Fig. 3. The flowchart of APCHO

We excluded the CHO remove condition from this flow and focused instead on evaluating the CHO decision logic and performance.

4. SIMULATION RESULTS

In this section, we compare APCHO in a simulation against a standard CHO, a velocity and cell outage-based version. The evaluation is based on handover errors and RLF, averaged over 50 simulation runs using the topology shown in Fig. 4. We started with 100 users and increased the number of users by 100 for each of the 50 runs. The velocity was randomly assigned when placing users at the beginning of the simulation. Additionally, we ensured that 50% of the users had a velocity of less than 80 km/h. Table 1 outlines the parameters used in the simulation. The network topology was implemented with macrocells and microcells added at random locations. The topology advantage of 5G and beyond networks is the microcell and its performance. In [29], the authors introduced the usage of microcell's mode, power consumption, and a heterogeneous dense network topology.

Fig. 5 shows the average RLF ratio, defined as the ratio between the number of RLFs and the number of users. The graph illustrates that RLF ratio for the standard CHO and cell outage mechanism is 2%-6% higher than that of APCHO as the number of handover attempts increases. This is because the static parameters of CHO and cell-outage condition have low effectiveness in preventing too late handovers.

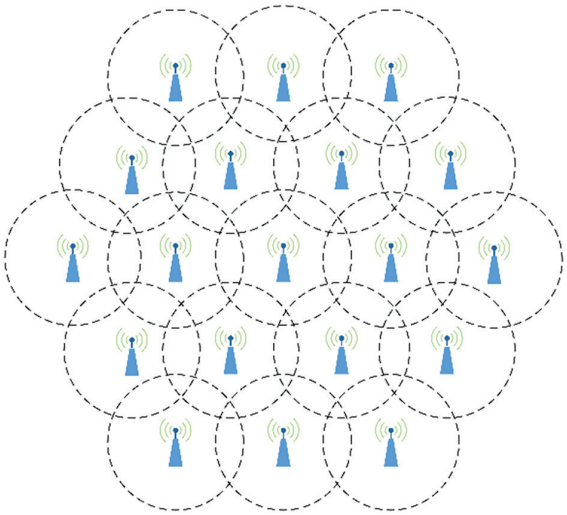


Fig. 4. The simulation topology with 19 macrocell

Note that the number of users and number of attempted handovers are directly related. The velocity-based mechanism produces an RLF ratio is 1%-3% higher than APCHO because o_{exec} is adjusted based on only on velocity. As a result, at 1000 users, the RLF ratio in APCHO is 2%,5%, 6% lower than in the other three mechanisms. This is because APCHO optimizes the parameters using three mechanisms: execution condition parameters, cell outage condition, and HPI calculation-based changes.

Table 1. Simulation parameters

Parameters	Macrocell	Microcell
Carrier frequency (GHz)	2.1	2.1
Bandwidth (MHz)	20	100
The numbers of cells	19	30
Cell radius (m)	500	200
Path loss model	128.1+37.6 log10 (d)	128.1+37.6 log10 (d)
Transmit Power (dBm)	43	21
Overlapping zone (%)	30	0
Antenna Gain (dBi)	5	5
Antenna Gain of UE (dBi)	0	0
Shadowing (dB)	12	7.8
The number of UEs	Up to 1000 (each simulation runs 100 to 1000)	
Velocity of UEs, V_{max} (km/h)	Up to 200 (each user randomly selected)	
UE's mobility model	Random direction	
HPI threshold (%)	2	
o_{add}, o_{remove}, $o_{threshold}$ (dB)	9 dB, 6 dB, -8 dB	
ω_{HPP}, ω_{RLF}, ω_{HOF}	1,1,1	

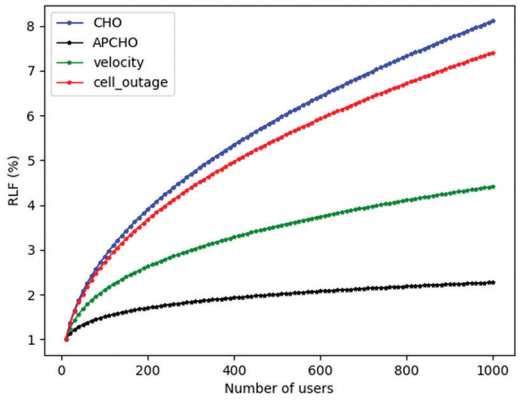


Fig. 5. RLFs versus Number of users

Fig. 6 shows the results for RLFs at velocities ranging from 10 km/h to 200 km/h for all four mechanisms. Below 40 km/h, all four mechanisms show low differences between standard CHO, the velocity-based mechanism, the cell outage-based mechanism and APCHO. The cell-outage mechanism and standard CHO showed similar results at low velocities. On the contrary, it reduced the number of handovers from macrocell to microcell. Starting from 50 km/h, our APCHO and the velocity-based mechanism show a lower growth rate in RLFs compared to the other two mechanisms. The proposed APCHO maintains low RLFs at all velocities, even at 200 km/h.

This effect is achieved through the dynamic adjustment of o_{exec} and $o_{threshold}$. However, as velocity increases beyond 150 km/h, RLFs begin to increase due to the lower RSRP of the serving cell and delayed handovers. CHO and cell-outage mechanisms show a higher growth rate in RLFs. This is because, in these mechanisms, the cell-outage condition is only one of the factors influencing handover decisions, which allows the UE to begin a handover with the target cell too late. We show the average percentage of handover errors with respect to the number of users and their velocity.

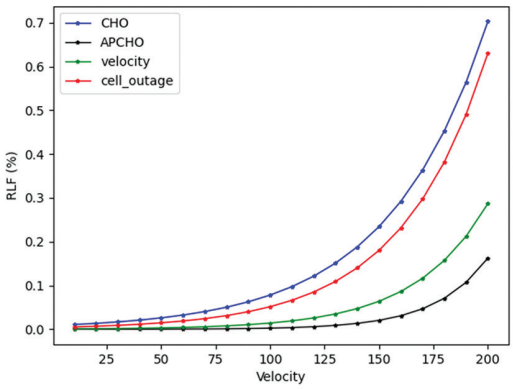


Fig. 6. RLFs versus velocity

Handover errors refer to the number of procedure failures that occur when the UE's handover procedure fails during the execution or completion phases.

Fig. 7 illustrates the comparison of handover errors across four combinations of UE numbers and velocity: low UE density with low velocity (low-low), low UE density with high velocity (low-high), high UE density with low velocity (high-low), and high UE density with high velocity (high-high) environments. Low UE density refers to 100-250 users, while high density refers to up to 1000 users in the simulation area.

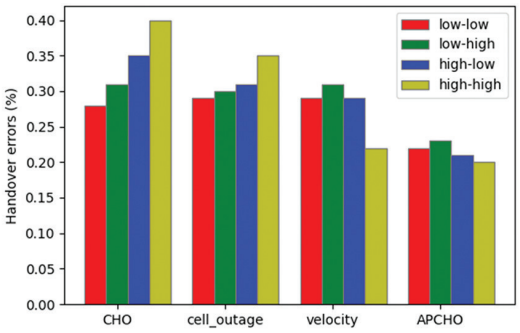


Fig. 7. Handover errors versus the four combinations of UE and velocity

Similarly, low velocity means 10-80 kmph and high velocity means 80-200 kmph. As observed in Figure 6, the baseline CHO shows handover errors of approximately 0.4% in the high-high environment. This is due to the higher number of users, which results in more ping-pong handovers, handover failures, too early handovers, and too late handovers. Additionally, as UE density increases, more handover procedures are attempted, which affects the percentage of handover errors. The velocity-based and APCHO show reduced handover errors in high-low and high-high combinations. The proposed APCHO has few handover errors at all combinations. This is because the server calculates HPI based on handover historical data, and APCHO adjusts the handover parameters for each handover and combinations of target and serving cell.

Fig. 8 illustrates handover performance over the 40-minute test duration. At the start of the simulation

(before 4 minutes), there were minimal differences between APCHO and the velocity-based mechanisms.

As handover failures and the impact of moving velocity began to emerge, the handover error for CHO and cell outage-based mechanism increased noticeably. Notably, CHO displayed a peak and continuous increase, reaching 12% handover errors at 40 minutes. In contrast, the proposed APCHO improved performance by adjusting o_{exec} for each attempted handover. This led to a reduced HPI compared to CHO, especially after 5 minutes into the simulation.

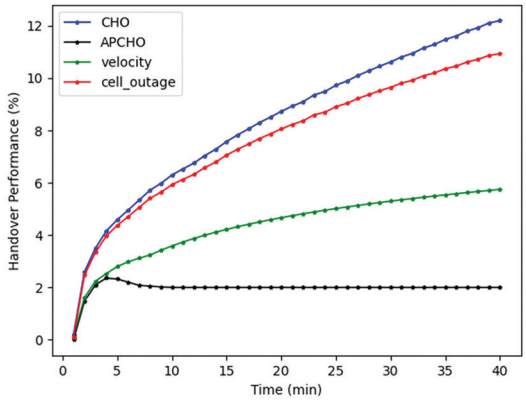


Fig. 8. Performance comparison of mechanisms

As shown in Table 2, the proposed APCHO reduces RLF, ping-pong handovers, and overall handover errors. For example, when ω_{HPP} and ω_{HOF} are set to 1 and ω_{RLF} is changed from 1 to 10, RLF decreased by 2.142%, while HPP and HOF increased by 2.147% and 0.08%, respectively. When adjusting the weighting factors, HOF variation is smaller compared to the variation in HPP and RLFs. For example, when ω_{RLF} and ω_{HPP} are set to 1 and ω_{HOF} is increased from 0.1 to 10, HOF changes by only 0.4%. Therefore, by adjusting the weights of the reward function, the RLFs and HPP experienced by users can be reduced through the optimization of specific types of HO errors.

Table 2. Effect of weight parameters

Weights	Too-early and ping-pong handovers	Too late handovers or RLF	Handover errors
ω_{HPP}=1, ω_{RLF}=1, ω_{HOF}=1	5.364	4.385	1.158
ω_{HPP}=1, ω_{RLF}=10, ω_{HOF}=1	7.511	2.243	1.156
ω_{HPP}=10, ω_{RLF}=1, ω_{HOF}=1	3.318	4.391	1.206
ω_{HPP}=1, ω_{RLF}=10, ω_{HOF}=0.1	5.465	2.158	1.201
ω_{HPP}=10, ω_{RLF}=1, ω_{HOF}=0.1	3.248	4.355	1.92
ω_{HPP}=1, ω_{RLF}=1, ω_{HOF}=10	5.344	3.401	1.116

5. CONCLUSION

We have introduced Autotuning-based Parameters for Conditional Handover (APCHO), an enhanced version of CHO that incorporates autotuning parameters and second handover trigger to improve 5G handover performance. In addition, we calculated the HPI to automatically adjust parameters when error thresholds were exceeded. Our proposed APCHO mechanism also uses mobility management performance data to dynamically adjust the weight of parameters to mitigate handover errors impact. We evaluated the mobility performance using a simulator for 5G HetNets. The simulation results showed that APCHO reduced HOF and RLFs compared to standard CHO.

ACKNOWLEDGEMENTS

This work was supported by the National University of Mongolia, grant number P2022-4381.

6. REFERENCES:

[1] U. Gustavsson, P. Frenger, C. Fager, T. Eriksson, H. Zirath, F. Dielacher, C. Studer, A. Pärssinen, R. Correia, J. N. Matos, D. Delo, N. B. Carvalho, "Implementation challenges and opportunities in beyond-5G and 6G communication", IEEE Journal of Microwaves, Vol. 1, No. 1, 2021, pp. 86-100.

[2] I. M. Abbas, T. Y. Hoe, C. Ali, A. D. Jamal, "Handover Optimization Framework for Next-Generation Wireless Networks: 5G, 5G– Advanced and 6G", Proceedings of the IEEE International Conference on Automatic Control and Intelligent Systems, Shah Alam, Malaysia, 29 June 2024, pp. 409-414.

[3] I. Begić, A. S. Kurdija, Željko Ilić, "A Framework for 5G Network Slicing Optimization using 2-Edge-Connected Subgraphs for Path Protection", International Journal of Electrical and Computer Engineering Systems, Vol. 15, No. 8, 2024, pp. 675-685.

[4] D. Chandramouli, R. Liebhart, J. Pirskanen, "5G for the Connected World", 1st Edition, John Wiley & Sons, 2019.

[5] H. Martikainen, I. Viering, A. Lobinger, T. Jokela, "On the basics of conditional handover for 5G mobility", Proceedings of the IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications, Bologna, Italy, 9-12 September 2018, pp. 1-7.

[6] I. Viering, H. Martikainen, A. Lobinger, B. Wegmann, "Zero-zero mobility: Intra-frequency handovers with zero interruption and zero failures", IEEE Network, Vol. 32, No. 2, 2018, pp. 48-54.

[7] Intel, "3gpp tdocs, performance evaluation of conditional handover", Intel Corporation, Chengdu, China, Technical Report R2-1814052, No. R2-103b, 2018.

[8] Nokia, "3gpp, conditional handover basic aspects and feasibility in rel-15", Nokia Shanghai Bell, Berlin, Technical Report R2-1708586, No. R2-99, 2017.

[9] U. Karabulut, A. Awada, I. Viering, A. N. Barreto, G. P. Fettweis, "Analysis and performance evaluation of con-

ditional handover in 5G beamformed systems", arXiv:1910.11890, 2019.

[10] J. Stanczak, U. Karabulut, A. Awada, "Conditional handover in 5G-principles, future use cases and FR2 performance", Proceedings of the International Wireless Communications and Mobile Computing, Dubrovnik, Croatia, 30 May - 3 June 2022, pp. 660-665.

[11] W. K. Saad, I. Shayea, B. J. Hamza, H. Mohamad, Y. I. Daradkeh, W. A. Jabbar, "Handover parameters optimisation techniques in 5G networks", Sensors, Vol 21, No. 15, 2021, pp. 1-22.

[12] S. B. Iqbal, A. Awada, U. Karabulut, I. Viering, P. Schulz, G. P. Fettweis, "On the modeling and analysis of fast conditional handover for 5g-advanced", Proceedings of the IEEE 33rd Annual International Symposium on Personal, Indoor and Mobile Radio Communications, Kyoto, Japan, 12-15 September 2022, pp. 595-601.

[13] S. B. Iqbal, S. Nadaf, A. Awada, U. Karabulut, P. Schulz, G. P. Fettweis, "On the analysis and optimization of fast conditional handover with hand blockage for mobility", IEEE Access, Vol. 11, 2023, pp. 30040-30056.

[14] C. Fiandrino, D. J. Martínez-Villanueva, J. Widmer, "A study on 5G performance and fast conditional handover for public transit systems", Computer Communications, Vol. 209, 2023, pp. 499-512.

[15] J. Stanczak, U. Karabulut, A. Awada, "Conditional handover modelling for increased contention free resource use in 5g-advanced", Proceedings of the IEEE 34th Annual International Symposium on Personal, Indoor and Mobile Radio Communications, Toronto, ON, Canada, 5-8 September 2023, pp. 1-6.

[16] E. Kim, I. Joe, "Handover triggering prediction with the two-step xgboost ensemble algorithm for conditional handover in non-terrestrial networks", Electronics, Vol. 12, No. 16, 2023, pp. 1-24.

[17] A. Gharouni, U. Karabulut, A. Enqvist, P. Rost, A. Maeder, H. Schotten, "Signal overhead reduction for AI-assisted conditional handover preparation", Proceedings of Mobile Communication-Technologies and Applications; 25th ITG-Symposium, Osnabrueck, Germany, 3-4 November 2021, pp. 1-6.

[18] A. Alhammadi, M. Roslee, M. Y. Alias, I. Shayea, S. Alraih, K. S. Mohamed, "Auto tuning self-optimization algorithm for mobility management in LTE-A and 5G HetNets", IEEE Access, Vol. 8, 2019, pp. 294-304.

[19] S. Deb, M. Rathod, R. Balamurugan, S. K. Ghosh, R. K. Singh, S. Sanyal, "Performance evaluation of conditional handover in 5G systems under fading scenario", arXiv:2403.04379, 2024.

[20] S. Deb, M. Rathod, R. Balamurugan, S. K. Ghosh, R. K. Singh, S. Sanyal, "Evaluating Conditional handover for 5G networks with dynamic obstacles", Computer Communications, Vol. 233, 2025, p. 108067.

[21] C. F. Kwong, S. Chenhao, L. Qianyu, Y. Sen, C. David, K. Pushpendu, "Autonomous handover parameter optimisation for 5G cellular networks using deep deterministic policy gradient", Expert Systems with Applications, Vol. 246, 2024, pp. 1-8.

[22] C. Lee, H. Cho, S. Song, J. M. Chung, "Prediction-based conditional handover for 5G mm-wave networks: A deep-learning approach", IEEE Vehicular Technology Magazine, Vol. 15, No. 1, 2020, pp. 54-62.

[23] T. Y. Arif, A. Fitri, "Adaptive Strategies and Handover Optimization in 5G Networks to Address High Mobility Challenges", Proceedings of the IEEE International Conference on Electrical Engineering and Informatics, Banda Aceh, Indonesia, 12-13 September 2024, pp. 153-157.

[24] Z. Yin, "Mobility Performance Analysis for Mixed Handover", Proceedings of the IEEE Wireless Communications and Networking Conference, Dubai, United Arab Emirates, 21-24 April 2024, pp. 1-6.

[25] S. C. Sundararaju, R. Shrinath, P. B. Dandra, P. Vanama, "Advanced Conditional Handover in 5G and Beyond Using Q-Learning", Proceedings of the IEEE Wireless Communications and Networking Conference, Dubai, United Arab Emirates, 21-24 April 2024, pp. 1-6.

[26] Y. S. Junejo, K. S. Faisal, S. C. Bhawani, E. Waleed, "Adaptive Handover Management in High-Mobility Networks for Smart Cities", Computers, Vol. 14, No. 1, 2025, pp. 1-27.

[27] S. O. Hasan, S. K. Ezzulddin, O. S. Hammd, R. H. Mahmud, "Design and Performance Analysis of Rectangular Microstrip Patch Antennas Using Different Feeding Techniques for 5G Applications", International Journal of Electrical and Computer Engineering Systems, Vol. 14, No. 8, 2023, pp. 833-841.

[28] I. M. Abbas, T. Y. Hoe, C. Ali, A. D. Jamal, "Self-optimization of handover control parameters for 5G wireless networks and beyond", IEEE Access, Vol. 12, 2023, pp. 6117-6135.

[29] J. Premalatha, A. S. A. Nisha, S. E. V. Somanathan Pillai, A. Bhuvanesh, "Performance Measurement of Small Cell Power Management Mechanism in 5G Cellular Networks using Firefly Algorithm", International Journal of Electrical and Computer Engineering Systems, Vol. 15, No. 5, 2024, pp. 437-448.

Experimental Study and Modeling of Radio Wave Propagation for IoT in Underground Wine Cellars

Original Scientific Paper

Snježana Rimac-Drlje*
J. J. Strossmayer University of Osijek,
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Kneza Trpimira 2B, Osijek, Croatia
snjezana.rimac@ferit.hr

Tomislav Keser
J. J. Strossmayer University of Osijek,
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Kneza Trpimira 2B, Osijek, Croatia
tomislav.keser@ferit.hr

Vanja Mandrić
J. J. Strossmayer University of Osijek,
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Kneza Trpimira 2B, Osijek, Croatia
vanja.mandric@ferit.hr

Slavko Rupčić
J. J. Strossmayer University of Osijek,
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Kneza Trpimira 2B, Osijek, Croatia
slavko.rupcic@ferit.hr

*Corresponding author

Abstract *– This paper presents the results of a study of radio wave propagation in underground wine cellars in the context of the optimal use of wireless communication systems for the application of the Internet of Things (IoT) in wine production environments. Electric field strength measurements were carried out in two subterranean line-of-sight (LOS) and non-line-of-sight (NLOS) conditions at 860 MHz, 2400 MHz, and 3600 MHz. The measured results were compared with predictions from seven existing propagation models, including site-general models (Free-space, ITU-R P.1238-13, ITU-R P.1411-12, ETSI TR 138 901 V16.1.0) and site-specific models (ITU-R P.1411-12, tunnel and knife-edge diffraction). Statistical analysis determined that the ITU-R P.1238-13 model, which estimates path loss in corridors, and the tunnel model have the best agreement with measurements in LOS conditions, with average Root Mean Square Error (RMSE) values of 2.8 dB and 3.56 dB, respectively. For the NLOS regions, the knife-edge diffraction model achieved the highest accuracy (average RMSE = 2.77 dB). Furthermore, based on the measurement results, the coefficients of the general path-loss model were adjusted to the data using the least-squares method, yielding RMSEs of 2.02 dB for LOS and 2.65 dB for NLOS conditions. The analysis showed that combining the ITU-R P.1238-13 or tunnel model for LOS conditions with a diffraction-based model for NLOS conditions provides a good basis for modeling radio propagation in subterranean wine cellars. These findings support the design of efficient and robust IoT communication networks in winery environments.*

Keywords: *Internet of Things (IoT), underground wine cellar, radio wave propagation, path loss modeling*

Received: October 16, 2025; Received in revised form: November 11, 2025; Accepted: November 12, 2025

1. INTRODUCTION

The application of smart technologies in production processes improves product quality while reducing energy consumption and conserving other essential resources. The application of sensors plays a major role in this, and the rapid progress of the Internet of Things (IoT), sensor technologies, and various communication solutions that enable sensor connectivity have created significant opportunities for the application of wireless sensor networks even in traditional industrial sectors such as wine production [1–6]. Due to the inherent variability of grapes, whose properties are influenced by meteorological conditions, soil composition, and agronomic practices during cultivation, the wine production process is not standardized and must be continually adapted to the current state of the must or wine. Consequently, continuous monitoring and measurement of production parameters are essential to identify potential deviations and take timely corrective actions. The collection and processing of various data requires reliable transmission of measurements to a centralized unit where predictive models developed using machine learning techniques can be used to predict the behavior of parameters and thus support process management and optimization [1-4].

Given that a large number of wine containers are typically distributed across a wide area, wireless communication is the most practical and efficient means of data transmission. Considering that the volume of transmitted data is relatively modest, robust communication systems such as LoRa or NB-IoT are suitable; however, WiFi-based solutions can also be acceptable alternatives. When designing a radio communication system, it is essential to select an appropriate model for estimating propagation losses of radio waves. Propagation models are intensively researched in different scenarios (indoor, outdoor, outdoor-to-indoor), and environments (urban, suburban, rural, indoor corridors/ rooms/multi-floor, tunnels), and for different frequencies, depending on the model's area of application [7-14]. However, there are only a few studies on radio wave propagation in basements with architectural features such as wine cellars, [15, 16].

Wine production facilities can generally be divided into two categories: traditional cellars and purpose-built structures. These two environments differ significantly in architectural design and construction materials, which in turn lead to variations in radio wave propagation behavior. This paper focuses on the propagation of radio waves in traditional wine cellars, which are typically characterized by their subterranean location, thick brick or stone walls, absence of windows, elongated corridors with low ceilings, and frequently vaulted structures. The high humidity commonly found in wine cellars influences the electromagnetic properties of the building materials [17], consequently affecting the absorption and reflection of electromagnetic waves. Moreover, the wine containers themselves have a distinctive impact on wave propagation, depending on their size, spatial arrangement, and material composition. Stainless steel tanks, which are widely used in modern wineries, act as reflectors of radio waves due to their high conductivity, and direct the reflections in a specific way because of their rounded shape. As additional causes of reflection and diffraction of radio waves, they make propagation conditions in the cellars even more complex.

Given the similar architectural characteristics, models developed for radio wave propagation in tunnels or corridors [18–22] represent potential candidates for modeling propagation losses in wine cellars. However, to the best of our knowledge, the applicability and performance of such models in wine cellars have not yet been systematically analyzed. Accordingly, the objective of this study is to investigate how the unique structural and environmental features of underground wine cellars influence radio wave propagation, and to assess the suitability of existing propagation models for predicting attenuation in these environments.

In [16], we presented measurements of electric field strength in a basement environment with characteristics similar to a wine cellar, made for two frequencies, 860 MHz and 2.4 GHz. Those results were compared with predictions derived from the Free-Space (FS) propagation model [7], the tunnel propagation model [11], and the knife-edge diffraction model [13]. The present study extends this earlier research by performing measurements in two distinct basement environments, at three operating frequencies, and by comparing the results with seven existing propagation models.

The main contributions of this work are as follows:

- Radio wave propagation loss measurements were performed in two types of cellars, one semi-empty and one containing stainless-steel wine vessels, under both Line-of-Sight (LOS) and Non-Line-of-Sight (NLOS) conditions, at three frequencies: 860 MHz, 2400 MHz, and 3600 MHz.
- Seven propagation models, originally developed for environments with characteristics comparable to underground wine cellars, were analyzed and evaluated.
- The measured propagation losses and the corresponding values predicted by each model were compared, separately for LOS and NLOS conditions. Based on statistical analysis, the models that most accurately estimated losses as a function of the distance between the transmitter and receiver were identified, making them the most suitable for application in the design of radio networks in wine cellars.

The remainder of this paper is organized as follows. An overview of related work on modeling radio wave propagation in environments similar to underground wine cellars (tunnels, corridors, and street canyons) is provided in Section 2. Section 3 presents the experimental setup and measurement methodology. The propagation models used in this study are presented in Section 4, as well as a statistical analysis of the measurement results and propagation losses obtained by the propagation models. Finally, Section 5 provides the main conclusions of the study.

2. RELATED WORK

Radio wave propagation modeling in complex and confined environments has been the subject of extensive research, particularly in relation to emerging wireless technologies such as LoRa, NB-IoT, and 5G systems. Azevedo and Mendonça [23] provided a comprehensive critical review of propagation models employed in LoRa systems, identifying their suitability and limitations for different deployment scenarios. Their analysis found that most empirical models use logarithmic attenuation dependence on distance, as well as the need for environment-specific modeling, especially for NLOS conditions and constrained spaces. Similarly, Alobaidy et al [14] examined channel propagation models for IoT technologies, highlighting the challenges of achieving reliable wireless transmission in difficult conditions. These findings are consistent with broader studies on IoT and Industrial IoT (IIoT) communications, such as Ji-

ang et al. [13], who reviewed standardized 5G channel models for industrial environments, illustrating how 3GPP-defined frameworks can support next-generation industrial communications networks.

Over the years, a significant amount of research has been conducted on radio wave propagation in underground spaces and tunnels, environments that share physical and electromagnetic similarities with wine cellars. Emslie et al. [11] developed theoretical models for UHF propagation in coal mine tunnels, laying the foundation for subsequent empirical and semi-deterministic models for similar environments. Subrt and Pechac [18] proposed a semi-deterministic model for underground galleries and tunnels, while Rak and Pechac [24] analyzed UHF propagation in caves and underground passages, showing the influence of geometry and material properties on attenuation and multipath behavior. A comprehensive review by Hrovat et al. [20] and a more recent review by Samad et al. [19] provide systematic reviews of propagation modeling techniques in tunnel environments, including deterministic, empirical, and hybrid approaches. Empirical investigations of wireless sensor network (WSN) deployment in complex utility tunnels by Celaya-Echarri et al. [25] combined radio wave propagation analysis with network optimization methods. Li et al. [26] analyzed tunnel channels with human presence at 6 GHz, enhancing the understanding of body-induced multipath effects relevant to 5G communications.

Measurements in indoor environments provide further insight into modeling radio propagation in heterogeneous and obstructed spaces. Tan et al. [27] performed multipath delay measurements in multi-floor wireless communication scenarios. Similarly, Samad et al. [21] evaluated large-scale propagation models in indoor corridors at 3.7 and 28 GHz, validating their performance for 5G mobile network deployments.

Several studies have explored propagation in outdoor-to-deep-indoor settings relevant to underground industrial and agricultural facilities. Malarski et al. [28] and [29] investigated NB-IoT signal attenuation in deep-indoor conditions, providing empirical data on the importance of specific environmental features for sub-GHz signal strength prediction. Ali et al. [30] proposed a propagation loss model for neighborhood area networks in smart grids, integrating environmental variability into large-scale path loss estimation for an outdoor-to-deep-indoor scenario.

The propagation of radio waves in natural or artificial underground environments has also been examined through field measurements. Branch [31] conducted measurements of LoRa propagation in an underground gold mine and analyzed attenuation and multipath effects at 915 MHz. Soo et al. [15] investigated propagation behavior within a natural cave adapted for wine storage, providing one of the earliest studies linking radio wave propagation to wine cellar conditions. Together, these studies highlight the need for environment-specific propagation models tailored to the unique structural and material properties of cellars used in wine production

Although previous research has made significant progress in understanding the propagation of radio waves in various environments, including corridors and tunnels, the electromagnetic characterization of underground wine cellars remains insufficiently explored. The specific architectural features, high humidity levels, and presence of conductive storage vessels introduce propagation phenomena that differ from those observed in typical tunnel or building environments. These gaps motivate further efforts aimed at accurately characterizing the behavior of radio waves in wine cellars and optimizing wireless communication systems for smart wine production.

3. EXPERIMENTAL SETUP AND MEASUREMENT METHODOLOGY

The measurements were carried out in two different cellars. The first one, hereafter Cellar A, is the basement of a 19th-century building, completely built below ground level. This cellar is not used for wine production, but has all the architectural and environmental characteristics of a traditional underground wine cellar. Although it does not contain wine vessels, the space includes several wooden tables and shelves with metal components, as well as metal objects such as radiators and fire extinguishers (Fig. 1a). These objects are not typical of wine cellars, but their reflective and diffractive properties adequately mimic objects that similarly affect the propagation of radio waves in such environments.

Cellar A consists of six rooms (Fig. 1a), separated by brick walls approximately 74 cm thick. The ceiling is slightly arched, with a maximum height of 2.60 m at the center. At the beginning of the measurement campaign, the humidity in the basement was 65.3 %, and the temperature was 24.5 °C; by the end of the measurement, these values had dropped to 56.7 % and 21.8 °C due to additional ventilation. These conditions are typical for underground wine cellars, where high humidity inside the cellars indicates high humidity in the walls due to the porosity of the building material (brick, plaster) from which they are made [32]. The bricks' conductivity and relative dielectric permittivity, the quantities which determine the reflection coefficient, increase with the percentage of moisture [17, 33]. Consequently, this directly affects the power and phase shift of the reflected waves, and thus the total radio signal strength at the receiving site.

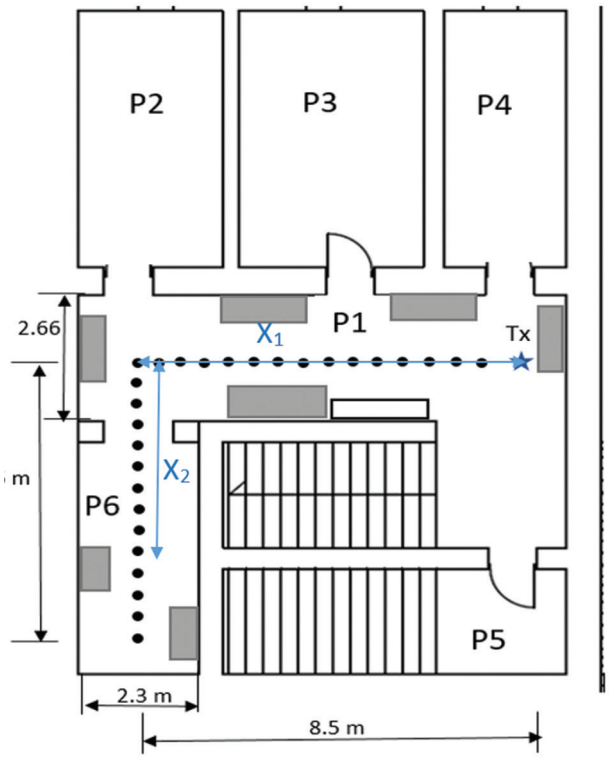
Measurements in Cellar A were carried out in rooms P1 and P6, with LOS propagation conditions achieved in P1 and NLOS in P6. The transmitting antenna was placed at the entrance to room P1 (marked with a star in Fig. 1a), and the measurement was carried out through the middle of the room at measurement points that were 0.5 m apart, starting from a distance of 1 m from

the transmitting antenna (the measurement locations are marked with dots in Fig. 1a).

The second facility, Cellar B, is an operational underground wine cellar of the Horvat Winery located in the Slavonija - Baranja County in eastern Croatia. Cellar B consists of a single elongated room containing 15 wine vessels of various sizes (Fig. 2a). The ceiling is slightly arched, with a maximum height of 3.18 m at the center. Ambient humidity and temperature were measured as 78.3 % and 18.7 °C, respectively. The measurement in Cellar B was carried out through the middle of the room, under LOS conditions of radio wave propagation.

The distance between the measurement points was also 0.5 meters, and their positions as well as the position of the transmitting antenna are shown in Fig. 2b, using the same notations as for Cellar A.

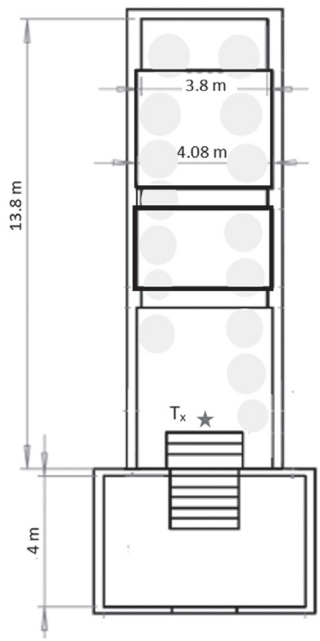
(a)



(b)

Fig. 1. a) Measurement setup in Cellar A; **b**) Cellar A plan

(a)



(b)

Fig. 2. a) Cellar B; **b**) Cellar B plan

The characteristics of Cellars A and B are summarized in Table 1.

Table 1. Architectural and environmental characteristics of the cellars

	Key characteristics of the cellars
Cellar A (Fig. 1)	• completely underground • 6 rooms, separated by 74 cm thick brick walls • several desks and shelves with metallic parts • LOS measurements in P1 and NLOS in P2 • P1 width W_{A1}=2.66 m, P2 width W_{A2}=2.33 m • arched ceiling with maximum height H_A=2.60m • humidity 56.7 % - 65.3 %, temperature 21.8 °C - 24.5 °C
Cellar B (Fig. 2)	• completely underground • 1 room, filled with large steel containers • LOS measurements • width W_B =4.08 m • arched ceiling with maximum height H_B=3.18m • humidity 78.3%, temperature 18.7°

Given the specific architectural and material features of both cellars, it was expected that signal strength at a given location would be significantly influenced by reflections from walls, ceilings, floors, and internal objects, as well as by diffraction at object edges and edges of walls at room transitions (particularly between P1 and P2 in Cellar A). Due to such multipath propagation, large variations in field strength can occur over small distances, even below $\lambda/2$. To mitigate these small-scale fluctuations, multiple measurements were taken within a circular area of 12 cm in diameter centered at each measurement location. The mean value of these measurements was used for further analysis.

To encompass most of the frequency bands used by Wi-Fi, LoRa, and NB-IoT systems, measurements were performed at 860 MHz, 2400 MHz, and 3600 MHz. Microwave signal generator ANA-PICO APSIN20G served as the transmitter, paired with vertically polarized dipole and logarithmic antennas, in dependence on frequency. A calibrated isotropic antenna Rhode&Schwarz TSEMF B1, and a spectrum analyzer Rhode&Schwarz FSH8 were used for electric field strength measurements. Detailed specifications of all equipment are provided in Table 2. Both the transmitting (Tx) and receiving (Rx) antennas were mounted on wooden tripods at a height of 1.5 m (Fig. 1a). The transmitter power was set to 15 dBm for all measurements.

The path loss (PL), expressed in decibels (dB), representing the attenuation of the radio wave between the Tx and Rx antennas, was calculated as:

$$PL[\mathrm{dB}] = P_{Tx}[\mathrm{dBW}] + G_{Tx}[\mathrm{dB}] - L_{cTx}[\mathrm{dB}] - (P_{Rx}[\mathrm{dBW}] + L_{cRx}[\mathrm{dB}] - G_{Rx}[\mathrm{dB}]) \quad (1)$$

P_{Tx} is the transmitter power, G_{Tx} and G_{Rx} are the antenna gains, L_{cTx} and L_{cRx} are cable losses, and P_{Rx} is received power. Antenna gains were derived from the antenna factors (AF) according to [34]:

$$G_{Tx}[\mathrm{dB}] = -149.7 + 20\log_{10} f[\mathrm{Hz}] - AF[\mathrm{dB/m}]. \quad (2)$$

Received power was computed from the measured electric field strength (E) using the following relations:

$$P_{Rx}[\mathrm{dBW}] = 10\log_{10}\frac{U^2}{R} = 10\log_{10}\frac{E^2}{AF_{lin}^2 R} \quad (3)$$

$$P_{Rx}[\mathrm{dBW}] = 20\log_{10}E[\mathrm{V/m}] - 20\log_{10}AF_{lin}[\mathrm{m^{-1}}] - 10\log_{10}R[\Omega] \quad (4)$$

or equivalently,

$$P_{Rx}[\mathrm{dBW}] = E[\mathrm{dBV/m}] - AF[\mathrm{dB/m}] - 10\log_{10}50 \quad (5)$$

Table 2. Experiment parameters

	Key data
Transmitter	Microwave signal generator ANA-PICO, APSIN20G P_{Tx} =15 dBm
Receiver	Spectrum analyzer ROHDE SCHWARZ FSH8
Operating frequencies	860 MHz, 2400 MHz and 3600 MHz
Transmitter antenna	Vertically polarized dipole antenna Seibersdorf Laboratories Precision Conical Dipole PCD 8250 for 860 MHz and 2400 MHz Vertically polarized antenna HyperLOG 30180 Aaronia for 3600 MHz
Receiver antenna	Isotropic antenna Rohde & Schwarz TSEMF B1 EMF Measurement Isotropic Probe
Extended uncertainty of measuring set	3.49 dB for f < 3 GHZ 3.99 dB for f ≥ 3 GHZ
Antenna height	h_{Tx} = 1.5 m, h_{Rx} = 1.5 m
Measurement points	28 at Cellar A, 21 at Cellar B 5-10 measurements at each measuring point - total of 560 and 315 measurements in cellar A and B, respectively

Cable losses and the receiving antenna factor were obtained from its calibration data, while the transmitting antenna factor was taken from [34].

For clarity, measurement results are expressed as path gain (PG), defined as the negative of path loss:

$$PG[\mathrm{dB}] = -PL[\mathrm{dB}] \quad (6)$$

Since path loss increases with distance, path gain correspondingly decreases, reflecting the reduction of received power as the receiver moves away from the transmitter.

4. PROPAGATION MODELS AND COMPARISON WITH MEASUREMENTS

4.1. RADIO WAVE PROPAGATION MODELS

We tested seven standardized propagation models for path loss prediction in underground wine cellars. For testing, we chose propagation models that were developed for environmental configurations similar to the characteristics of wine cellars.

The International Telecommunication Union Radiocommunication Sector (ITU-R) Recommendation ITU-R P.525-3, [7], provides the reference free-space propagation model, which is useful as a baseline for short-range links or to quantify the additional attenuation

introduced by the cellar environment. However, this model does not account for reflections, absorption, or scattering from walls and metallic objects, which are dominant effects in enclosed underground spaces. The formulas used in this paper for the free-space propagation loss model, as well as for the other path-loss models, are given in Table 3.

Indoor environments with corridor-like geometries are more appropriately represented by ITU-R P.1238-13, [8], which defines a site-general model for propagation in buildings and corridors. The vaulted structure and narrow passages of wine cellars closely resemble the conditions assumed in this recommendation, making it a suitable starting point for empirical path loss estimation. Nevertheless, parameter tuning is typically required, since ITU-R P.1238 is based on modern building materials rather than stone or brick walls typical of historic cellars.

The ITU-R P.1411-12, [9], site-general and site-specific models, originally developed for street-canyon propagation, can also be considered when the cellar exhibits elongated and narrow geometries. Namely, street-canyon propagation is characterized by multiple reflections from buildings and the ground that cause a waveguide effect, which can also be expected for long basement corridors.

Since wine cellars contain large steel wine tanks, the radio wave propagation conditions could be similar to those in industrial buildings. To investigate this, we have considered the site-general model for indoor propagation in factory environments with dense clutter and low base station heights, defined in ETSI TR 138 901 V16.1.0, [10]. The higher path loss exponents and increased shadowing predicted by this model are consistent with environments dominated by large reflective objects and strong multipath components.

The tunnel propagation model, originally proposed by Emslie et al. [11], describes the behavior of UHF (Ultra High Frequency) radio waves in tunnel-like environments where multiple reflections from the walls, floor, and ceiling create a waveguiding effect. The model divides propagation into two regions: a free-space region close to the transmitter and a waveguide region beyond the Fresnel breakpoint distance, where attenuation increases approximately linearly with distance. This formulation effectively captures the transition from spherical to guided propagation and has since been widely applied to model radio wave behavior in underground and enclosed environments with elongated geometries.

In addition, ITU-R P.526-14 [12], which describes the classical knife-edge diffraction method, is relevant for modeling wave propagation at transitions between rooms, doorways, or the edges of large metallic containers. In such cases, diffraction losses can be estimated separately and combined with the baseline statistical model to improve prediction accuracy.

4.2. PROPAGATION MODELS VS. MEASUREMENTS

The measurement results are presented in terms of path gain, calculated using the equations described in Section 3. The electric field strength at each measurement point represents the mean of five to ten measurements taken within a circular area of 12 cm in diameter centered on that point. Accordingly, the reported results show the average path gain as a function of the distance between the transmitting and receiving antennas. Table 4 lists the standard deviations of the individual measurements related to their respective mean values, which range from 2.09 dB to 3.06 dB, indicating the level of the field strength variations in that small area.

The measurement results for both cellars are presented in Fig. 3. In each case, the path gain decreases with increasing frequency, as expected. A pronounced reduction in path gain is also evident in Cellar A within the NLOS region.

A direct comparison between Cellar A and Cellar B is possible only under LOS conditions, i.e., for all points in Cellar B and up to 9 m in Cellar A. The steel wine containers in Cellar B were expected to cause stronger attenuation. This was confirmed at 2.4 GHz, where the average field strength in Cellar B was 4.3 dB lower than in Cellar A. At 860 MHz, however, the difference was only 1.4 dB, likely due to the longer wavelength and weaker interaction of the wave with metallic surfaces.

Propagation was also influenced by transmitter antenna placement. In Cellar A, the transmitter was located 1 m from a brick wall, with openings to adjacent rooms on both sides. In Cellar B, it was positioned near a wide opening to another chamber, bounded laterally by a wall and a metal wine container. These geometric differences contributed to the observed variation in path gain.

Measured path gains are compared with the theoretical free-space model in Fig. 3. In Cellar B, the agreement at a 1 m Tx–Rx separation is within 1 dB. In Cellar A, measured path gains at distance of 1 m exceed free-space predictions by 2.7 dB at 860 MHz, 4.3 dB at 2.4 GHz, and 4.2 dB at 3.6 GHz, primarily due to reflections from nearby surfaces. For larger distances, the free-space model significantly overestimates path loss in both cellars, confirming the strong effects of multipath propagation and absorption in enclosed environments.

In addition to the free-space propagation model, the measured path gain results were compared with the predictions obtained from six additional propagation models, which can be categorized into two groups.

The first group comprises site-general models, which incorporate environmental characteristics indirectly through empirical parameters α, β, and γ, fitted for the type of environment under consideration. These parameters appear in the general path loss formulation

$$PL(d,f) = 10\alpha \cdot \log_{10}(d) + \beta + 10\gamma \cdot \log_{10}(f) \quad (7)$$

Table 3. Selected propagation path-loss models

Propagation model	LOS/ NLOS	Path loss, PL (dB) (f is in GHz, d is in m)	Shadow fading std (dB)	Frequency range (GHz)	Distance range (m)
ITU-R P.525-3 Free space model, [7]	LOS	$PL_{FS}(d,f) = 92.4 + 20 \cdot \log_{10}(d) + 20 \cdot \log_{10}(f)$			
ITU-R P.1238-13 Site-general model for propagation in corridors, [8]	LOS	$PL_b(d,f) = 15.7 \cdot \log_{10}(d) + 29.46 + 22.4 \cdot \log_{10}(f)$	σ=3.77	0.3-300	2-160
	NLOS	$PL_b(d,f) = 27.8 \cdot \log_{10}(d_{3D}) + 28.62 + 25.4 \cdot \log_{10}(f)$	σ=7.58	0.625-159	3-94
ITU-R P.1411-12 Site-general model for propagation within street canyons, [9]	LOS	$PL_b(d,f) = 21.2 \cdot \log_{10}(d) + 29.2 + 21.1 \cdot \log_{10}(f)$	σ=5.06	0.8-82	5-660
	NLOS	$PL_b(d,f) = 40 \cdot \log_{10}(d) + 10.2 + 23.6 \cdot \log_{10}(f)$	σ=7.6	0.8-82	30-715
ITU-R P.1411-12 Site-specific model for propagation within street canyons, [9]	LOS	$PL_{LOS,m} = L_{bp} + 6 + \begin{cases} 20 \cdot log_{10}\left(\frac{d}{R_{bp}}\right) & for\ d \le R_{bp} \\ 40 \cdot log_{10}\left(\frac{d}{R_{bp}}\right) & for\ d > R_{bp} \end{cases}$ $L_{bp} = 10 \cdot \log_{10}\left\lvert\frac{\lambda^2}{8\pi h_1 h_2}\right\rvert \quad R_{bp} \approx \frac{4h_1h_2}{\lambda}$		0.3-3	
	NLOS	$PL_{NLOS2} = 10 \cdot \log_{10}(10^{-\frac{L_r}{10}} + 10^{-\frac{L_d}{10}})$ $L_r = 20\log_{10}(x_1 + x_2) + x_1 \cdot x_2 \frac{f(\alpha)}{W_1 W_2} + 20\log_{10}\left(\frac{4\pi}{\lambda}\right)$ $f(\alpha) = \frac{3.86}{\alpha^{3.5}} \quad 0.6 < \alpha(\text{rad}) < \pi$ $L_d = 10\log_{10}(x_1 x_2 (x_1 + x_2) + 2D_a - 0.1\left(90 - \alpha\frac{180}{\pi}\right) + 20\log_{10}\left(\frac{4\pi}{\lambda}\right)$ $D_a = \frac{40}{2\pi}\left[\arctan\left(\frac{x_1}{W_1}\right) + \arctan\left(\frac{x_2}{W_2}\right) - \frac{\pi}{2}\right]$ $PL_{NLOS2} = L_{LOS} + L_c + L_{att}$ $L_c = \begin{cases} \frac{L_{corner}}{\log_{10}(1 + d_{corner})}\log_{10}\left(x_2 - \frac{W_1}{2}\right) & \frac{W_1}{2} + 1 < x_2 \le \frac{W_1}{2} + 1 + d_c \\ L_{corner} & x_2 > \frac{W_1}{2} + 1 + d_{corner} \end{cases}$ $L_{att} = \begin{cases} 60\log_{10}\left(\frac{x_1 + x_2}{x_1 + \frac{W_1}{2} + d_{corner}}\right) & x_2 > \frac{W_1}{2} + 1 + d_{corner} \\ 0 & x_2 \le \frac{W_1}{2} + 1 + d_{corner} \end{cases}$		0.8-2 2-28	
ETSI TR 138 901 V16.1.0 Site-general model for indoor propagation in factory with dense clutter and low BS (InF-DL), [10]	LOS	$PL_b(d,f) = 21.5 \cdot \log_{10}(d) + 31.84 + 19 \cdot \log_{10}(f)$	σ=4	0.5-100	1-600
	NLOS	$PL_b(d,f) = 35.7 \cdot \log_{10}((d_{3D})) + 18.6 + 20 \cdot \log_{10}(f)$	σ=7.2	0.5-100	1-600
Tunnel model, [11]	LOS	$PL_{Tunnel}(d) = \begin{cases} L_{FS}(d) & for \quad d < d_{BP} \\ L_{FS}(d_{BP}) + \alpha \cdot (d_{BP} - d) & for \quad d \ge d_{BP} \end{cases}$ $\alpha = 5.13 \cdot \lambda^2 \cdot \left(\frac{1}{W^3\sqrt{\varepsilon_v - 1}} + \frac{\varepsilon_h}{H^3\sqrt{\varepsilon_h - 1}}\right), \quad d_{BP} = \frac{4(H - h_A)^2}{\lambda}$			
ITU-R P.526-14 Knife-edge diffraction model, [12]	NLOS	$PL_d = 6.9 + 20\log_{10}(\sqrt{(\nu - 0.1)^2 + 1} + \nu - 0.1 \quad \text{for } \nu > -0.78$ $\nu = h\sqrt{\frac{2}{\lambda}\frac{d_2 + d_1}{d_1 d_2}}$			

Although these models account for the specific type of environment (e.g., indoor, urban, or suburban), they do not explicitly consider the geometric properties of that environment (see Table 3).

Conversely, site-specific models explicitly incorporate geometric parameters of the propagation environment (such as street width, tunnel height, and tunnel width) into the path loss computation.

Fig. 4 presents the measurement results alongside the path loss values predicted by the site-general models ITU-R P.1238-13, ITU-R P.1411-12 (site-general), and ETSI TR 138 901 V16.1.0.

Among these, the ITU-R P.1238-13 model demonstrated the closest agreement with the measured data, while the other models tended to overestimate path losses. The ITU-R P.1238-13 model provided an accu-

rate estimate of losses in LOS region, yielding an average Root Mean Squared Error (RMSE) of 2.8 dB (Table 5). In NLOS region, however, the error increased significantly to 12.37 dB (Table 6). This is to be expected, as the model is designed for indoor propagation in buildings, where in NLOS areas, significant signal strength comes from wave components passing through walls. In Cellar A, for NLOS measurement points in room P2, radio waves need to penetrate two 74 cm thick walls and propagate partially through the ground.

The results obtained from the site-specific models—namely ITU-R P.1411-12 (site-specific), the tunnel model, and the ITU-R P.526-14 (knife-edge diffraction) model—are illustrated in Fig. 5.

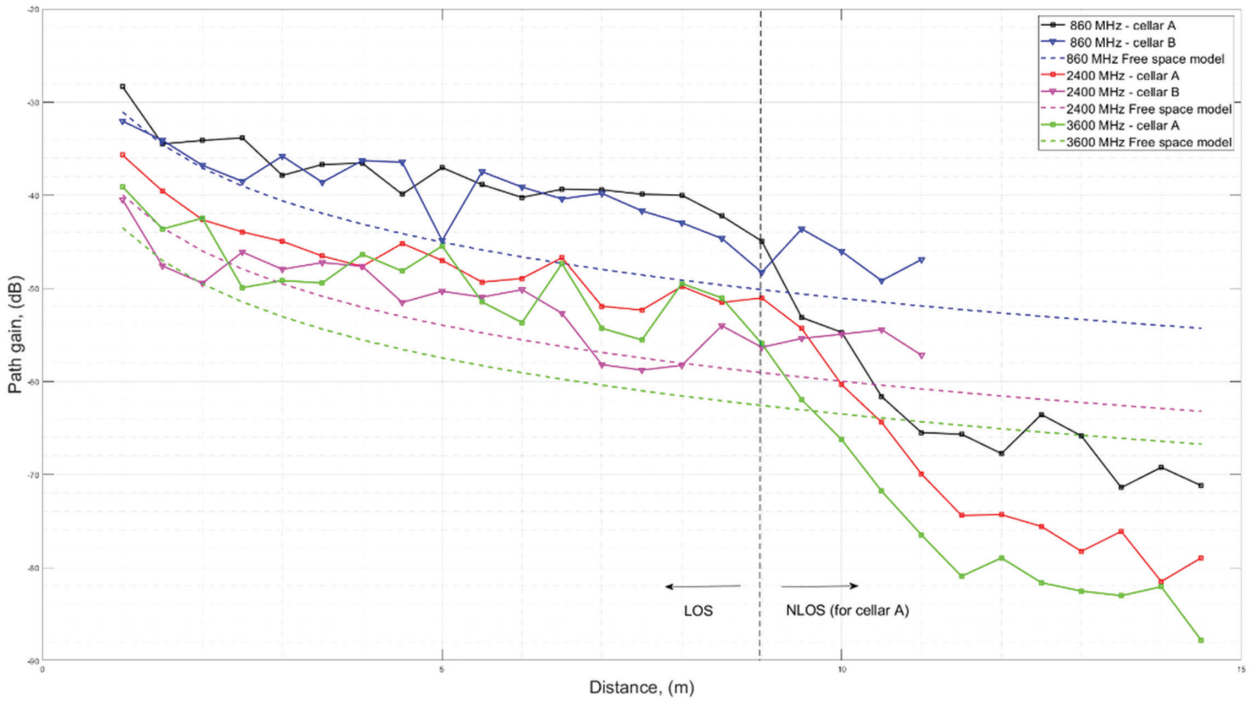


Fig. 3. Results of measurements in Cellar A and Cellar B presented as Path gain in dB

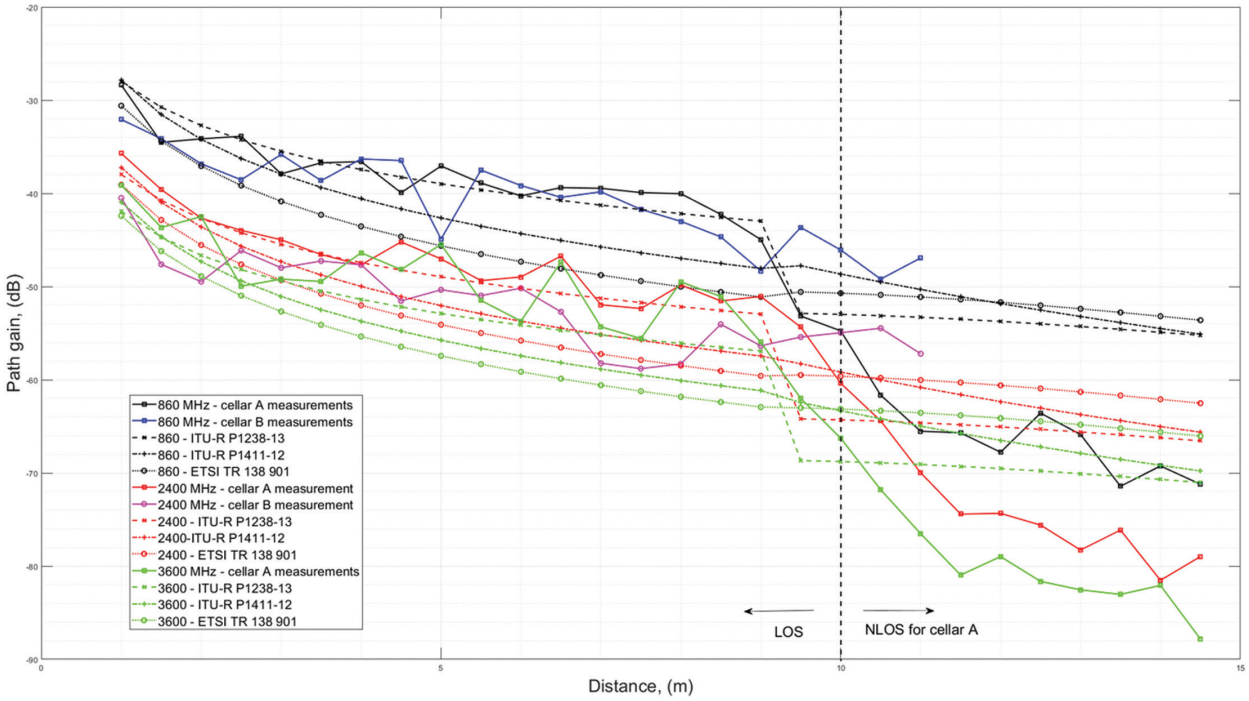


Fig. 4. Comparison of measured results and site-general models' ITU-R P.1238-13, ITU-R P.1411-12 (site-general), and ETSI TR 138 901 V16.1.0. results

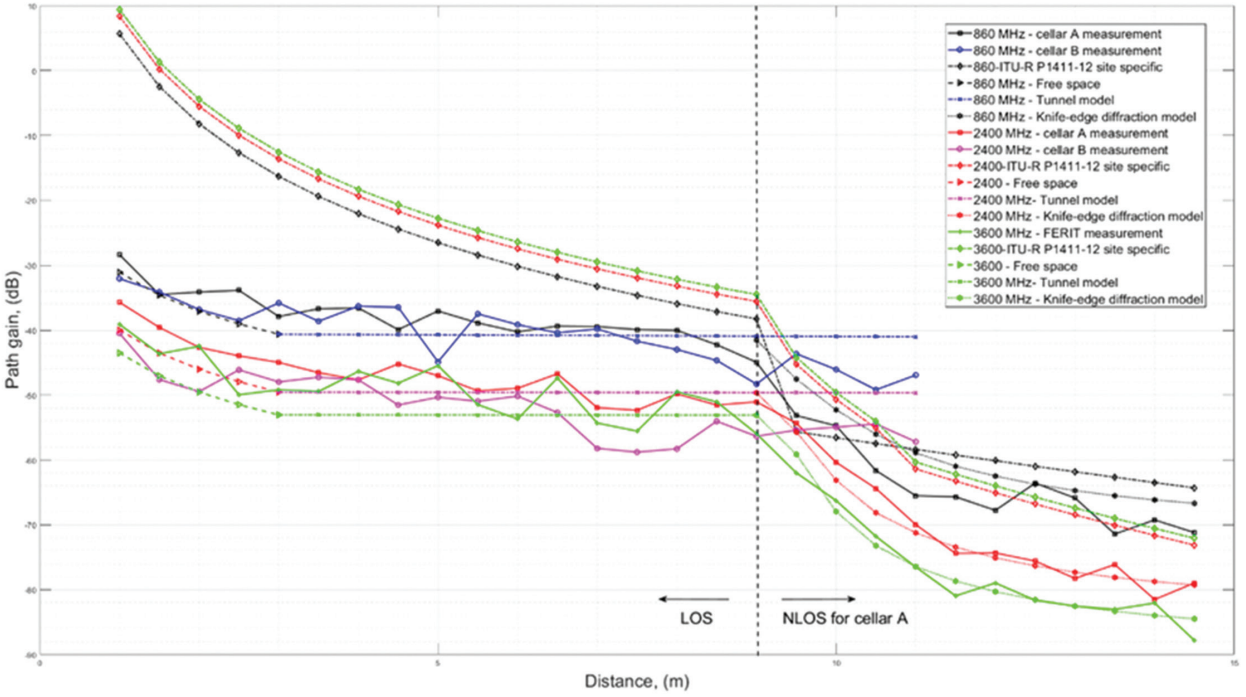


Fig. 5. Comparison of measured results and site-specific ITU-R P.1411-12 (site-specific) model, the tunnel model, and the ITU-R P.526-14 (knife-edge diffraction) model results

Table 4. Standard deviations of the individual measurements from their respective mean values

Measurements standard deviation				
860 MHz		2400 MHz		3600 MHz
Cellar A	Cellar B	Cellar A	Cellar B	Cellar A
2.79 dB	2.14 dB	3.06 dB	2.99 dB	2.09 dB

Table 5. RMSE for different models against measurement results for LOS propagation

LOS conditions	RMSE [dB]					
	860 MHz		2400 MHz		3600 MHz	Average
	Cellar A	Cellar B	Cellar A	Cellar B	Cellar A	
Free space model (FS)	6.20	5.35	6.07	3.52	7.79	5.79
ITU-R P.1238-13	1.66	2.99	1.69	3.74	3.90	2.80
ITU-R P.1411-12 Site-general	4.26	3.77	4.29	3.04	6.26	4.32
ETSI TR 138 901 V16	6.83	5.96	6.17	3.82	7.74	6.10
ITU-R P.1411-12 Site-specific	17.26	16.94	27.40	29.50	30.87	24.39
FS+Tunnel	2.83	3.18	2.95	4.61	4.25	3.56
Fitted model	1.42	2.44	1.30	2.29	2.65	2.02

The tunnel model was applied exclusively to the LOS region, while the knife-edge diffraction model was applied to the NLOS region, consistent with their respective domains of validity.

The tunnel model predicts propagation behavior consistent with the free-space model up to a distance shorter than the Fresnel breakpoint distance, d_{BP} (Table 3). For Cellar A, d_{BP} = 13.9 m, and for Cellar B, d_{BP} = 32.4 m at 860 MHz; at higher frequencies (2.4 GHz and 3.6 GHz), the breakpoint distances are even greater.

However, the measurement results indicate that the tunnel effect, characterized by a slow increase in path loss with distance, occurs much earlier, at approximately 2 – 2.5 m. Consequently, the tunnel model results in Fig. 5 were adjusted by setting d_{BP} = 2.5 m to better align with empirical observations. To calculate the α parameter for this model, we used $\varepsilon_h = \varepsilon_v = 3.75$, which corresponds to the values for bricks according to [35].

Table 6. RMSE for different models against measurement results for NLOS propagation

NLOS conditions	RMSE [dB]			
	860 MHz	2400 MHz	3600 MHz	Average
	Cellar A	Cellar A	Cellar A	
ITU-R P.1238-13	11.84	9.90	10.47	10.74
ITU-R P.1411-12 Site-general	13.42	11.22	12.47	12.37
ETSI TR 138 901 V16	13.62	13.12	14.89	13.88
ITU-R P.1411-12 Site-specific	5.71	8.99	28.99	14.56
Knife-edge diffraction	4.56	1.98	1.78	2.77
Fitted model	2.54	2.76	2.64	2.65

Further adjustments were made to the ITU-R P.1411-12 (site-specific) model based on the measurement data. Specifically, it was observed that the parameter d_{corner}, which denotes the transition zone between LOS and NLOS regions and where the path gain decreases by 20 dB, corresponds to the width of room P2. This width is 2.5 m, substantially smaller than the 30 m specified in the original model. The ITU-R P.1411-12 model was originally developed for urban street canyon environments, where 30 m represents an average street width. By analogy, for propagation within Cellar A, the width of room P2 serves as the corresponding environmental parameter governing the LOS-to-NLOS transition.

It should be noted that in Fig. 3-6, within the NLOS region, the distance (d) between the transmitter (Tx) and each measurement point is expressed as the sum of two components, X_1 and X_2, as illustrated in Fig. 1(b). Specifically, $X_1 = 8.5$ m represents the distance from the transmitter to the last measurement point located in room P1, while X_2 denotes the distance from that point to the corresponding measurement position in room P2. For propagation models in which the path loss is defined as a function of the three-dimensional distance d_{3D} this distance is computed as follows:

$$d_{3D} = \sqrt{X_1^2 + X_2^2 + \Delta h^2} \tag{8}$$

where Δh presents the height difference between the transmitting and receiving antennas.

For the site-specific propagation models, the tunnel model for LOS region and the knife-edge diffraction model for NLOS region yielded the most accurate results, with RMSE of 3.56 dB and 2.77 dB, respectively. As shown in Fig. 5, the tunnel model accurately predicts the average path gain variation up to a distance of 7 meters at 860 MHz for both basement environments. However, for longer distances in Cellar B, the model tends to overestimate the path gain. In Cellar A, the model also provides good agreement with measurements at 2.4 GHz, whereas in other scenarios the discrepancy between measured and predicted values increases substantially (RMSE > 4.2 dB).

The knife-edge diffraction model performs well at 2400 MHz and 3600 MHz, achieving RMSE values below 2 dB, while at 860 MHz the RMSE increases to 4.56 dB. Part of this error at 860 MHz stems from inaccuracies in the tunnel model's path gain estimate at a distance of 9 meters, which serves as the initial condition for the knife-edge diffraction model's calculation.

In contrast, the ITU-R P.1411-12 site-specific model exhibits significantly higher prediction errors. In the LOS region, the average RMSE is approximately 24 dB, and the model even predicts positive path gains within 2 meters of the transmitter—a physically implausible outcome. For the NLOS region, the RMSE reaches 14.56 dB, largely due to the over 6 dB error in the LOS path gain at 9 meters, which propagates into the NLOS region calculations since this value is used as the initial reference.

The final comparison was by fitting the coefficients α, β, and γ of the path-loss model (6) to the measurement results using least-squares regression. We compared the obtained with the corresponding coefficients in the general-site models. The results of the fitted models are presented in Fig. 6, and the corresponding coefficients are summarized in Table 7. For reference, Table 7 also includes the coefficient values defined in the site-general models: Free-Space model, ITU-R P.1238-13, ITU-R P.1411-12 (site-general), and ETSI TR 138 901 V16.1.0. models.

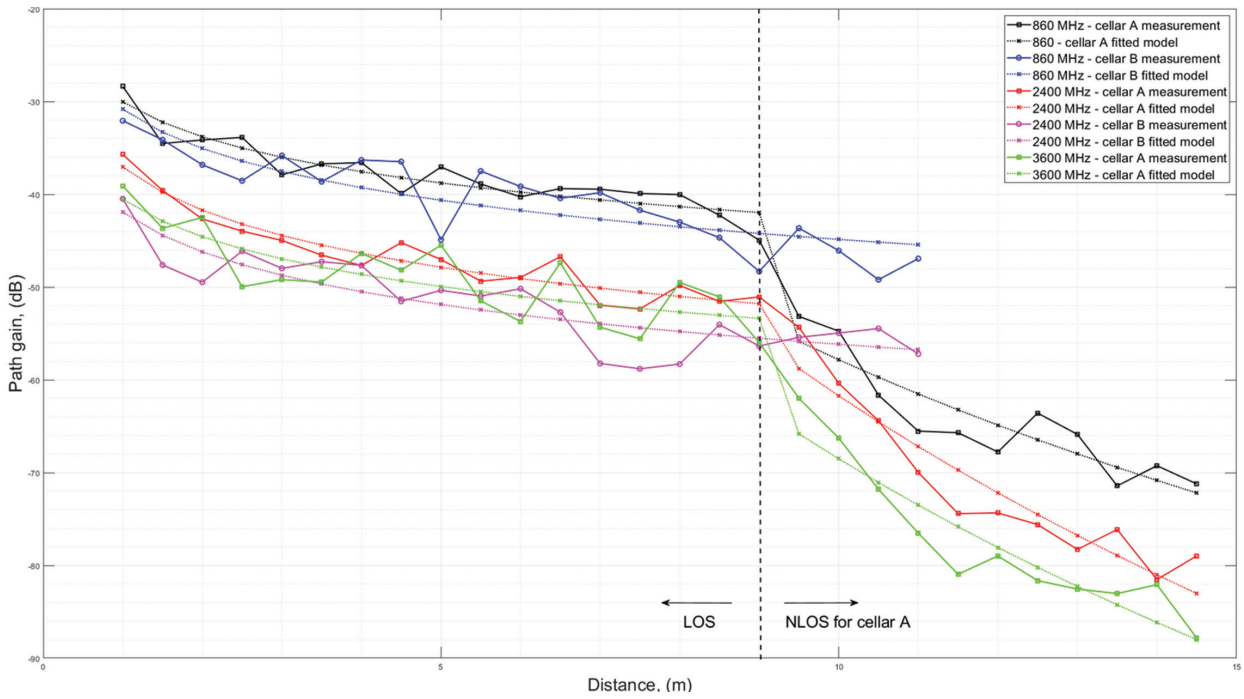


Fig. 6. Measurement results and results of corresponding fitted models

Table 7. Comparison of propagation model parameters

		Parameters for the model $PL(d, f)=10\alpha\cdot\log_{10}(d)+\beta+10\gamma\cdot\log_{10}(f)$					
		LOS			NLOS		
		α	β	γ	α	β	γ
Free space model		2	92.4	2	-	-	-
ITU-R P.1238-13		1.57	29.46	2.24	2.78	28.62	2.54
ITU-R P.1411-12 Site-general		2.12	29.2	2.11	4	10.2	2.36
ETSI TR 138 901 V16		2.15	31.84	1.9	3.57	18.6	1.9
Fitted models A	860 MHZ	1,25	31.68	1.66	8.9	-31.14	-
	2400 MHz	1.54	31.68	1.66	13.2	-70.29	-
	3600 MHz	1.35	31.68	1.66	12.01	-52.22	-
Fitted models B	860 MHZ	1.4	31.68	1.66	-	-	-
	2400 MHz	1.42	31.68	1.66	-	-	-

Under line-of-sight (LOS) conditions, the fitted coefficients were found to be in the ranges α = 1.25 - 1.54, β = 31.68, and γ = 1.66. These values are closest to those specified in the ITU-R P.1238-13 model, corroborating its strong predictive performance. For NLOS conditions, the fitted coefficients vary considerably, with α = 8.9-13.2, β = -31.14 – (-70.29), while γ could not be determined. These values differ substantially from those in the corresponding site-general models, highlighting their inadequacy for accurate prediction in NLOS environments.

The average RMSE of the fitted models is 2.02 dB for LOS and 2.65 dB for NLOS conditions (Tables 5 and 6).

5. CONCLUSIONS

This paper presents the results of path loss measurements of radio waves in underground wine cellars to assess the applicability of existing propagation models for IoT-based monitoring and control systems in smart wineries. The measurements were performed at frequencies 860 MHz, 2400 MHz, and 3600 GHz under LOS and NLOS conditions in two different cellar environments, one semi-empty and one filled with metal wine containers.

By comparing measurement results with the path loss values predicted by seven propagation models, it was found that for LOS conditions, the ITU-R P.1238-13 model and the tunnel model provide the most accurate predictions. The ITU-R P.1238-13 model achieved an average RMSE of 2.8 dB, while the tunnel model provided comparable accuracy (average RMSE = 3.56 dB), showing that these models successfully capture the dominant mechanisms of radio wave propagation in elongated basement spaces.

For NLOS conditions, the knife-edge diffraction model achieved the best correspondence with experimental data, with an average RMSE of 2.77 dB. In contrast, the site-specific ITU-R P.1411-12 model exhibited large deviations and physically inconsistent results, indicating its limited applicability to underground cellars.

Using least squares regression, coefficients for a general path-loss model fitted to the measurement data were obtained, which achieved average RMSE values of 2.02 dB (LOS) and 2.65 dB (NLOS). The obtained model coefficients closely match the parameters of the ITU-R P.1238-13 model for LOS conditions, confirming its applicability to basement environments.

It can be concluded that a hybrid modeling approach, i.e., a combination of the ITU-R P.1238-13 or tunnel model for LOS and a diffraction-based model for NLOS conditions, offers the most reliable and physically based prediction of path loss in underground wine cellars. This approach enables accurate signal estimation for low-power IoT communication systems such as LoRa, NB-IoT, and Wi-Fi, providing a practical basis for the application of smart production process monitoring technologies in the wine industry.

In future work, we plan to investigate some aspects of the presented research further. Given that in real wireless sensor networks antennas are not isotropic, as the receiving antenna used in our measurements, it can be expected that some correction/offset is required for the model. The necessity for the correction is expected both in the path loss estimation and in the field strength standard deviation.

Namely, an isotropic antenna captures reflections and diffractions from all directions, in contrast to real IoT antennas that are directional and have specific polarization. The impact of different transmitting antennas on path losses in such a complex environment could also be investigated.

Another aspect worth considering is how well wireless network planning simulation tools can estimate propagation for spaces as underground wine cellars. Modeling such spaces, considering their geometry and the electrodynamic properties of the building materials in conditions of increased humidity, presents a particular challenge.

ACKNOWLEDGMENT

This work results from implementing research activities on the project Development of an expert system for food production and processing management (ESIA-Expert System for Intelligent Agriculture) KK.01.1.1.07.0036 funded by the European Regional Development Fund.

The authors would like to thank Ivana Kovačević, Robert Miling, and Ivan Vučina for their support and assistance during the measurements.

6. REFERENCES:

[1] S. Kontogiannis, M. Tsoumani, G. Kokkonis, C. Pikridas, Y. Kotseridis, "Proposed SmartBarrel System for Monitoring and Assessment of Wine Fermentation Processes Using IoT Nose and Tongue Devices", Sensors, Vol. 25, No. 13, 2025.

[2] J. Nelson, R. Boulton, A. Knoesen, "Automated Density Measurement with Real-Time Predictive Modeling of Wine Fermentations", IEEE Transactions on Instrumentation and Measurement, Vol. 71, 2022, pp. 1-7.

[3] D. Oreški, I. Pihir, K. Cajzek, "Smart Agriculture and Digital Transformation on Case of Intelligent System for Wine Quality Prediction", Proceedings of the 44th International Convention on Information, Communication and Electronic Technology, Opatija, Croatia, 27 September - 1 October 2021, pp. 1370-1375.

[4] I. Kovačević, M. Orić, I. Hartmann Tolić, E. K. Nyarko, "Modelling the Fermentation Process in Winemaking using Temperature and Specific Gravity", Proceedings of the Central European Conference on Information And Intelligent Systems, Dubrovnik, Croatia, 20-22 September 2023, pp. 347-352.

[5] I. Kovačević, I. Aleksi, T. Keser, T. Matić, "Winnie: A Sensor-Based System for Real-Time Monitoring and Quality Tracking in Wine Fermentation", Applied Sciences, Vol. 15, No. 21, 11317, 2025.

[6] G. Masetti, F. Marazzi, L. Di Cecilia, L. Rovati, "IOT-Based Measurement System for Wine Industry", Proceedings of Workshop on Metrology for Industry 4.0 and IoT, Brescia, Italy, 16-18 April 2018, pp. 163-168.

[7] ITU-R P.525-3 Calculation of Free-space Attenuation, International Telecommunication Union – Radiocommunication Sector (ITU-R), Geneva, Switzerland, 2019.

[8] ITU-R P.1238-13 Propagation Data and Prediction Methods for the Planning of Indoor Radiocommunication Systems and Radio Local Area Networks in the Frequency Range 300 MHz to 450 GHz, International Telecommunication Union – Radiocommunication Sector (ITU-R), Geneva, Switzerland, 2023.

[9] ITU-R P.1411-12 Propagation Data and Prediction Methods for the Planning of Short-Range Outdoor Radiocommunication Systems and Radio Local Area Networks in the Frequency Range 300 MHz to 100 GHz, International Telecommunication Union – Radiocommunication Sector (ITU-R), Geneva, Switzerland, 2023.

[10] ETSI TR 138 901 V16.1.0 Study on channel model for frequencies from 0.5 to 100 GHz (3GPP TR 38.901 version 16.1.0 Release 16), European Telecommunications Standards Institute (ETSI), Sophia Antipolis, France, 2020

[11] A. Emslie, R. Lagace, P. Strong, "Theory of the Propagation of UHF Radio Waves in Coal Mine Tunnels", IEEE Transactions on Antennas and Propagation, Vol. 23, No. 2, 1975, pp. 192-205.

[12] ITU-R.526-14 Propagation by Diffraction, International Telecommunication Union - Radiocommunication Sector (ITU-R), Geneva, Switzerland, 2018.

[13] T. Jiang, J. Zhang, P. Tang, L. Tian, Y. Zheng, J. Dou, H. Asplund, L. Raschkowski, R. D'Errico, T. Jämsä, "3GPP Standardized 5G Channel Model for IIoT Scenarios: A Survey", IEEE Internet of Things Journal, Vol. 8, No. 11, 2021, pp. 8799-8815.

[14] H. A. H. Alobaidy, M. Jit Singh, M. Behjati, R. Nordin, N. F. Abdullah, "Wireless Transmissions, Propagation and Channel Modelling for IoT Technologies: Applications and Challenges", IEEE Access, Vol. 10, 2022, pp. 24095-24131.

[15] Q. P. Soo, S. Y. Lim, D. W. G. Lim, K. M. Yap, S. L. Lau, "Propagation Measurement of a Natural Cave-turned-wine-cellar", IEEE Antennas and Wireless Propagation Letters, Vol. 17, No. 5, 2018, pp. 743-746.

[16] S. Rimac-Drlje, S. Rupčić, T. Keser, I. Kovačević, R. Miling, V. Mandrić, "Propagation of Radio Waves in Wine Cellars", Proceedings of 30th International Conference on Systems, Signals and Image Processing, Ohrid, North Macedonia, 27-29 June 2023, pp. 1-5.

[17] L. Mollo, R. Greco, "Moisture Measurements in Masonry Materials by Time Domain Reflectometry", Journal of Materials in Civil Engineering, Vol. 23, No. 4, 2011, pp. 441-444.

[18] L. Subrt, P. Pechac, "Semi-Deterministic Propagation Model for Subterranean Galleries and Tunnels", IEEE Transactions on Antennas and Propagation, Vol. 58, No. 11, 2010, pp. 3701-3706.

[19] M. A. Samad, S. W. Choi, C. S. Kim, K. Choi, "Wave Propagation Modeling Techniques in Tunnel Environments: A Survey", IEEE Access, 2023, Vol. 11, 2023, pp. 2199-2225.

[20] A. Hrovat, G. Kandus, T. Javornik, "A Survey of Radio Propagation Modeling for Tunnels", IEEE Communications Surveys and Tutorials, Vol. 16, No. 2, 2014, pp. 658-669.

[21] M. A. Samad, F. D. Diba, Y. J. Kim, D. Y. Choi, "Results of Large-Scale Propagation Models in Campus Corridor at 3.7 and 28 GHz", Sensors, Vol. 21, No. 22, 2021.

[22] Z. Changsen, M. Yan, "Effects of Cross Section of Mine Tunnel on the Propagation Characteristics of UHF Radio Wave", Proceedings of the 7th International Symposium on Antennas, Propagation & EM Theory, Guilin, China, 26-29 October 2006, pp. 1-5.

[23] J. A. Azevedo, F. Mendonça, "A Critical Review of the Propagation Models Employed in LoRa Systems", Sensors, Vol. 24, No. 12, 2024.

[24] M. Rak, P. Pechac, "UHF Propagation In Caves And Subterranean Galleries", IEEE Transactions on Antennas and Propagation, Vol. 55, No. 4, 2007, pp. 1134-1138.

[25] M. Celaya-Echarri, L. Azpilicueta, P. Lopez-Iturri, I. Picallo, E. Aguirre, J. J. Astrain, J. Villandagos, F. Falcone, "Radio Wave Propagation and WSN Deployment in Complex Utility Tunnel Environments", Sensors, Vol. 20, No. 23, 2020.

[26] S. Li, Y. Liu, L. Lin, W. Ji, Z. Zhu, "Measurement, Simulation and Modeling in the Tunnel Channel with Human Bodies at 6 GHz for 5G Wireless Communication System", Proceedings of the 12th International Symposium on Antennas, Propagation and EM Theory, Hangzhou, China, 3-6 December 2018, pp. 1-4.

[27] S. Y. Tan, M. Y. Tan, H. S. Tan, "Multipath Delay Measurements and Modeling for Interfloor Wireless Communications", IEEE Transactions on Vehicular Technology, Vol. 49, No. 4, 2000, pp. 1334-1341.

[28] K. M. Malarski, J. Thrane, M. G. Bech, K. Macheta, H. L. Christiansen, M. N. Petersen, and S. Ruepp "Investigation of Deep Indoor NB-IoT Propagation Attenuation", Proceedings of the IEEE 90th Vehicular Technology Conference, Honolulu, HI, USA, 22-25 September 2019, pp. 1-5.

[29] K. M. Malarski, J. Thrane, H. L. Christiansen, S. Ruepp, "Understanding Sub-GHz Signal Behavior in Deep-Indoor Scenarios", IEEE Internet of Things Journal, Vol. 8, No. 8, 2021, pp. 6746-6756.

[30] M. B. Ali, W. Endemann, R. Kays, "Propagation Loss Model for Neighborhood Area Networks in Smart Grids", Journal of Communications and Networks, Vol. 24, No. 3, 2022, pp. 313-323.

[31] P. Branch, "Measurements and Models of 915 MHz LoRa Radio Propagation in an Underground Gold Mine", Sensors, Vol. 22, No. 22, 2022.

[32] D. D'Ayala, Y. D. Aktas, "Moisture Dynamics in the Masonry Fabric of Historic Buildings Subjected to Wind-Driven Rain and Flooding", Building and Environment, Vol. 104, 2016, pp. 208-220.

[33] R. Agliata, T. A. Bogaard, R. Greco, L. Mollo, E. C. Slob, S. C. Steele-Dunne, "Non-invasive estimation of moisture content in tuff bricks by GPR", Construction and Building Materials, Vol. 160, 2018, pp. 698-706.

[34] Z. Živković, A. Šarolić, "Measurements of Antenna Parameters in GTEM Cell", Journal of Communications Software and Systems, Vol. 6, No. 4, 2010, pp. 125-132.

[35] ITU-R, "Recommendation ITU-R P.2040-1 Effects of Building Materials and Structures on Radiowave Propagation Above About 100 MHz, P Series Radiowave propagation", 2015.

INTERNATIONAL JOURNAL OF ELECTRICAL AND COMPUTER ENGINEERING SYSTEMS

Published by Faculty of Electrical Engineering, Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia.

About this Journal

The International Journal of Electrical and Computer Engineering Systems publishes original research in the form of full papers, case studies, reviews and surveys. It covers theory and application of electrical and computer engineering, synergy of computer systems and computational methods with electrical and electronic systems, as well as interdisciplinary research.

Topics of interest include, but are not limited to:

- Power systems
- Renewable electricity production
- Power electronics
- Electrical drives
- Industrial electronics
- Communication systems
- Advanced modulation techniques
- RFID devices and systems
- Signal and data processing
- Image processing
- Multimedia systems
- Microelectronics
- Instrumentation and measurement
- Control systems
- Robotics
- Modeling and simulation
- Modern computer architectures
- Computer networks
- Embedded systems
- High-performance computing
- Parallel and distributed computer systems
- Human-computer systems
- Intelligent systems
- Multi-agent and holonic systems
- Real-time systems
- Software engineering
- Internet and web applications and systems
- Applications of computer systems in engineering and related disciplines
- Mathematical models of engineering systems
- Engineering management
- Engineering education

Paper Submission

Authors are invited to submit original, unpublished research papers that are not being considered by another journal or any other publisher. Manuscripts must be submitted in doc, docx, rtf or pdf format, and limited to 30 one-column double-spaced pages. All figures and tables must be cited and placed in the body of the paper. Provide contact information of all authors and designate the corresponding author who should submit the manuscript to https://ijeces.ferit.hr. The corresponding author is responsible for ensuring that the article's publication has been approved by all coauthors and by the institutions of the authors if required. All enquiries concerning the publication of accepted papers should be sent to ijeces@ferit.hr.

The following information should be included in the submission:

- paper title;
- full name of each author;
- full institutional mailing addresses;
- e-mail addresses of each author;
- abstract (should be self-contained and not exceed 150 words). Introduction should have no subheadings;
- manuscript should contain one to five alphabetically ordered keywords;
- all abbreviations used in the manuscript should be explained by first appearance;
- all acknowledgments should be included at the end of the paper:
- authors are responsible for ensuring that the information in each reference is complete and accurate. All references must be numbered consecutively and citations of references in text should be identified using numbers in square brackets. All references should be cited within the text;
- each figure should be integrated in the text and cited in a consecutive order. Upon acceptance of the paper, each figure should be of high quality in one of the following formats: EPS, WMF, BMP and TIFF;
- corrected proofs must be returned to the publisher within 7 days of receipt.

Peer Review

All manuscripts are subject to peer review and must meet academic standards. Submissions will be first considered by an editor-in-chief and if not rejected right away, then they will be reviewed by anonymous reviewers. The submitting author will be asked to provide the names of 5 proposed reviewers including their e-mail addresses. The proposed reviewers should be in the research field of the manuscript. They should not be affiliated to the same institution of the manuscript author(s) and should not have had any collaboration with any of the authors during the last 3 years.

Author Benefits

The corresponding author will be provided with a .pdf file of the article or alternatively one hardcopy of the journal free of charge.

Units of Measurement

Units of measurement should be presented simply and concisely using System International (SI) units.

Bibliographic Information

Commenced in 2010.
ISSN: 1847-6996
e-ISSN: 1847-7003

Published: semiannually

Copyright

Authors of the International Journal of Electrical and Computer Engineering Systems must transfer copyright to the publisher in written form.

Subscription Information

The annual subscription rate is 50€ for individuals, 25€ for students and 150€ for libraries.

Postal Address

Faculty of Electrical Engineering,
Computer Science and Information Technology Osijek,
Josip Juraj Strossmayer University of Osijek, Croatia
Kneza Trpimira 2b
31000 Osijek, Croatia

IJECES Copyright Transfer Form

(Please, read this carefully)

This form is intended for all accepted material submitted to the IJECES journal and must accompany any such material before publication.

TITLE OF ARTICLE (hereinafter referred to as "the Work"):

COMPLETE LIST OF AUTHORS:

The undersigned hereby assigns to the IJECES all rights under copyright that may exist in and to the above Work, and any revised or expanded works submitted to the IJECES by the undersigned based on the Work. The undersigned hereby warrants that the Work is original and that he/she is the author of the complete Work and all incorporated parts of the Work. Otherwise he/she warrants that necessary permissions have been obtained for those parts of works originating from other authors or publishers.

Authors retain all proprietary rights in any process or procedure described in the Work. Authors may reproduce or authorize others to reproduce the Work or derivative works for the author's personal use or for company use, provided that the source and the IJECES copyright notice are indicated, the copies are not used in any way that implies IJECES endorsement of a product or service of any author, and the copies themselves are not offered for sale. In the case of a Work performed under a special government contract or grant, the IJECES recognizes that the government has royalty-free permission to reproduce all or portions of the Work, and to authorize others to do so, for official government purposes only, if the contract/grant so requires. For all uses not covered previously, authors must ask for permission from the IJECES to reproduce or authorize the reproduction of the Work or material extracted from the Work. Although authors are permitted to re-use all or portions of the Work in other works, this excludes granting third-party requests for reprinting, republishing, or other types of re-use. The IJECES must handle all such third-party requests. The IJECES distributes its publication by various means and media. It also abstracts and may translate its publications, and articles contained therein, for inclusion in various collections, databases and other publications. The IJECES publisher requires that the consent of the first-named author be sought as a condition to granting reprint or republication rights to others or for permitting use of a Work for promotion or marketing purposes. If you are employed and prepared the Work on a subject within the scope of your employment, the copyright in the Work belongs to your employer as a work-for-hire. In that case, the IJECES publisher assumes that when you sign this Form, you are authorized to do so by your employer and that your employer has consented to the transfer of copyright, to the representation and warranty of publication rights, and to all other terms and conditions of this Form. If such authorization and consent has not been given to you, an authorized representative of your employer should sign this Form as the Author.

Authors of IJECES journal articles and other material must ensure that their Work meets originality, authorship, author responsibilities and author misconduct requirements. It is the responsibility of the authors, not the IJECES publisher, to determine whether disclosure of their material requires the prior consent of other parties and, if so, to obtain it.

- The undersigned represents that he/she has the authority to make and execute this assignment.
- For jointly authored Works, all joint authors should sign, or one of the authors should sign as authorized agent for the others.
- The undersigned agrees to indemnify and hold harmless the IJECES publisher from any damage or expense that may arise in the event of a breach of any of the warranties set forth above.

_____ _____

Author/Authorized Agent **Date**

CONTACT
International Journal of Electrical and Computer Engineering Systems (IJECES)
Faculty of Electrical Engineering, Computer Science and Information Technology Osijek
Josip Juraj Strossmayer University of Osijek
Kneza Trpimira 2b
31000 Osijek, Croatia
Phone: +38531224600,
Fax: +38531224605,
e-mail: ijeces@ferit.hr