Proceedings of

2025 8th International Conference on Circuits, Systems and Simulation

(ICCSS 2025)

May 16-18, 2025

Ho Chi Minh City, Vietnam

ISBN: 979-8-3315-9431-2

Sponsored by

IEEE

INDUSTRIAL UNIVERSITY OF HOCHIMINH CITY

Co-sponsored by

CAS VIETNAM CHAPTER

IEEE VIETNAM SECTION

Patron

國立中山大學 National Sun Yat-sen University

2025 8th International Conference on Circuits, Systems and Simulation (ICCSS 2025)

Compliant PDF Files

IEEE Catalog Number: CFP25IAI-ART
ISBN: 979-8-3315-9432-9

Conference USB Version

IEEE Catalog Number: CFP25IAI-USB
ISBN: 979-8-3315-9431-2

2025 8th International Conference on Circuits, Systems and Simulation

(ICCSS 2025)

Table of Contents

Mathematical Model of Blood Pressure Regulation using Delay Differential Equations

Anna Yu. Pyrkova*
Computer Sciences Department
Al-Farabi Kazakh National University
Almaty, Kazakhstan
ORCID: 0000-0001-8483-451X
*Corresponding author
mimoza_30_a@mail.ru

Aida M. Seitaliyeva
Fundamental Medicine Department
Al-Farabi Kazakh National University
Almaty, Kazakhstan
ORCID: 0000-0003-0177-5599

Zhanerke E. Temirbekova
Cybersecurity and Cryptology Department
Al-Farabi Kazakh National University
Almaty, Kazakhstan
ORCID: 0000-0003-3909-0210

Gulzinat K. Ordabayeva
Cybersecurity and Cryptology Department
Al-Farabi Kazakh National University
Almaty, Kazakhstan
ORCID: 0000-0001-9952-1620

Yekaterina A. Zuyeva
Automation and Information Technologies Department
Gumar Daukeyev Almaty University of Power Engineering and Telecommunications
Almaty, Kazakhstan
ORCID: 0000-0003-0762-6260

***Abstract* —This paper presents a mathematical model of blood pressure regulation described by a system of delay differential equations. The model incorporates two key physiological mechanisms — the baroreceptor and renal regulatory loops — each represented as control channels with distinct response times. The state variable is the mean systolic pressure, while the control actions are modelled through peripheral vascular resistance and renal fluid excretion rate. To ensure system stability, a finite-time stabilization (FTS) method is developed based on linear approximation and regression-based parameter identification using experimental data. The resulting system is solved numerically using the 4th-order Runge–Kutta method. The model's reliability is supported by statistical data provided by a cardiology institute, demonstrating its applicability for analyzing blood pressure dynamics under both physiological and pathological conditions.**

Keywords—Blood pressure regulation, Mathematical modeling, Differential equations with delay, Cardiovascular system, Baroreceptor mechanism.

I. INTRODUCTION

Arterial blood pressure (BP) is one of the key physiological parameters ensuring adequate perfusion of organs and tissues. Its maintenance within homeostatic ranges is achieved through a complex regulatory system involving neural, humoral, and mechanical mechanisms. Disruptions in this system can lead to the development of arterial hypertension, which, according to the World Health Organization, affects a significant portion of the adult population and is a leading risk factor for cardiovascular diseases.

Mathematical modelling of physiological processes, including BP regulation, serves as a powerful tool for analyzing and predicting the behaviour of complex biological systems. Modern models strive to account for the multilevel organization and interaction of various components within the regulatory system. For instance, Kutumova E. et al. [1] developed a modular agent-based model of the human cardiovascular and renal systems, meticulously calibrated using extensive experimental data, enabling the simulation of both physiological and pathological states.

Ishbulatov Y.M. et al. [2] proposed a model of autonomic cardiovascular regulation that incorporates baroreflex mechanisms and was validated using data from passive head-up tilt tests in healthy subjects. This model facilitated the analysis of adaptive process dynamics, including norepinephrine concentration, sympathetic and parasympathetic tone, peripheral vascular resistance, and BP.

Similarly, Celant M. and colleagues [3] developed a closed-loop mathematical model simulating hypertensive conditions, capturing changes in the cardiovascular system characteristic of arterial hypertension. The model encompasses large systemic arteries, the heart, microcirculation, pulmonary circulation, and the venous system.

Kutumova E. et al. [4] focused on modelling the effects of antihypertensive therapy, including various drug mechanisms of action. Their model allows for the assessment of therapy efficacy based on in silico predictions, paving the way for personalized medicine approaches.

Ewing G.W. [5] investigated the influence of sensory-visual stimuli on the neuroregulation of BP, aligning with our description of the central role of the medulla oblongata in the baroreceptor loop.

Michel Kana's work [6] emphasizes the importance of a comprehensive approach to modelling the interaction between neural, humoral, and vascular components, which aligns with

our consideration of regulators with varying response times and delays.

Study of Cavalcanti S. et al. [7] highlights the significance of incorporating fluid balance and vascular capacity, which corresponds with our model's inclusion of renal mechanisms and reservoir volumes.

Despite significant advancements in mathematical modelling of BP regulation, unresolved issues remain concerning the consideration of time delays in signal transmission between different system components and the integration of various regulatory mechanisms into a unified model. Notably, regulators do not directly control BP but influence cardiovascular parameters such as peripheral resistance, cardiac output, vascular stiffness, and blood volume, collectively affecting BP levels.

This study proposes a mathematical model of BP regulation described by a system of delay differential equations [8]. The model accounts for two key physiological mechanisms—the baroreceptor and renal regulatory loops—each represented as control channels with distinct response times. The system's state variable is the mean systolic pressure, while control actions involve changes in peripheral resistance and renal fluid excretion rate. To ensure model stability, a finite-time stabilization (FTS) method is developed based on linear approximation and regression-based parameter identification using experimental data [9].

The proposed model enables a more accurate depiction of BP dynamics by considering time delays and the interaction of various regulatory mechanisms, making it a valuable tool for analyzing both physiological and pathological states of the cardiovascular system.

II. Problem Statement

As the criterion for arterial blood pressure, we have chosen the mean systolic pressure, as it is the most susceptible to changes.

The cardiovascular system is presented as follows [10]. There are reservoirs A and V with arterial and venous blood connected to each other. The reservoir walls are elastic, and therefore the volumes of arterial and venous reservoirs may vary within some limits. The flow of fluid Q from the arterial reservoir to the venous is controlled by a throttle simulating peripheral resistance R. Passing through this throttle, arterial blood becomes venous. The heart plays the role of the pump (its capacity Q_h) driving the system. In addition, a small circulation circle is included, so that it is believed that after passing through the pump, venous blood becomes arterial. The venous reservoir receives liquid from other systems of the body (Q_{in} - the flow of this liquid). The fluid is removed from the blood reservoir by the kidneys (the flow of the removed fluid is Q_{out}).

Many variables can be distinguished in the blood system. They all ultimately depend on one another. Therefore, as independent variables (sufficient coordinates) it is advisable to choose those through which all others are expressed most simply. Such variables are blood and venous reservoir volumes $V_a(t)$ and $V_v(t)$:

$$V_a(t) = V_{aw}(t) - V_{a0}, \tag{1}$$

where $V_{aw}(t)$ – the volume of all arterial blood; V_{a0} - the volume of arterial reservoir in the absence of deformation (i.e. when blood pressure is zero).

Similarly

$$V_v(t) = V_{vw}(t) - V_{v0}, \tag{2}$$

where $V_{vw}(t)$ – the volume of all venous blood; V_{v0} - the volume of the venous reservoir in the absence of deformation (i.e. when the venous pressure is zero).

The pressure-to-stressed-volume dependence approaches reasonably well by the following linear dependence:

$$\begin{cases} P_a(t) = k_a V_a(t) \\ P_v(t) = k_v V_v(t) \end{cases} \tag{3}$$

where $P_a(t)$ – the mean blood pressure, $P_v(t)$ - the mean venous pressure, k_a and k_v - the stiffness of the arterial and venous rummets, respectively. If you ignore the accelerations of the liquid, then

$$P_a(t) - P_v(t) = R(t)Q(t), \tag{4}$$

where $Q(t)$ - the flow of fluid from the arterial reservoir to the venous, $R(t)$ - the resistance of the throttle, the channel passing which arterial blood becomes venous.

As is known, cardiac output is determined by the level of venous pressure, the cardiac contractility (inotropic state), and the afterload (resistance) in the arterial vessels. It should be noted, however, that the latter factor plays a significant role only at high blood pressure (Anrepa phenomenon), so it can be neglected in this model.

Under Frank-Starling Law [11]

$$Q_h(t) = k_1 P_v(t), \tag{5}$$

where k_1 – the constant that characterizes the cardiac contractility (inotropic state).

The continuity equations are of the form [12]:

$$\begin{cases} \dot{V}_a(t) = Q_h(t) - Q_{out}(t) - Q(t) \\ \dot{V}_v(t) = Q_{in}(t) + Q(t) - Q_h(t) \end{cases}, \tag{6}$$

where $Q_{in}(t)$ – the incoming flow of liquid into the venous reservoir, $Q_{out}(t)$ - the flow of removed liquid from the arteries into the kidneys, $Q_h(t)$ - a heart-pump derivative that drives the system.

It should be noted that the incoming fluid flow can be considered practically constant, since all fluctuations related to external fluid loss or intake are buffered by the extracellular fluid reserve, which amounts to about 20 litres, whereas the total volume of the circulatory system is only about 5 litres. Let 's mark $Q_{in}(t) = Q_{in} = \text{const}$.

Consider how blood pressure is regulated in a given system. Various mathematical models are known in literature describing to some extent the physiological mechanisms of blood pressure regulation. Generally, eight of the most significant blood pressure control mechanisms are generally isolated. Some of them quickly practice perturbatory effects and within a few seconds offset the resulting pressure surges. Others are included with different delay times, and the transition period for them is from a few minutes to several hours and even days. In addition, the different control loops have different boundary areas of blood pressure beyond which the pressure change does not cause changes in the regulator.

It can be assumed that the mean systolic arterial pressure may vary from 100 to 150 mmHg. Consequently, control circuits such as the chemoreceptor system, the ischemic response of the central nervous system, and the renin-angiotensin system can be neglected. The "stress relaxation" mechanism and the capillary water filtration mechanism are implicitly accounted for in the stiffness coefficients of arterial and venous reservoirs.

Thus, on this model it is sufficient to take into account the operation of only two control loops.

A. Baroreceptor mechanism

Baroreceptors are located in external layer of vascular wall of aorta and carotid sinus. The reason for the impulse activity occurring in the baroreceptor is the deformation of the vessel wall caused by blood pressure. The signal generated in nerve endings of receptors is transmitted to the central nerve regulator of oblong brain, at the output of which a signal is formed, which transmits to peripheral vessels and changes their hydraulic resistance.

From the point of view of the control object, the baroreceptor mechanism measures the deformation of the walls of arterial vessels associated with the stressed volume of the arterial reservoir, and according to the measurement results controls the peripheral resistance according to an inverse-proportional S-shaped dependence $R(V_a)$. In the model's assumption of blood pressure change boundaries, this relationship can be approximated to linear.

Note also that some time passes between the moment of change of the stressed volume (arterial pressure) and the corresponding relaxation of the peripheral vessels. According to experimental data, it is 5-7 seconds. Therefore, you can write:

$$R(t) = k_2 - k_3 V_a(t - \tau), \tag{7}$$

where τ - delay of baroreceptor mechanism; k_2, k_3 - constants determined by regression analysis methods.

B. Renal mechanism of water-salt balance preservation

When arterial pressure drops below the normal level, kidney release of water and salt decreases. This mechanism entails a progressive increase in fluid in the extracellular space of the body, which leads to an increase in the intracular volume of blood, an increase in the stress-strain state of the vessels and a return of pressure to some initial level.

As the pressure increases, the release of water and salt from the kidneys increases, resulting in a decrease in the volume of extracellular fluid, a decrease in the stressed volume of the vascular system, and a decrease in pressure to a normal level.

From the point of view of the control object, the renal mechanism affects the flow of outgoing liquid $Q_{out}(t)$. This dependency can be roughly described by a linear function:

$$Q_{out}(t) = k_4 P_a(t) - k_5, \tag{8}$$

where k_4, k_5 - the constants determined by regression analysis methods.

This control loop has no delay, however, this does not contradict the fact that the renal mechanism is considered the slowest arterial pressure control loop. The fact is that the rate of fluid removal by the kidneys is quite small - only about 1 ml/min at normal pressure.

Thus, based on (3)-(5), (7), (8), the arterial pressure control model (6) can be represented as a system of differential equations with the delaying argument [9]:

$$\begin{cases} \dot{P_a}(t) = k_1 k_a P_v(t) - k_4 k_a P_a(t) + k_5 k_a - {k_a}^2 \frac{P_a(t) - P_v(t)}{k_2 k_a - k_3 P_a(t-\tau)} \\ \dot{P_v}(t) = k_v Q_{in} + k_v k_a \frac{P_a(t) - P_v(t)}{k_2 k_a - k_3 P_a(t-\tau)} - k_1 k_v P_v(t) \end{cases} \tag{9}$$

$$t_0 < t < T, \text{ ?}$$

$$\begin{cases} P_a(t) = P_{a0}, t \in [t_0 - \quad , t_0] \\ P_v(t_0) = P_{v0} \end{cases}, \tag{10}$$

where k_i, $i = \overline{1,5}$ - coefficients determined by regression methods, P_{a0} and P_{v0} are values of arterial and venous pressures at initial moment of time.

In terms of physical meaning, the values $R(t)$ and $Q_{out}(t)$ must be positive. Therefore, in linear approximation of constraints $R(V_a(t))$ and $Q_{out}(P_a(t))$, constraints should be imposed on the quantity $P_a(t)$:

$$\sigma_1 = \frac{k_5}{k_4} \le P_a(t) < k_a k_2 k_3 = \sigma_2. \tag{11}$$

This condition limits the scope of the built model.

III. Methodology

A non-linear approximation of these dependencies is required to extend this region.

Stationary solution of this system:

$$\begin{cases} P_a{}^* = \frac{Q_{in} + k_5}{k_4} = x^* \\ P_v{}^* = [k_1 \left(k_2 - \frac{k_3}{k_a} x^*\right) + 1]^{-1} \left[Q_{in} \left(k_2 - \frac{k_3}{k_a} x^*\right) + x^*\right] = y^* \end{cases} \tag{12}$$

We will replace the variables as follows:

$$\begin{cases} P_a(t) = P_a^* + x_1(t) \\ P_v(t) = P_v^* + x_2(t) \end{cases}, \quad (13)$$

Having replaced variables and having connected control, we have:

$$\begin{cases} \dot{x_1}(t) = -c_1c_0x_1(t) + c_1c_2x_2(t) \\ +\big(-c_1x_1(t) + c_1x_2(t) - c_3x_1(t-\tau)\big)\alpha\big(x_1(t-\tau)\big) + u_1 \\ \dot{x_2}(t) = -c_4c_2x_2(t) \\ +\big(c_4x_1(t) - c_4x_2(t) - c_5x_1(t-\tau)\big)\alpha\big(x_1(t-\tau)\big) + u_2 \end{cases} \quad (14)$$

$$t_0 < t < T,$$

with initial conditions:

$$\begin{cases} x_1(t) = P_{a0} - P_a^*, t \in [t_0 - \tau, t_0] \\ x_2(t_0) = P_{v0} - P_v^* \end{cases} \quad (15)$$

and stabilization condition:

$$x_i(T) = 0, i = \overline{1,2} \quad (16)$$

where

$$c_0 = k_4,\ c_1 = k_a,\ c_2 = k_1, c_3 = k_3(k_1P_v^* - k_4P_a^* + k_5), c_4 = k_v, \quad (17)$$

$$c_5 = \frac{k_3k_v}{k_a}(Q_{in} - k_1P_v^*), \alpha\big(x_1(t-\tau)\big) = \frac{k_a}{k_2k_a - k_3(P_a^* + x_1(t-\tau))}$$

According to the theorem 2.4.1 [8] about stabilization of the system movement on the finite-time interval (FTI), it is possible to determine the stabilizing control of the arterial pressure system (9), (10).

When searching for the method control proposed in the theorem 2.4.1, it is most convenient for the matrix of the linear part of the system to lead to a diagonal view. Therefore, we will use the replacement:

$$\begin{cases} x_1(t) = y_1(t) + \frac{c_1c_2}{c_1c_0 - c_4c_2}y_2(t) \\ x_2(t) = y_2(t) \end{cases} \quad (18)$$

After replacement and simplest algebraic transformations, the system (14) will take the form:

$$\begin{cases} \dot{y_1}(t) = -c_1c_0y_1(t) + \bar{f_1}\big(t, y_1(t), y_2(t), y_1(t-\tau)\big) \\ \quad +w_1(t, y_1(t), y_2(t), y_1(t-\tau)) \\ \dot{y_2}(t) = -c_4c_2y_2(t) + \bar{f_2}\big(t, y_1(t), y_2(t), y_1(t-\tau)\big) \\ \quad +w_2(t, y_1(t), y_2(t), y_1(t-\tau)) \end{cases} \quad (19)$$

where $w_1\big(t, y_1(t), y_2(t), y_1(t-\tau)\big) = u_1\big(t, x(t), x_1(t-\tau)\big) + \frac{c_1c_2}{c_2^2 - c_1c_0}u_2\big(t, x(t), x_1(t-\tau)\big)$,

$w_2\big(t, y_1(t), y_2(t), y_1(t-\tau)\big) = u_2\big(t, x(t), x_1(t-\tau)\big)$,

$\bar{f_1}(t, y_1(t), y_2(t), y_1(t-\tau))$ and $\bar{f_2}(t, y_1(t), y_2(t), y_1(t-\tau))$ - non-linear parts of the system.

$$\dot{\Phi}(t) = \bar{A}(t)\Phi(t),\ \Phi(t_0) = E_n, \quad (20)$$

where $\bar{A}(t)$ - the matrix of the linear part of the system (19).

$$\Phi(t) = \begin{pmatrix} e^{-c_1c_0t} & 0 \\ 0 & e^{-c_4c_2t} \end{pmatrix}, \quad (21)$$

$$Q(t) = \Phi^{-1}(t)\bar{B}(t) \quad (22)$$

where $\bar{B}(t)$ - the system control coefficient matrix (19).

$$Q(t) = \begin{pmatrix} e^{c_1c_0t} & 0 \\ 0 & e^{c_4c_2t} \end{pmatrix} \quad (23)$$

$$R(t,T) = \begin{pmatrix} \frac{e^{2c_1c_0T} - e^{2c_1c_0t}}{2c_1c_0} & 0 \\ 0 & \frac{e^{2c_4c_2T} - e^{2c_4c_2t}}{2c_4c_2} \end{pmatrix}, \quad (24)$$

$$W(t,T) = \begin{pmatrix} \frac{e^{2c_1c_0(T-t)} - 1}{2c_1c_0} & 0 \\ 0 & \frac{e^{2c_4c_2(T-t)} - 1}{2c_4c_2} \end{pmatrix}, \quad (25)$$

$$K(t) = W^{-1}(t,T) = \begin{pmatrix} \frac{2c_1c_0}{e^{2c_1c_0(T-t)} - 1} & 0 \\ 0 & \frac{2c_4c_2}{e^{2c_4c_2(T-t)} - 1} \end{pmatrix} \quad (26)$$

$$w^0\big(t, y(t)\big) = -B^*(t)K(t)y(t) = \begin{pmatrix} \frac{2c_1c_0y_1(t)}{1 - e^{2c_1c_0(T-t)}} \\ \frac{2c_4c_2y_2(t)}{1 - e^{2c_4c_2(T-t)}} \end{pmatrix} \quad (27)$$

When returning to variables $x(t) = \begin{pmatrix} x_1(t) \\ x_2(t) \end{pmatrix}$, replace (18), we have:

$$\begin{cases} u_1^0\big(t, x(t)\big) = \frac{2c_1c_0x_1(t)}{1 - e^{2c_1c_0(T-t)}} - \frac{2c_1c_2x_2(t)}{c_1c_0 - c_4c_2} \\ \quad * \left(\frac{c_1c_0}{1 - e^{2c_1c_0(T-t)}} - \frac{c_2c_4}{1 - e^{2c_4c_2(T-t)}}\right) \\ u_2^0\big(t, x(t)\big) = \frac{2c_4c_2x_2(t)}{1 - e^{2c_4c_2(T-t)}} \end{cases} \quad (28)$$

and $v(t, x(t), x_1(t-\tau))$ can be defined from identity $x^*(t)K(t)B(t)v\big(t, x(t), x_1(t-\tau)\big) = -x^*(t)K(t)f\big(t, x(t), x_1(t-\tau)\big) - {x_1}^2(t)$.

$$\begin{cases} v_1\big(t,x(t),x_1(t-\tau)\big) = \frac{e^{2c_1c_0(T-t)}-1}{2c_1c_0x_1(t)} - \\ \big(-c_1x_1(t)+c_1x_2(t)-c_3x_1(t-\tau)\big)\alpha(x_1(t-\tau)) \\ v_2\big(t,x(t),x_1(t-\tau)\big) = \frac{1-e^{2c_4c_2(T-t)}}{2c_4c_2x_2(t)}(x_1^2(t)+1) \\ -\big(c_4x_1(t)-c_4x_2(t)-c_5x_1(t-\tau)\big)\alpha(x_1(t-\tau)) \end{cases} \quad (29)$$

Stabilizing control of the control system according to theorem 2.4.1 is as follows: $u\big(t,x(t),x_1(t-\tau)\big) = u^0\big(t,x(t)\big) + v(t,x(t),x_1(t-\tau))$ has the following view:

$$\begin{cases} u_1\big(t,x(t),x_1(t-\tau)\big) = \frac{2c_1c_0x_1(t)}{1-e^{2c_1c_0(T-t)}} - \frac{2c_1c_2x_2(t)}{c_1c_0-c_4c_2} \\ \left(\frac{c_1c_0}{1-e^{2c_1c_0(T-t)}} - \frac{c_2c_4}{1-e^{2c_4c_2(T-t)}}\right) + \frac{e^{2c_1c_0(T-t)}-1}{2c_1c_0x_1(t)} \\ -\big(-c_1x_1(t)+c_1x_2(t)-c_3x_1(t-\tau)\big)\alpha(x_1(t-\tau)) \\ u_2\big(t,x(t),x_1(t-\tau)\big) = \frac{2c_4c_2x_2(t)}{1-e^{2c_4c_2(T-t)}} + \\ + \frac{1-e^{2c_4c_2(T-t)}}{2c_4c_2x_2(t)}(x_1^2(t)+1) \\ -\big(c_4x_1(t)-c_4x_2(t)-c_5x_1(t-\tau)\big)\alpha(x_1(t-\tau)) \end{cases} \quad (30)$$

The system (14), (15), (16) was solved, together with the obtained management, by the numerical Runge-Kutta method of the 4th order, the parameters of the system were determined by means of regression analysis methods. Statistical data was provided by Kazakh Scientific Research Institute of Cardiology (Almaty) under a research collaboration agreement.

IV. Discussion

The proposed mathematical model for arterial pressure regulation, based on a system of delay differential equations, offers a more precise depiction of blood pressure dynamics by incorporating time delays and the interplay of various regulatory mechanisms. Unlike prior models, such as the modular agent-based framework by Kutumova E. et al. [1] and the autonomic regulation model by Ishbulatov Y.M. et al. [2], our model integrates both baroreceptor and renal control loops with distinct temporal responses, providing a more realistic representation of physiological processes.

The finite-time stabilization method (FTSM) developed within this study ensures the model's stability and facilitates its adaptation to experimental data. This is particularly crucial for analyzing pathological conditions like arterial hypertension, where signal transmission delays play a significant role.

Comparisons with other studies, such as the closed-loop model by Celant M. et al. [3], which simulates hypertensive states, indicate that incorporating time delays and multiple control loops enhances the model's accuracy and predictive capabilities. Furthermore, our model can be employed to assess the efficacy of antihypertensive therapies, akin to the approach proposed by Kutumova E. et al. [4], thereby opening avenues for personalized medicine.

However, certain limitations of our model should be acknowledged. Firstly, it is based on simplified assumptions of linearity and does not account for potential nonlinear interactions among different components of the regulatory system. Secondly, the model necessitates further validation using a broader range of experimental data, including data from patients with various forms of hypertension.

Future work aims to expand the model by incorporating additional regulatory mechanisms, such as neurohumoral and endothelial factors, and to conduct clinical studies for its validation and customization to individual patient characteristics.

Acknowledgment

The authors would like to thank the Cardiology Institute of Almaty for providing access to patient data, which was essential for this study.

References

[1] Kutumova E, Kiselev I, Sharipov R, Lifshits G, Kolpakov F. Thoroughly calibrated modular agent-based model of the human cardiovascular and renal systems. Front Physiol. 2021;12:746300. https://doi.org/10.3389/fphys.2021.746300

[2] Ishbulatov YM, Karavaev AS, Kiselev AR, Simonyan MA, Prokhorov MD, Ponomarenko VI, et al. Mathematical modeling of the cardiovascular autonomic control in healthy subjects during passive head-up tilt test. Sci Rep. 2020;10(1):71532. https://doi.org/10.1038/s41598-020-71532-7

[3] Celant M, Toro EF, Bertaglia G, Cozzio S, Caleffi V, Valiani A, et al. Modeling essential hypertension with a closed-loop mathematical model. Math Med Biol. 2022;39(2):3748. https://doi.org/10.1002/cnm.3748

[4] Kutumova E, Kiselev I, Sharipov R, Lifshits G, Kolpakov F. Mathematical modeling of antihypertensive therapy. Front Physiol. 2022;13:1070115. https://doi.org/10.3389/fphys.2022.1070115

[5] Ewing GW. Mathematical modeling the neuroregulation of blood pressure using a cognitive top-down approach. N Am J Med Sci. 2010;2(5):2341. https://doi.org/10.4297/najms.2010.2341

[6] Kana M. Mathematical models of cardiovascular control by the autonomic nervous system [Preprint]. arXiv. 2019. Available from: https://doi.org/10.48550/arXiv.1901.05071

[7] Cavalcanti S, Cavani S, Ciandrini A, Avanzolini G. Mathematical modeling of arterial pressure response to hemodialysis-induced hypovolemia. Comput Biol Med. 2006;36(2):128–44. https://doi.org/10.1016/j.compbiomed.2004.08.004

[8] Pyrkova A.Yu. Solving the problems of medicine and biology using high-performance computing. Almaty: Kazakh University; 2019. 102 p.

[9] Pyrkova A.Yu., Ivachshenko A.T. Bioinformation in biotechnology. Almaty: Qazaq University; 2020. 149 p.

[10] Arkhipova O.Y., Godin E.A., Kolmanovsky V.B., Shtengold E.Sh. Regulation of arterial pressure and hypertension. Autom Remote Control. 1990;(8):129–38.

[11] Magder S. Central Venous Pressure: A Useful but Not So Simple Measurement. Critical Care Medicine. 2006;34(8):2224-2227.

[12] Guyton A.C., Hall J.E. Textbook of Medical Physiology. Elsevier; 14th edition (2020).

A Calibration-free 12-bit 1-GS/s Pipelined ADC with High-Linearity Input Buffer

Haotian Zhang
State Key Laboratory of Millimeter Waves
Southeast University
Nanjing, China
e-mail: 230228248@seu.edu.cn

Peng Miao*
State Key Laboratory of Millimeter Waves
Southeast University
Nanjing, China
e-mail:miaopeng123@seu.edu.cn

Fei Li
State Key Laboratory of Millimeter Waves
Southeast University
Nanjing, China
e-mail: lifei@seu.edu.cn

Yahan Yu
State Key Laboratory of Millimeter Waves
Southeast University
Nanjing, China
e-mail: 230208539@seu.edu.cn

Di Wang
State Key Laboratory of Millimeter Waves
Southeast University
Nanjing, China
e-mail: 230209036@seu.edu.cn

Ankang Ding
State Key Laboratory of Millimeter Waves
Southeast University
Nanjing, China
e-mail: 230218611@seu.edu.cn

***Abstract*—This paper presents a 12-bit 1-GS/s pipelined analog-to-digital converter (ADC) implemented in 40nm CMOS technology, featuring a high-linearity input buffer and high-gain residue amplifier design. The proposed architecture employs three key innovations: (1) A current-feedback input buffer with constant drain-source voltage (V_{DS}) control technique that achieves 60-80% input capacitance reduction while maintaining high linearity through negative feedback loops. (2) A folded quarter-range 2.5-bit MDAC architecture that reduces amplifier bandwidth requirements by 50% through output swing range optimization. (3) A two-stage push-pull operational amplifier with gain-boosting that achieves 80 dB gain and 12 GHz unity-gain bandwidth. Fabricated prototype measurements demonstrate 10.2-bit ENOB and 76.8 dB SFDR at 496 MHz input frequency. The core consumes 466 mW power and occupies 750×250 μm² core area. The front-end sample-and-hold circuit is eliminated through SHA-less architecture, while the implementation of a high-gain wide-bandwidth operational amplifier obviates the need for complex calibration circuitry.**

Keywords—ADC, pipelined ADC, input buffer, residue amplifier, high-linearity

I. INTRODUCTION

As a critical component in digital signal processing systems, Analog-to-Digital Converters (ADCs) play an essential role across multiple application domains. With technological advancements, the demand for high-speed and high-precision data conversion continues to escalate. The pipelined ADC architecture, a prevalent configuration for achieving both speed and resolution, is widely employed in high-performance systems including high-speed data acquisition, communication systems, and radar applications [1-4]. The increasingly stringent performance requirements imposed on ADCs underscore the significance of researching and refining pipelined ADC designs to address these technical challenges.

Supported by State Key Laboratory of Millimeter Waves Open Project (Z202503-07)

Positioned preceding the sampling network, the integrated input buffer proves critical for maintaining sampling-phase linearity through its dual functionality as a low-output-impedance driver. As demonstrated in prior studies, this configuration establishes signal isolation while simultaneously maintaining high input impedance-essential characteristics that optimize the sampling network's dynamic performance [5, 6]. The residue amplifier is the core of the MDAC circuit and also a key module in the design of the pipelined ADC. The performance of the residue amplifier determines the quality of the output signal of the MDAC, which in turn affects the output of each subsequent pipelined stage, playing a decisive role in the conversion performance of the final ADC.

In this paper, a 12-bit 1GS/s pipelined ADC with high-linearity input buffer is proposed. The current-feedback input buffer with constant-V_{DS} control technique can achieve high bandwidth and low distortion. To avoid using complex calibration circuits, the design employs a two-stage push-pull amplifier with gain-boosting as the residue amplifier. This approach effectively meets the critical need for both high settling speed and precision in the system.

This paper is organized as follows. Section II discusses the proposed pipelined ADC architecture. Section III describes the key block circuit implementation. The simulation results and post-layout are shown in Section IV, followed by conclusion in Section V.

II. ADC ARCHITECTURE

Fig. 1 illustrates the proposed pipelined ADC architecture, comprising three primary components: an input buffer, a multi-stage ADC core, and a digital error correction unit. The ADC core consists of five 2.5-bit multiplying digital-to-analog converter (MDAC) stages and a 2-bit flash ADC. A SHA-less architecture is used in the first 2.5-bit MDAC. In this design, the inter-stage gain is strategically scaled down from the

979-8-3315-9431-2/25/$31.00 ©2025 IEEE

conventional factor of 4 to 2 in the initial two stages, thereby relaxing the operational amplifier's design constraints while enhancing linearity characteristics.

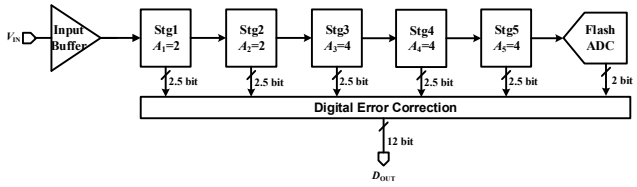


Fig. 1. The pipelined ADC architecture

A. MDAC Architecture

The conventional 2.5-bit pipeline stage shares structural similarities with the 1.5-bit architecture. As depicted in Fig. 2(a), both the input and output voltage swings span from -V_{REF} to V_{REF}. The conventional configuration employs six comparators to partition the input range into seven quantization sub-ranges, with each segment exhibiting an inter-stage gain of 4. This design achieves a comparator offset tolerance of $V_{REF}/8$.

In contrast, A folded quarter-range 2.5-bit pipeline stage is introduced in this work. As shown in Fig. 2(b), the last two sub-ranges are limited to half the size of the other sub-ranges, thereby ensuring uniform output ranges across all sub-ranges. The extra code requires an extra comparator in the stage and an extra bit in the output. In addition, the output signal swing of this pipeline stage is reduced to a quarter of the full-scale range, thereby relaxing the performance specifications of the amplifier in the MDAC. Specifically, the required unity-gain bandwidth is reduced by approximately 50% compared to conventional MDAC designs. This bandwidth reduction proves critical for minimizing the ADC's overall power consumption, given that amplifiers typically account for over 65% of the total power budget in pipelined ADC.

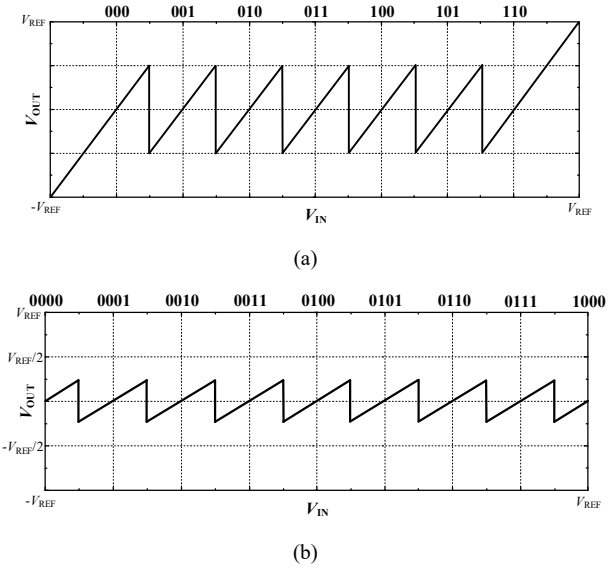


Fig. 2. A plot of a 2.5-bit stage's output residue as a function of its input. (a) conventional. (b) folded quarter-range

III. ADC Circuit Implementation

A. High-Linearity Input Buffer Design

The input buffer not only provides better isolation for the input signal but also ensures a high input impedance, thereby enhancing the performance of the sampling system. In high-resolution ADCs, employing a sufficiently large sampling capacitance is crucial for mitigating thermal noise. The input buffer drives the sampling capacitance within a very short sampling time, ensuring high linearity and minimal added noise.

In the conventional design, the input buffer is based on an NMOS source follower. The transfer function can be conclude in the equation as [7]:

$$\frac{v_{out}}{v_{in}} \cong \frac{g_m Z_L}{1+g_m Z_L} \tag{1}$$

Where V_{out} and V_{in} is the input and output of the buffer, g_m represents the transconductance of the follower device and Z_L is the load impedance. Due to the signal-dependence of g_m, the nonlinearity in the input buffer can be expressed as a variation in the output voltage. From (1), the variation of output voltage can be deduced as:

$$\frac{\delta v_{out}}{v_{out}} \cong \frac{\delta g_m / g_m}{1+g_m Z_L} \tag{2}$$

Enhancing the g_m of the SF is also critical for minimizing the distortion. Beyond this, the buffer's parameters (e.g. g_m) should remain constant irrespective of input signal variation.

The proposed input buffer architecture, shown in Fig. 3, combines current feedback [8] with stable drain-source voltage regulation. The key innovations include a single-stage current mirror configuration (M3-M4 pair) for current feedback, and an auxiliary PMOS source follower (M2) used to clamp the V_{DS} of the main follower transistor M1. To mitigate the body effect, both the main source follower (M1) and the auxiliary one (M2) connect their bulk to the source terminal.

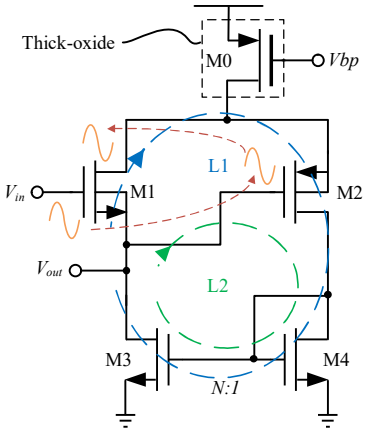


Fig. 3. Proposed current-feedback input buffer (single-ended for simplicity)

To improve the frequency response and distortion performance, the design prioritizes parasitic capacitance

minimization at critical nodes, particularly at the buffer output. Concurrently, thin-oxide MOS devices are selectively deployed in the follower network and current mirror to optimize transconductance efficiency (g_m/I_D) under low-supply-voltage constraints. Notably, a thick-oxide transistor (M0) is strategically incorporated in the biasing path to ensure robust overvoltage protection, effectively eliminating gate oxide reliability concerns.

Fluctuations in the signal current through the sampling capacitance can induce nonlinear distortion due to g_m variations in the source follower. In this work, when the input signal leads to an increase in the current of M1, the current of M2 correspondingly decrease due to the fixed total bias current of the circuit. This current reduction in M2 is mirrored through the current mirror pair (M3-M4) back to M1's biasing path, thereby ensuring that the current through M1 remains independent of input signal variations. This negative feedback mechanism significantly improves the linearity of the input buffer.

The buffer linearity degradation caused by channel-length modulation effects from V_{DS} variations is addressed through the auxiliary source follower M2. This auxiliary circuit forces the drain voltage of M1 to track its source voltage variations, which helps to maintain a constant V_{DS} for M1. Furthermore, the introduction of M2 reduces the input buffer's input capacitance by 60%-80% due to the Miller effect.

As depicted in Fig. 3, the input buffer circuit contains two feedback loops: a negative feedback loop (L1) and a positive feedback loop (L2). To ensure system stability, the negative feedback factor must be greater than the positive feedback factor. In this topology, both loop L1 and loop L2 share transistors M2 and M3. Therefore, stability is guaranteed by maintaining the transconductance of M1 (g_{m1}) greater than that of M2 (g_{m2}).

B. Residue Amplifier

Residue amplifiers serve as the core component of MDAC circuits and represent critical modules in pipelined ADC designs. The performance of the residue amplifier directly determines the output signal quality of the MDAC, which subsequently affects subsequent pipeline stage outputs, ultimately determining the complete ADC's conversion performance [9].

Therefore, when designing Residue amplifiers for MDAC applications, it is essential to comprehensively analyze nonideal effects. These include finite DC gain, limited unity-gain bandwidth, and slew rate constraints. Such considerations serve as the foundation for setting the design specifications of the amplifier's performance metrics. The amplifier design must allow for sufficient performance margins to ensure the robustness of the overall system.

Both open-loop amplifiers and ring amplifiers exhibit relatively low power consumption; however, they are notably sensitive to variations in process, voltage, and temperature (PVT) [10]. To reduce the system complexity, a high - gain closed - loop operational transconductance amplifier (OTA) is selected as the residue amplifier in this design. This approach can avoid the use of complex calibration circuits.

To mitigate the impact of nonideal effects on overall ADC performance, the total error should be restricted to within 1 least significant bit (LSB). For the MDAC, a margin should be reserved for other nonideal factors, and it is specified that the error introduced by the finite gain error should be less than half of VLSB of the subsequent stage. Therefore, the requirement for the open-loop gain A of the operational amplifier is:

$$A > \frac{2 \cdot 2^{N-M}}{\beta} \tag{3}$$

Where N represents the resolution of ADC, M represents the actual quantization bits at the current stage and β is the feedback factor of the MDAC. Similarly, the settling error caused by the finite unity-gain bandwidth should also be less than half of the VLSB of the subsequent stage. Therefore, the requirement for the unity-gain bandwidth f_u of the operational amplifier should be:

$$f_u = \frac{\omega_u}{2\pi} > \frac{\ln(2 \cdot 2^{N-M})}{2\pi \cdot \beta \cdot t_s} \tag{4}$$

Here, t_s is the small-signal settling time of the operational amplifier. The first-stage MDAC employs a folded quarter-range 2.5-bit architecture, where the feedback factor $\beta = 1/2$ and the closed-loop gain is 2. By using (3) and ensuring an appropriate design margin, the open-loop gain specification of the operational amplifier is determined. Through this analysis, it is found that the first-stage operational amplifier requires a minimum open-loop gain of over 80 dB.

Accounting for the comparator's comparison time, decision propagation delay, and large-signal settling characteristics, the operational amplifier's small-signal settling time is specified as 250 ps. Consequently, the first-stage operational amplifier requires the unity-gain bandwidth exceeding 12 GHz to satisfy this transient performance constraint.

To satisfy the high-gain, high-bandwidth requirements, a two-stage push-pull amplifier architecture with gain-boosting techniques is implemented, achieving enhanced open-loop gain characteristic as demonstrated in Fig. 4.

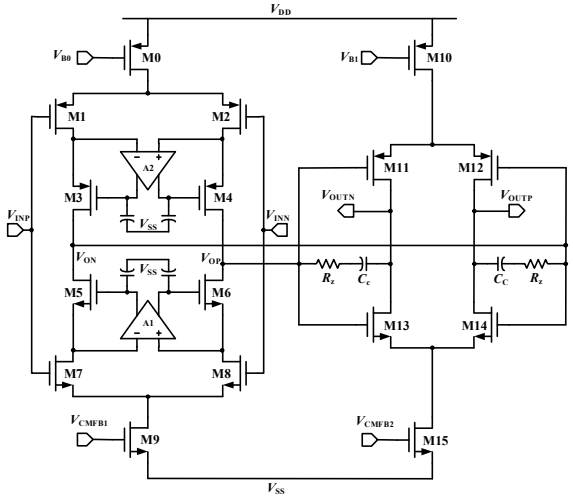


Fig. 4. Schematic of the operational amplifier in this work

The first amplifier stage implements a modified telescopic cascode architecture, where the cascode load transistors (M7 and M8) are configured as complementary input pairs to form push-pull amplification. An NMOS transistor M9 functions as the tail current source with common-mode feedback (CMFB) control beneath this structure. Analogously, the second stage utilizes a common-source amplifier with a push-pull configuration. Here, the load transistors M13 and M14 act as input pairs, and an NMOS transistor M15 is incorporated as the CMFB-regulated tail current source. Gain-boosting is accomplished by embedding auxiliary amplifiers A1 and A2 in the N-type and P-type cascode branches of the first stage, respectively. This approach effectively enhances the overall open-loop gain.

The two-stage operational amplifier architecture necessitates compensation circuitry to ensure stability. A Miller compensation capacitor (C_C) in series with a resistor (R_Z), connected between the outputs of the first and second stages, guarantees sufficient phase margin under closed-loop operation. Furthermore, the implemented gain-boosting technique introduces complex pole-zero pairs (doublelet), necessitating load capacitors at the differential outputs of both auxiliary amplifiers for dominant-pole compensation.

Given the reduced resolution requirements in pipeline stages 3-5, the specifications for gain error tolerance can be appropriately relaxed. Consequently, the gain-boosting technique is omitted in these subsequent stages, thereby enhancing the closed-loop phase margin. This design simplification eliminates auxiliary amplifiers, achieving measurable reductions in area and power consumption compared to architectures employing gain-boosting.

IV. Post Simulation Results

The proposed pipelined ADC is fabricated in 40nm CMOS process with a total chip area of 800 μm×400 μm, as shown in Fig. 5. The area of the core is 750 μm×250 μm. The core ADC totally dissipates 466 mW with a 2.5 V supply and 1.2 V_{pp} input swing.

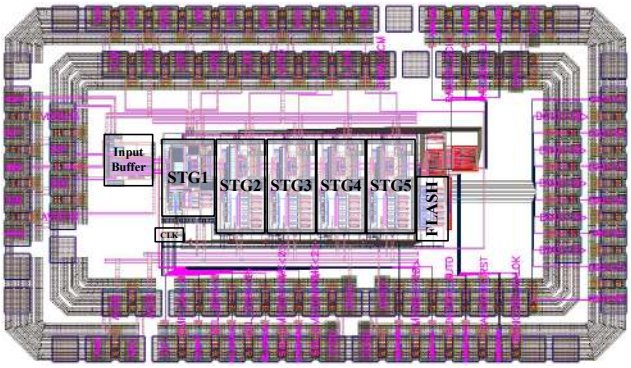


Fig. 5. The layout of the proposed ADC

Fig. 6 shows the output spectrum of the proposed pipelined ADC operating at a high input frequency of 496 MHz. The ADC achieves 10.2-bit ENOB and 76.8 dB SFDR without calibration. Fig. 7 plots ENOB, SFDR, and SNDR variations versus input frequency without calibration.

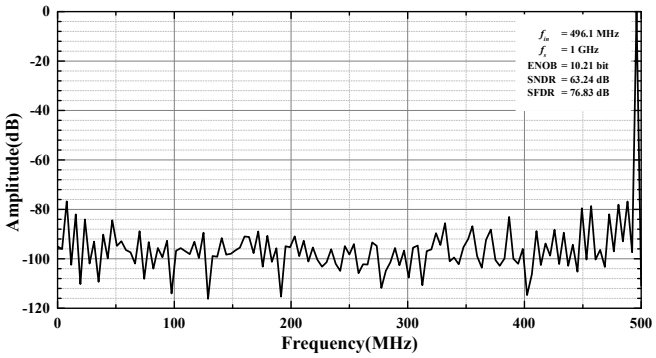


Fig. 6. The output spectrum of proposed ADC at high input frequency

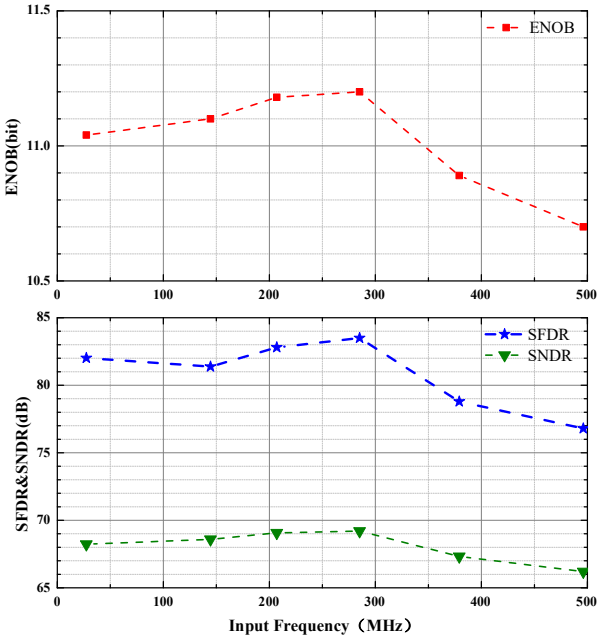


Fig. 7. Simulated ENOB, SNDR and SFDR of ADC versus input frequency

V. Conclusion

This paper presents a 12-bit 1-GS/s SHA-less pipelined ADC implemented in 40nm CMOS technology that achieves calibration-free operation through three key innovations: a current-feedback input buffer with constant-VDS control technique that reduces input capacitance by 60-80% while maintaining high linearity, a folded quarter-range 2.5-bit MDAC architecture that relaxes amplifier bandwidth requirements by 50%, and a two-stage push-pull operational amplifier with gain-boosting that delivers 80 dB gain and 12 GHz unity-gain bandwidth. Simulation results demonstrate 10.2-bit ENOB and 76.8 dB SFDR at 496 MHz input frequency with 466 mW power consumption, making this design particularly suitable for high-speed applications such as 5G communications and radar systems where both precision and speed are critical. The compact core area of 750×250 μm² further enhances its practical applicability in integrated systems. Future research directions could explore extending this architecture to higher sampling rates beyond 2-GS/s using more advanced process nodes, investigating digital background calibration techniques to further improve linearity caused by CDAC and RA,

and reducing the overall power consumption of the system to make it applicable to time - interleaved systems and low - power applications. These advancements would build upon the solid foundation established by this work while addressing emerging requirements in high-performance data conversion systems.

REFERENCES

[1] A.M.A. Ali, “High Speed Data Converters” (The Institution of Engineering and Technology, London, 2016) 165-181.

[2] A.M.A. Ali, et al., “A 14 bit 1GS/s RF sampling pipelined ADC with background calibration,” IEEE J. Solid-State Circuits, vol. 49, no. 12, pp. 2857-2867, Dec. 2014

[3] A. M. A. Ali et al., "A 14-bit 2.5GS/s and 5GS/s RF sampling ADC with background calibration and dither," 2016 IEEE Symposium on VLSI Circuits (VLSI-Circuits), Honolulu, HI, USA, 2016, pp. 1-2

[4] A. M. A. Ali et al., "A 12-b 18-GS/s RF Sampling ADC With an Integrated Wideband Track-and-Hold Amplifier and Background Calibration," in IEEE Journal of Solid-State Circuits, vol. 55, no. 12, pp. 3210-3224, Dec. 2020

[5] L. Fang, X. Wen, T. Fu and P. Gui, "A 12-Bit 1 GS/s RF Sampling Pipeline-SAR ADC With Harmonic Injecting Cross-Coupled Pair Achieving 7.5 fj/Conv-Step," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 69, no. 8, pp. 3225-3236, Aug. 2022.

[6] L. Fang et al., "A 2.56-GS/s 12-bit 8x-Interleaved ADC With 156.6-dB FoMS in 65-nm CMOS," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 30, no. 2, pp. 123-133, Feb. 2022.

[7] F. Cao, Y. Chen, Z. Dai, F. Ye and J. Ren, "An input buffer for 12bit 2GS/s ADC," 2017 IEEE 12th International Conference on ASIC (ASICON), Guiyang, China, 2017, pp. 750-753,.

[8] B. Vaz et al., "16.1 A 13b 4GS/s digitally assisted dynamic 3-stage asynchronous pipelined-SAR ADC," 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2017, pp. 276-277.

[9] M. Gu, Y. Zhong, L. Jie and N. Sun, "24.1 A 12b 3GS/s Pipelined ADC with Gated-LMS-Based Piecewise-Linear Nonlinearity Calibration," 2025 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2025, pp. 1-3, doi: 10.1109/ISSCC49661.2025.10904536

[10] W. Jiang, Y. Zhu, M. Zhang, C. -H. Chan and R. P. Martins, "A Temperature-Stabilized Single-Channel 1-GS/s 60-dB SNDR SAR-Assisted Pipelined ADC With Dynamic Gm-R-Based Amplifier," in IEEE Journal of Solid-State Circuits, vol. 55, no. 2, pp. 322-332, Feb. 2020, doi: 10.1109/JSSC.2019.2948170.

Power-effective TIA Design for PNN-based AI Edges

1st Chua-Chin Wang
Inst. of IC Design
National Sun Yat-Sen University
Kaohsoing, Taiwan
Email : ccwang@ee.nsysu.edu.tw

2nd Jyun-Wei Chen
Dept. of Electrical Eng.
National Sun Yat-Sen University
Kaohsiung, Taiwan
Email : david.cc.0412@gmail.com

3rd Yung-Jr Hung
Department of Photonics
National Sun Yat-Sen University
Kaohsiung, Taiwan
Email : yungjrhung@gmail.com

4th Zong-Ming Chang
Department of Photonics
National Sun Yat-Sen University
Kaohsiung, Taiwan
Email : tyuio9856@gmail.com

Abstract—CMOS-based or silicon-based deep neural networks (DNN) have been widely used recently in many real-time artificial intelligent (AI) hardware implementations. However, most present AI hardware platforms either suffer from serious area cost or large power dissipation. PNN (photonic neural network) has been studied to relax the high power dissipation when AI training and learning are executed. Transimpedance amplifiers (TIA), as the critical optical receiver frontend to carry out the optical-to-electrical (O-E) conversion, plays a key role in PNN-based systems. More specifically, TIA takes charge of converting the weak output current from the photodetector (PD) into voltage signals for further signal processing and computation. To faithfully realize optical computations in PNNs, TIAs shall attain large bandwidth and high sensitivity. An enhanced RGC-based TIA with a cross-coupled pair is proposed and realized in this investigation using cost-effective 180-nm CMOS process to meet the requirements of PNN systems. The on-silicon measurement results show that the bandwidth = 4.23 GHz, the gain = 64.9 dBΩ, and 3.1 mW power dissipation, which is adequate to be integrated within PNNs without any loss of performance.

Index Terms—PNN, TIA, high bandwidth, low power dissipation, CMOS process

I. Introduction and Overview

Artificial intelligence (AI) based on CNN, DNN, etc., has been booming in the past years owing to its capacity to discover unnoticed patterns and links within data sets, e.g., [1]. However, the severe power hungary issues for AI training and learning slow down the advance of related research progress. Strubell et al. has pointed out a fact that the CO2 emission of a typical learning using DNN or CNN might consume more than 300 times of that of a roung trip flight from New Work to San Francisco [2]. This fact directly tells the fact that AI applications demand large power and energy. For instance, Reck et al. presented one interesting design for chip-integrated optical neural networks [3]. So did Chang et al. proposed another hybrid optical solution [4]. The feature is that they proposed to use optical device, which are considered to attain less power consumption, to overcome the power hungry problems encountered by CMOS-based electronic platforms. However, the overhead caused by the interfacing circuits to carry out optical-electrical data conversion was ignored. For example, though the intensive convolution computation can be fully realized using MZIs (Mach-Zehnder interferometer) with appropriate biases, the summation of the optical power in all light paths with various wavelength must be converted into an electrical current using PD (photo diode) [5]. Then, a reliable TIA is required to restore the convolution outcome to be a electrical signal which will be processed with the electronic systems. A PNN prototype realized by a TSMC research team and our team is demonstrated in Fig. 1, where the positioning of TIA is highlight.

Many TIA designs have been proposed, e.g., MSTA-TIA [6] and RGC-TIA [7] to serve as solutions for various applications. However, none of the existing TIAs were designed mainly for the usage in PNNs to carry out the conversion of optical MAC (multiply accumulation) outcome into electrical signals. The reason is that due to the low amplitude of optical signals which is unfavorable for measurement. TIA circuits with high gain and sensitivity are definitely needed in PNNs. A typical Current Amplifier-based TIA with Capacitive Ladder Feedback [8] was reported, which can significantly reduce the noise effect. However, the complexity of this TIA and the possible mismatch of capacitors make it hard to be directly used in PNNs [9]. Thus, we propose a wide-bandwidth power-effective TIA in this work to meet the demand of general PNN optical signal conversion.

II. Proposed Power-Effective Wide-BW TIA for PNNs

A. Proposed TIA circuit analysis

Fig. 2 shows the schematic of the proposed TIA for PNNs is mainly composed of an RGC-TIA with a cross-coupled pair to reduce the overall input impedance and the effect of the input capacitance posing on the bandwidth

This investigation was also partially supported by National Science and Technology Council, Taiwan, under grant NSTC 110-2221-E-110-063-MY2 and 109-2221-E-032-001-MY3.

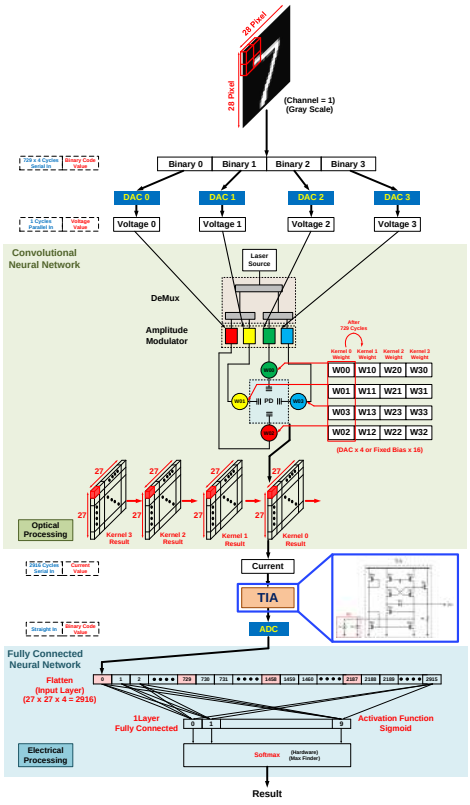


Fig. 1. TIA positioning in a typical PNN

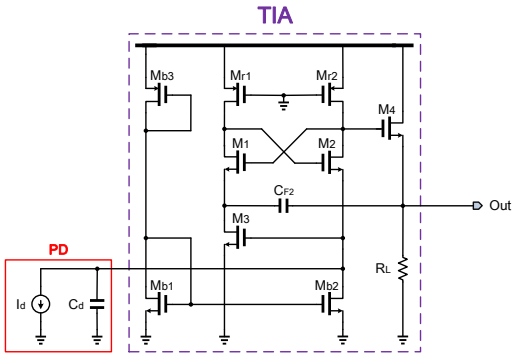


Fig. 2. Schematic of the proposed TIA

reduction. Notably, the cross-coupled pair is meant to provide a theoretical high gain to neutralize the parasitic capacitance so that the bandwidth is enhanced. Last but not least, a large capacitor is employed in the feedback loop to stabilize the output, which is deemed as an open circuit in DC analysis so that it will not affect the high frequency performance. In short, the proposed TIA consists of the following blocks (from left to right in Fig. 2) : the PD (photo detector) equivalent circuit, bias generator, cross-coupled gain stage, and the output stage to drive the external load.

To analyze the proposed TIA theoretically, the low-

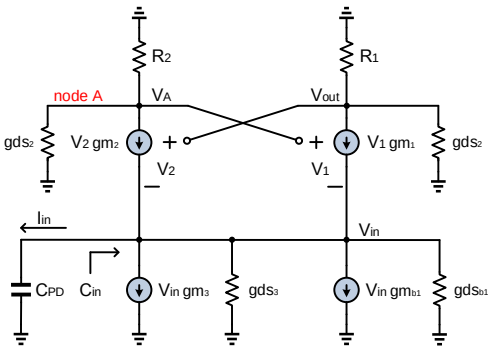


Fig. 3. Small-signal model of the proposed TIA

frequency small-signal model of the proposed TIA is given in Fig. 3. Based on the small-signal model, the DC gain can be quickly derived as follows.

$$\begin{aligned}
\frac{V_A}{V_{in}} &= \frac{-R_1 gm_1 \cdot [gm_{b1} + R_1 gm_2 (gds_{b1} \parallel gds_3)]}{gm_1 \cdot [1 - R_{out} R_1 gm_2 (gds_{b1} \parallel gds_3)] + gds_{b1}} \\
\frac{V_{out}}{V_A} &= \frac{-R_{out} gm_2 gm_1 \cdot [1 + R_1 gm_2 (gds_{b1} \parallel gds_3)]}{R_1 gm_1 \cdot [gm_{b1} + R_{out} gm_2 (gds_{b1} \parallel gds_3)]} \\
A_{V_{TIA}} &= \frac{-R_{out} gm_2 gm_1 \cdot [1 + R_1 gm_2 (gds_{b1} \parallel gds_3)]}{gm_1 \cdot [1 - R_{out} R_1 gm_2 (gds_{b1} \parallel gds_3)] + gds_{b1}} \\
&\cong \frac{R_{out} \cdot gm_2}{R_1} \cdot \frac{V_A}{V_{in}} \qquad (1)
\end{aligned}$$

The contribution of the proposed TIA is to reduce or eliminate the effect caused by the large input capacitance in prior RGC-TIAs. It is well known that the large capacitance of PD will dominate the first pole of TIAs to reduce the bandwidth thereof such that the conversion speed is deteriorated in certain circumstances, particularly in high-temperature scenarios. Thus, we add the cross-coupled pair made of PMOS devices to resolve this issue. The first pole of the proposed TIA, namely the 3-dB frequency, is then found in Eqn. (2). Notably, C_{PD} effect can be reduced by enlarging $A_{V_{TIA}}$, which is realized by the cross-couple pair.

$$\begin{aligned}
C_{in} &\cong C_{gsb1} + C_{gs3} + (1 - \frac{V_A}{V_{in}}) \cdot C_{gs2} \\
&\quad + (1 - gm_1 R_1) \cdot C_{gdb1} \\
\omega_{3dB} &\cong \frac{1}{R_{out} {A_{V_{TIA}}}^{-1} \cdot (C_{in} \parallel C_{PD})} \\
&\approx \frac{1}{R_{out} {A_{V_{TIA}}}^{-1} C_{PD}} \qquad (2)
\end{aligned}$$

B. Requirements of PD used in PNNs

In this investigation supported by TSMC, our PD is designated by our partners, including TSMC and Optoelectronics Lab. of Photonics Department in NSYSU, which is featured with the numbers specified in Table I.

TABLE I
PD characteristics in the PNN

I_{out} (μA)	1 - 500
Capacitance (pF)	> 0.5
Max. power (mW)	< 10

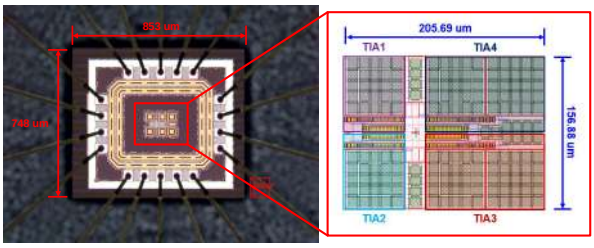


Fig. 4. Die photograph of the fabricated TIA

III. Post-layout Simulation and Measurement on Silicon

The proposed TIA prototype is fabricated on silicon using TSMC 180-nm CMOS technology, as shown in Fig. 4, where the chip area is 853×748 μm^2 and the core area is 206×167 μm^2.

A. All-PVT-corner post-layout simulation

An all-PVT-corner (process, supply voltage, temperature) post-layout simulation is required for the design approval before the prototype was fabricated. Notably, a total of at least 45 PVT corners are simulated, namely (SS, FS, TT, SF, FF)×(0.9×VDD, VDD, 1.1×VDD)×(0, 25, 75)^{o}C. Fig. 5 shows the AC simulations at all PVT corners, where the gain is 61.9 - 65.0 dBΩ, and the BW is 5.83 - 9.46 GHz. Fig. 6 demonstrates the DC simulations where the input range is 0 to 100 μA and the output voltage is 740 - 880 mV with very high linearity ($R^2 \geq 0.99$).

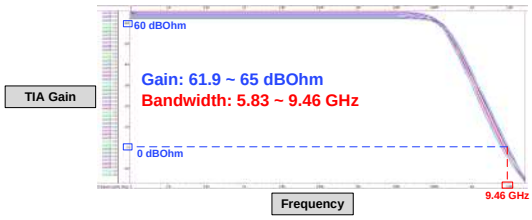


Fig. 5. Post-layout AC simulations

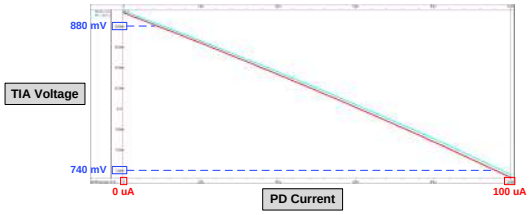


Fig. 6. Post-layout DC simulations

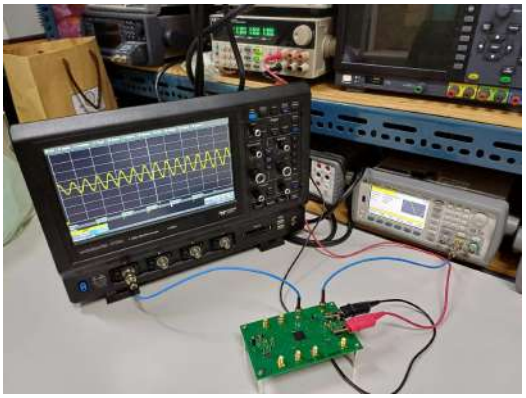


Fig. 7. Chip measurement setup for the TIA prototype

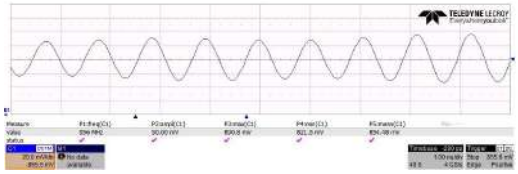


Fig. 8. Output waveform of the proposed TIA given 100 MHz input

B. TIA chip measurement

Referring to Fig. 7, the electrtical measurement setup for the prototype chip is presented. ITech IT6333A is the power supply used. The input signals are generated using Keysight 33600A function generator, and the outputs are measured using Teledyne LeCroy Wavesurfer 3104z. The measurements are done for 6 chips measured 10 times each. Fig. 8 shows the TIA output waveform given a 100 MHz input, which proves the correct functionality. The measurement outcomes are also included in Table II.

C. Optical signal conversion measurement

Besides the TIA prototype on-silicon measurement, an optical signal conversion experiment was conducted. Referring to Fig. 9, a complete measurement setup to test the performance of the proposed TIA is disclosed, where a tunable laser (Santec TSL-550), a polarization controller(Thorlabs FPC562 fiber polarization controller), a power meter SMU (Keithley-2401), DC power supply (ITech IT6333A), and an OSC (Teledyne LeCroy Wavesurfer 3104z) are used. The bird view of the DUT (device under test), a chip integrated with the proposed TIA and the grating coupler provided by TSMC, is demonstrated in Fig. 10.

Various PD currents are generated by the optoelectronic chip when different strength of laser light is given. The DC current measurement is summarized in Fig. 11, where the PD current is ranged from 50 to 300 μA to generate the TIA voltage from 1.05 to 0.40 V almost linearly. The laser power is also recorded as well simultaneously, which

TABLE II
Performance comparison with previous TIAs

	[10]	[11]	[12]	[13]	[14]	[15]	Ours
Year	2015	2016	2017	2019	2021	2023	2024
Process (nm)	130	180	180	65	11	180	180
V_{DD}	1.5	1.8	1.8	1.2	1.8	1.8	1.8
Max. Gain (dBΩ)	50.1	69.3	59	40	99	85.2	64.9
BW (GHz)	7	1	7.9	54	0.34	0.94	4.23
PD Cap. (pF)	0.25	N/A	0.3	0.05	1	0.65	1
Power Diss. (mW)	7.5	6.0	18.0	55.2	41	8.97	3.1
I/P Referred Noise ($pa/\sqrt{Hz}$)	31.3	9.33	23	19.8	3.67	N/A	14.29
Area (mm^2)	0.016	0.0075	0.11	0.6	0.023	0.988	0.638
Inductor	No	No	Yes	Yes	No	No	No
FOM1	46.8	11.55	25.9	39.1	0.8	8.9	88.57
FOM2	2.8	N/A	6.08	5.45	9.17	N/A	19.21

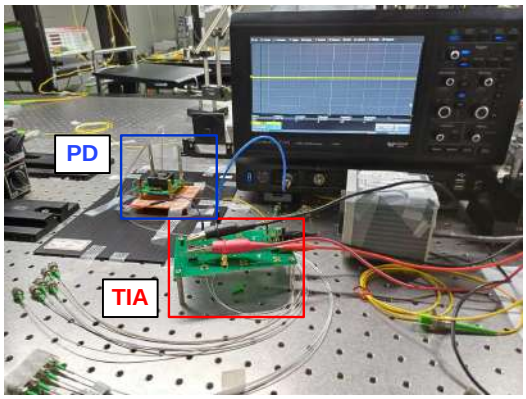


Fig. 9. Optical signal conversion experiment setup

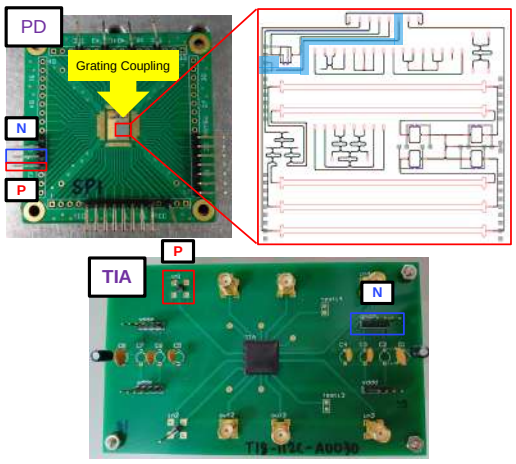


Fig. 10. Packaged TIA+grating coupler to be measured in the optical signal conversion

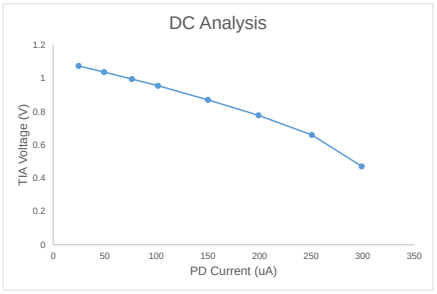


Fig. 11. TIA output voltage vs. PD current

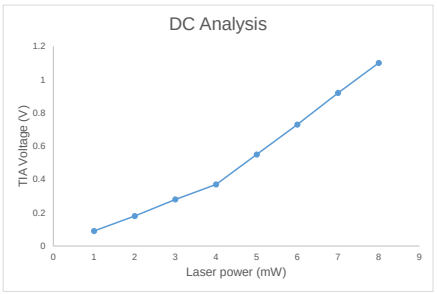


Fig. 12. TIA voltage vs. laser power

is shown in Fig. 12, where the power is ranged from 0.5 to 8 mW.

Lastly, Fig. 13, 14, and 15 demonstrate the stability of the TIA output voltage with different laser power ratings at 0.09, 0.37, and 0.92 mW, respectively. Apparently, these outcomes prove that the large capacitor in the feedback path works as expected to stabilize the output voltage. For example, the worst ripple at 0.37 mW laser power is found to be merely $\frac{960.2-951.5}{960.2+951.5}/2 < 0.5\%$.

D. Performance comparison and analysis

Table II presents a performance comparison with several prior TIA designs that were implemented and measured

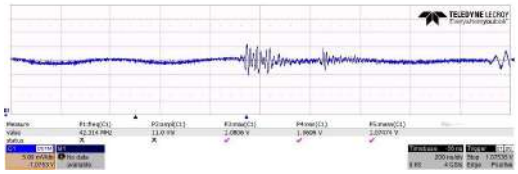


Fig. 13. TIA output voltage quality at laser power 0.09 mW

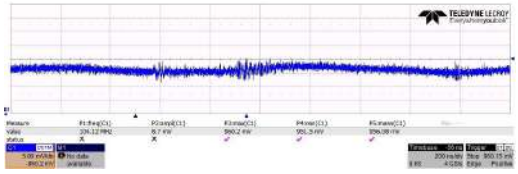


Fig. 14. TIA output voltage quality at laser power 0.37 mW

on silicon. The proposed TIA achieved the lowest power dissipation and the best FOM defined in 3.

$$\text{FOM1} = \frac{Gain(dB\Omega) \times BW(GHz)}{Power(mW)} \tag{3}$$

$$\text{FOM2} = \frac{Gain(dB\Omega) \times BW(GHz) \times \text{PD Cap.}(pF)}{\text{Input Referred Noise}(pA/\sqrt{Hz})} \tag{4}$$

The measured worst-case gain of the proposed TIA is 64.9 dBΩ, which is close to 65.0 dBΩ given by the post-layout simulations. The measured BW (4.23 GHz) is slightly lower than that given by the post-layout simulations, which is 5.83 GHz at the output load = 10 kΩ. The reason is the loading effect (50Ω, the impedance of OSC) of the measurement probes and equipment. However, it is still better than that of many prior TIAs.

IV. Conclusion

A wide-bandwidth and power effective TIA is reported in this investigation. Notably, since the TIA is meant to be integrated in PNNs so that the overall input capacitance effect is reduced so as to enhance the conversion speed required by the future optical MAC operations in future PNNs. The proposed TIA is featured with the addition of cross-coupled pair in RGC-TIA such that the impact of large input capacitance is relaxed. Moreover, a large capacitor is integrated in the feedback path to stabilize the output voltage.

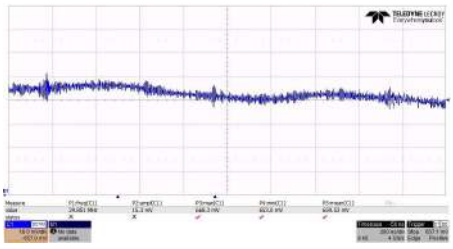


Fig. 15. TIA output voltage quality at laser power 0.92 mW

References

[1] C.-C. Wang, R. G. B. Sangalang, C.-P. Kuo, H.-C. Wu, Y. Hsu, S.-F. Hsiao, and C.-H. Yeh, "A 40.96-GOPS 196.8-mW digital logic accelerator used in DNN for underwater object recognition," IEEE Trans. on Circuits and Systems I: Regular Papers, vol. 69, no. 12, pp. 4860-4871, Dec. 2022.

[2] E. Strubell, A. Ganesh, and A. McCallum. "Energy and policy considerations for deep learning in NLP." arXiv preprint arXiv:1906.02243, 2019.

[3] M. Reck, A. Zeilinger, H. J. Bernstein, and P. Bertani, "Experimental realization of any discrete unitary operator," Physical Rev. Lett., vol. 73, no. 1, pp. 58 - 61, Jul. 1994.

[4] J. Chang, V. Sitzmann, X. Dun, W. Heidrich, and G. Wetzstein, "Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification," Scientific Rep., vol. 8, no. 1, pp. 1 - 10, Aug. 2018.

[5] R. G. B. Sangalang, S.-H. Luo, H.-C. Wu, B.-Q. He, S.-F. Hsiao, C.-C. Wang, C. Jou, H. Hsia, and D. Yu, "A power effective DLA for PBs in opto-electrical neural network architecture," in Proc. 2022 IEEE Asia Pacific Conference on Circuits and Systems (2022 APCCAS), pp. 46-49, Nov. 2022.

[6] D. Li, M. Liu, S. Gao, Y. Shi, Y. Zhang, Z. Li, P. Y. Chiang, F. Maloberti, and L. Geng, "Low-noise broadband CMOS TIA based on multi-stage stagger-tuned amplifier for high-speed high-sensitivity optical communication," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 66, no. 10, pp. 3676-3689, Oct. 2019.

[7] B. Abdollahi, B. Mesgari, S. Saeedi, E. Roshanshomal, A. Nabavi and H. Zimmermann, "Transconductance boosting technique for bandwidth extension in low-voltage and low-noise optical TIAs," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 69, no. 3, pp. 834-838, March 2022.

[8] E. Kang, M. Tan, J.-S. An, Z.-Y. Chang, P. Vince, N. Sénégond, T. Mateo, C. Meynier, M. A. P. Pertijs, "A variable-gain low-noise transimpedance amplifier for miniature ultrasound probes," IEEE Journal of Solid-State Circuits, vol. 55, no. 12, pp. 3157-3168, Dec. 2020.

[9] S. Elsaegh, C. Veit, U. Zschieschang, M. Amayreh, F. Letzkus, H. Sailer, M. Jurisch, J. N. Burghartz, U. Würfel, H. Klauk, H. Zappe, and Y. Manoli, "Low-power organic light sensor array based on active-matrix common-gate transimpedance amplifier on foil for imaging applications," IEEE Journal of Solid-State Circuits, vol. 55, no. 9, pp. 2553-2566, Sept. 2020.

[10] M. H. Taghavi, L. Belostotski, J. W. Haslett and P. Ahmadi, "10-Gb/s 0.13- m CMOS inductorless modified-RGC transimpedance amplifier," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 62, no. 8, pp. 1971-1980, Aug. 2015.

[11] R. Y. Chen, and Z. -Y. Yang, "CMOS transimpedance amplifier for gigabit-per-second optical wireless communications," IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 63, no. 5, pp. 418-422, May 2016.

[12] M. Seifouri, P. Amiri, and I. Dadras, "A transimpedance amplifier for optical communication network based on active voltage-current feedback ", Microelectronics Journal, vol. 67, pp. 25-31, Sep. 2017.

[13] S. G. Kim, C. Hong, Y. S. Eo, J. Kim and S. M. Park, "A 40-GHz mirrored-cascode differential transimpedance amplifier in 65-nm CMOS," IEEE Journal of Solid-State Circuits, vol. 54, no. 5, pp. 1468-1474, May 2019.

[14] F. Khoeini, B. Hadidian, K. Zhang and E. Afshari, "A transimpedance-to-noise optimized analog front-end with high PSRR for pulsed ToF Lidar receivers," IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 68, no. 9, pp. 3642-3655, Sept. 2021.

[15] X. Li, H. Wang, J. Zhu and C. P. Yue, "Dual-photodiode differential receivers achieving double photodetection area for gigabit-per-second optical wireless communication," IEEE Journal of Solid-State Circuits, vol. 58, no. 6, pp. 1681-1692, June 2023.

4-Port MIMO Antenna with Compact Size and High Gain Characteristics for 5G Applications

Dat Tran-Huy
Faculty of Electrical and Electronic Engineering
PHENIKAA University
Hanoi, Vietnam
dat.tranhuy@phenikaa-uni.edu.vn

Cuong Do-Manh
Faculty of Electrical and Electronic Engineering
PHENIKAA University
Hanoi, Vietnam
cuong.domanh@phenikaa-uni.edu.vn

Phuong Kim-Thi*
Faculty of Electrical and Electronics Engineering
Thuyloi University
Hanoi, Vietnam
*Corres. author: phuongkt@tlu.edu.vn

Duc-Nguyen Tran-Viet
Faculty of Radio-Electronic Engineering
Le Quy Don Technical University
Hanoi, Vietnam
tranvietducnguyen@lqdtu.edu.vn

***Abstract*—This paper presents a simple technique to enhance the gain of a multiple-input multiple-output (MIMO) patch antenna for 5G applications. Conventional approaches, such as utilizing a superstrate or frequency-selective surface, significantly increase the antenna size in both the vertical and horizontal directions. Similarly, combining multiple single-polarized radiating elements with a T-junction divider as one port of the MIMO system also results in large antenna dimensions. To address these disadvantages, the proposed method employs dual-polarized patches and T-junction power dividers. As a result, high-gain radiation can be achieved without requiring additional radiating elements. For demonstration, a 2-port MIMO antenna with two radiating patches and two T-junction dividers is investigated. The antenna has an overall size of $0.80\lambda \times 0.48\lambda \times 0.04\lambda$, an operating bandwidth of approximately 3%, isolation of 20 dB, and a peak broadside gain of 8.0 dBi.**

***Index Terms*—MIMO,high gain, patch**

I. Introduction

There is a strong demand for multiple-input-multiple-output (MIMO) antennas with planar structures and high gain radiation. Various types of MIMO patch antennas have been reported in the literature [1], [2], [3]. However, they suffer from the disadvantage of low gain radiation.

One of the effective methods to improve antenna gain is to use a superstrate or frequency selective surface (FSS) [4], [5], [6], [7], [8]. These structures are generally positioned at a proper distance from the radiating layer, and they also require large lateral dimensions to form a resonant cavity. Consequently, the antenna size in both the vertical and horizontal directions is considerably increased. To achieve high gain with a planar structure, the technique of using multiple single-polarized patches and T-junction power dividers has been proposed in [9], [10], [11], [12]. In such antennas, a one-port MIMO system consists of at least two radiating elements and one power divider. This means that to design a two-port MIMO, four radiating elements are required. It is obvious that gain enhancement involves a trade-off with the overall size.

This paper presents a different approach to the design of a MIMO antenna with a planar structure and high-gain radiation characteristics. The proposed 2-port MIMO design utilizes two dual-polarized patches and two T-junction power dividers. Accordingly, high gain radiation can be obtained with only two radiating elements, rather than the four elements as the method proposed in [9], [10], [11], [12]. An illustration of the proposed approach can be shown in Fig. 1. It is also noted that this configuration allows the MIMO antenna to radiate identical radiation patterns for both ports. Moreover, the radiation pattern is symmetric with the main beam pointing in the broadside direction.

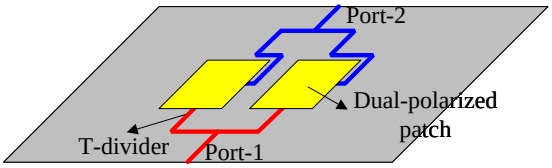


Fig. 1. Proposed approach to design high gain MIMO patch antenna.

II. Dual-polarized patch antenna

Fig. 2 shows the geometrical configuration of the dual-polarized patch antenna. The antenna is coaxially fed at two different feeding positions, F1 and F2. It is printed on the top layer of a Taconic RF-35 substrate (with a dielectric constant of 3.5 and a loss tangent of 0.002). The antenna has two ports, designated as Port-1 and Port-2. The optimal design parameters are as follows: W_s = 30, h_1 = 1.52, l = 15.6, d = 2.8 (unit: mm).

The simulated reflection and transmission coefficients ($|S_{11}|$, $|S_{21}|$), as well as the realized gain are presented in Fig. 3. As observed, the antenna has an impedance matching

979-8-3315-9431-2/25/$31.00 ©2025 IEEE

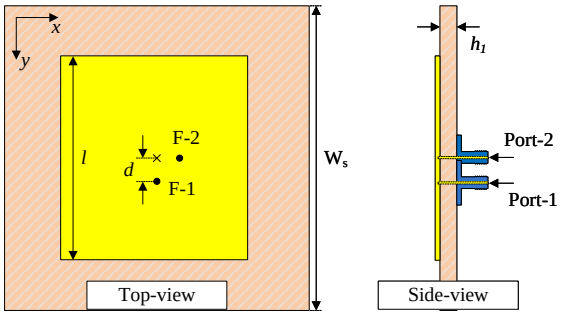
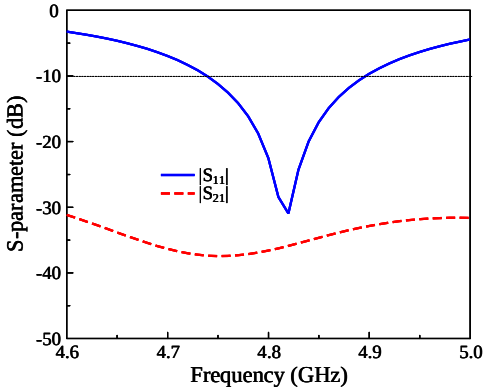
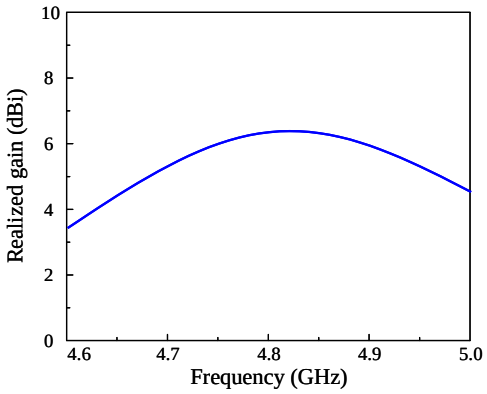


Fig. 2. Geometry of the proposed dual-polarized crossed patch antenna.



(a)



(b)

Fig. 3. Simulated (a) S-parameter and (b) realized gain of the dual-polarized patch antenna.

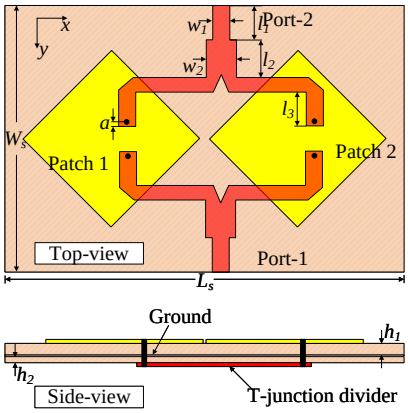
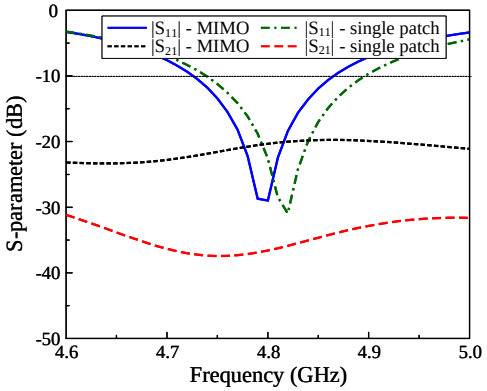
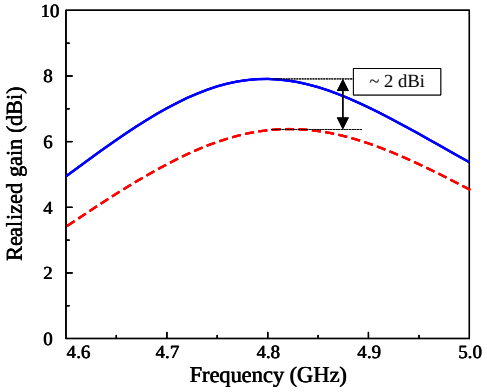


Fig. 4. Geometry of the proposed 2-port MIMO antenna.



(a)



(b)

Fig. 5. Simulated (a) S-parameter and (b) realized gain of the 2-port MIMO antenna.

bandwidth from 4.74 to 4.9 GHz. Moreover, as the ports are located at the null locus of each other, high isolation across the operating bandwidth is achieved. Specifically, the isolation is always better than 33 dB. In terms of gain, the simulated values are approximately 6.2 dBi.

III. 2-PORT MIMO ANTENNA

The dual-polarized patch antenna presented in Section 2 is utilized to design MIMO antenna. The geometry of the proposed 2-port MIMO antenna is depicted in Fig. 4. The MIMO configuration consists of two patches and two T-junction power dividers. The center-to-center spacing of the two patches are 24 mm, corresponding to about 0.38λ at 4.8 GHz. The feeding arrangement is shown in Fig. 4. Noted that these patches are rotated by 45° to ensure that the performance for both ports are identical. The optimal design parameters are as follows: L_s = 50, W_s = 30, h_1 = 1.52, h_2 = 0.76, l = 16, d = 3.6, l_1 = 4, w_1 = 1.7, l_2 = 4.2, w_2 = 2.9, l_3 = 3.8, a = 0.8 (unit: mm).

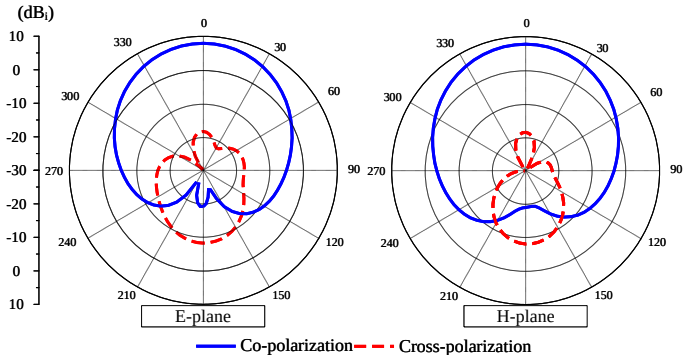


Fig. 6. Simulated radiation patterns of the proposed 2-port MIMO antenna at 4.8 GHz.

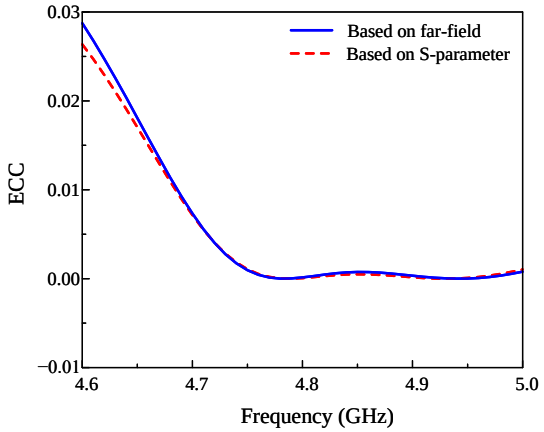


Fig. 7. Calculated ECC and DG of the proposed 2-port MIMO antenna.

The simulated performance of the proposed 2-port MIMO antenna is presented in Fig. 5. The data indicate that the impedance bandwidth ranges from 4.72 to 4.86 GHz. Across this band, the isolation is about 20 dB, which is degraded in comparison with the single patch. This is attributed to the coupling field between the excited position of Patch 1 to the non-excited position of Patch 2 and vice versa. In terms of gain, it can be seen clearly that this value for the MIMO antenna is 8.0 dBi, which represents about 2.0 dBi enhancement in comparison to the single patch.

The simulated gain radiation patterns in E- and H-plane at 4.8 GHz are plotted in Fig. 6. It can be observed that the antenna has a symmetrical radiation pattern around the broadside direction. The maximum gain also occurs on the broadside direction. The difference between co-polarization and cross-polarization is greater than 23 dB. Meanwhile, the front-to-back ratio is about 16 dB. In fact, this relatively low value is attributed to the small ground plane.

Finally, the MIMO diversity performance can be evaluated by the Envelope Correlation Coefficient (ECC). This parameter is calculated based on either far-field radiation pattern results or the S-parameter, as formulated accordingly in equation (1) and (2). The ECC of the 2-port MIMO antenna is depicted

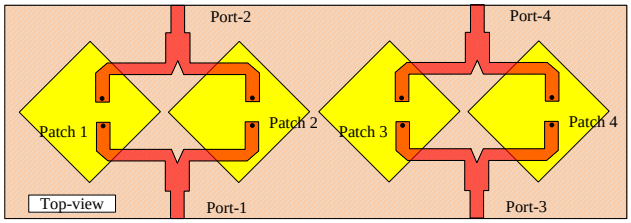


Fig. 8. Geometry of the proposed 4-port MIMO antenna.

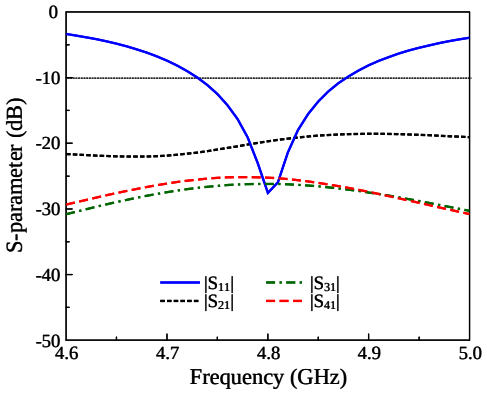


Fig. 9. Simulated S-parameter of the proposed 4-port MIMO array.

in Fig. 7. As seen in the frequency band from 4.72 to 4.86 GHz, the both ECC curves are always below the value of 0.01, which are much smaller than the acceptable value of 0.5.

$$ECC = \frac{\left|\int_{4\pi}\left[\vec{F}_1(\theta,\phi)\cdot\vec{F}_2(\theta,\phi)\right]d\Omega\right|^2}{\left(\int_{4\pi}\left|\vec{F}_1(\theta,\phi)\right|^2 d\Omega\right)\left(\int_{4\pi}\left|\vec{F}_2(\theta,\phi)\right|^2 d\Omega\right)} \quad (1)$$

$$ECC_{ij} = \frac{\left|S_{ii}^* * S_{ij} + S_{ji}^* * S_{jj}\right|^2}{\left(1-|S_{ii}|^2-|S_{ji}|^2\right)\left(1-|S_{jj}|^2-|S_{ij}|^2\right)} \quad (2)$$

IV. 4-PORT MIMO ANTENNA

Fig. 8 shows the geometry of the 4-port MIMO array. The antenna consists of four dual-polarized patches and four T-junction dividers. The element spacing between Patch 2 and Patch 3 is 2 mm. The overall dimensions of the antenna are 98 mm $\times$ 30 mm $\times$ 2.3 mm, corresponding to about 1.44 λ $\times$ 0.48 λ $\times$ 0.04 λ at 4.8 GHz.

Fig. 9 presents the simulated S-parameter of the proposed 4-port MIMO array. The data indicates that the proposed MIMO has operating bandwidth from 4.73 to 4.88 GHz, in which the matching performance is lower than -10 dB and the isolations between all ports are always better than 20 dB. The simulated radiation pattern with Port-1 operation at 4.8 GHz is depicted in Fig. 10, with the corresponding 2-dimensional plots in both E- and H-planes depicted in Fig. 11. It can be seen

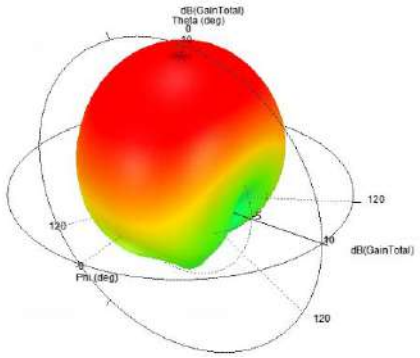


Fig. 10. Simulated 3-D radiation pattern of the 4-port MIMO array at 4.8 GHz.

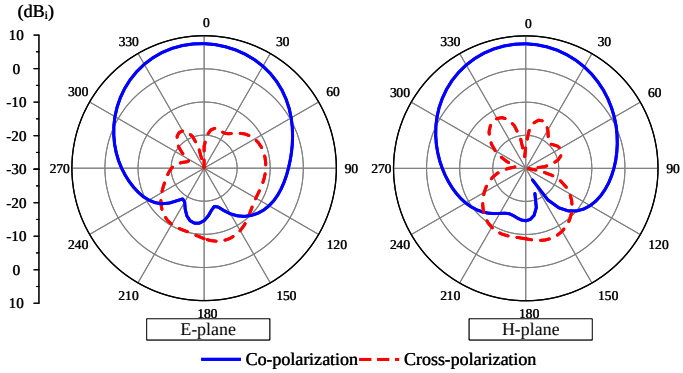
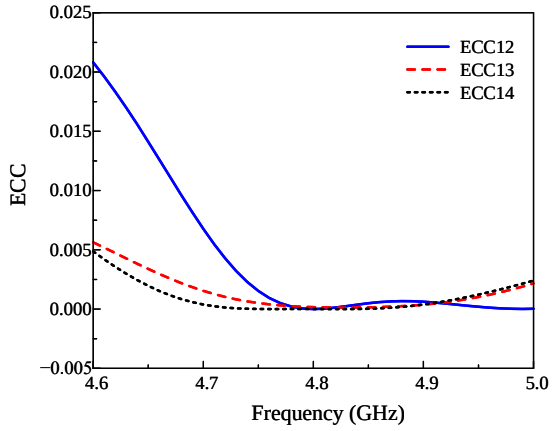
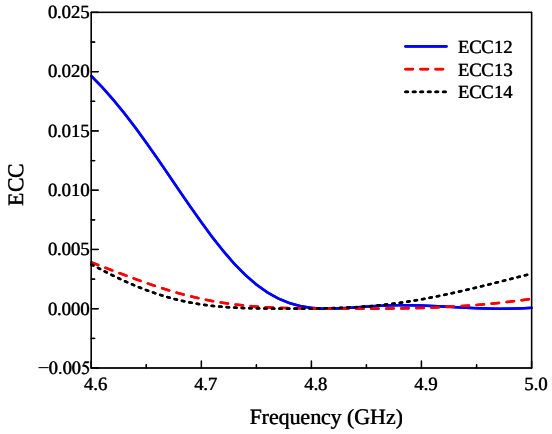


Fig. 11. Simulated radiation patterns of the proposed 2-port MIMO antenna at 4.8 GHz.



(a)



(b)

Fig. 12. Calculated ECC of the 4-port MIMO array. (a) based on far-field calculation, and (b) based on S-parameter calculation.

obviously that the antenna radiates unidirectional beam with a maximum gain of approximately 8.0 dBi. The patterns in E-plane and H-plane demonstrate strong and directional radiation characteristics with relatively low cross-polarization levels. In both planes, the radiation pattern exhibits a forward-directed main beam with good directivity and a largely symmetrical shape. Accordingly, the front-to-back ratio is notably high, with the back-lobe levels suppressed by more than 20 dB. In addition, the cross-polarization remains at least 20 dB below the co-polar component in the main beam region, indicating high polarization purity.

The MIMO diversity performance of the 4-port array has also been calculated and plotted in Fig. 12. Note that the ECC curves are determined based on both far-field and S-parameter results, and the investigation is done with the correlation between Port-1 and the other ports due to symmetrical property. As observed, the ECC values within the operational bandwidth is far lower than the acceptable threshold of 0.5, which fluctuates around 0.001. Besides, there is no difference between the results analyzed by far-field and S-parameter metrics. The low ECC throughout the operating band confirms that the antenna ports are effectively isolated in terms of radiation patterns and polarization. To conclude, the operation characteristics of 4-port and 2-port MIMO arrays are quite similar. It indicates that the proposed design approach has potential to extend to a 1-dimensional large-scale MIMO array.

V. Conclusion

The approach for designing high gain MIMO patch antenna with planar structure has been presented in this paper. Instead of using single-polarized patches, dual-polarized patches and T-junction power dividers are employed. The proposed 2-port MIMO antenna has operating bandwidth of about 3% (4.72–4.86 GHz) with an isolation of better than 2 dB. Additionally, the antenna achieves a high gain of 8.0 dBi and a symmetrical radiation pattern around the broadside direction. The antenna also exhibits low cross-polarization and good MIMO diversity performance. Furthermore, the proposed antenna design has also been tested in an array configuration with four feeding ports and four radiating elements. The simulation results of this 4-element MIMO antenna indicate that the proposed configuration has the potential for scaling up to larger 1-D MIMO antenna arrays.

REFERENCES

[1] D. Gao, Z.-X. Cao, S.-D. Fu, X. Quan, and P. Chen, "A novel slot-array defected ground structure for decoupling microstrip antenna array," *IEEE Transactions on Antennas and Propagation*, vol. 68, no. 10, pp. 7027–7038, 2020.

[2] H. Nguyen-Manh, D.-P. Pham, G. Nguyen-Hoai, H.-H. Tran, and N. Q. Dinh, "A design of mimo antenna with high isolation and compact size characteristics," *IEEE Access*, vol. 11, pp. 93 948–93 955, 2023.

[3] J. Fang, J. Li, P. Xiao, J. Dong, G. Li, S. Du, and W. T. Joines, "Coupling mode transformation-based dielectric surface and metasurface for antenna decoupler," *IEEE Transactions on Antennas and Propagation*, vol. 71, no. 1, pp. 1123–1128, 2023.

[4] S. Tariq, S. I. Naqvi, N. Hussain, and Y. Amin, "A metasurface-based mimo antenna for 5g millimeter-wave applications," *IEEE Access*, vol. 9, pp. 51 805–51 817, 2021.

[5] F. Francis, S. I. Rosaline, and R. S. Kumar, "A broadband metamaterial superstrate based mimo antenna array for sub-6 ghz wireless applications," *AEU-International Journal of Electronics and Communications*, vol. 173, p. 155015, 2024.

[6] M. Salehi and H. Oraizi, "Wideband high gain metasurface-based 4t4r mimo antenna with highly isolated ports for sub-6 ghz 5g applications," *Scientific Reports*, vol. 14, no. 1, p. 14448, 2024.

[7] M. U. Illahi, M. U. Khan, R. Hussain, and F. A. Tahir, "A highly compact fabry perot cavity-based mimo antenna with decorrelated fields," *Scientific Reports*, vol. 12, no. 1, p. 14021, 2022.

[8] T. Hassan, M. U. Khan, H. Attia, and M. S. Sharawi, "An fss based correlation reduction technique for mimo antennas," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 9, pp. 4900–4905, 2018.

[9] M. Pant and L. Malviya, "High gain and low ecc mimo antenna array for millimeter wave communication systems," *Research Square (Research Square)*, 2024.

[10] M. Bilal, S. I. Naqvi, N. Hussain, Y. Amin, and N. Kim, "High-isolation mimo antenna for 5g millimeter-wave communication systems," *Electronics*, vol. 11, no. 6, p. 962, 2022.

[11] M. Khalid, S. Iffat Naqvi, N. Hussain, M. Rahman, Fawad, S. S. Mirjavadi, M. J. Khan, and Y. Amin, "4-port mimo antenna with defected ground structure for 5g millimeter wave applications," *Electronics*, vol. 9, no. 1, p. 71, 2020.

[12] M. A. Haque, M. S. Ahammed, R. A. Ananta, K. Aljaloud, N. M. Jizat, W. M. Abdulkawi, K. H. Nahin, and S. S. Al-Bawri, "Broadband high gain performance mimo antenna array for 5 g mm-wave applications-based gain prediction using machine learning approach," *Alexandria Engineering Journal*, vol. 104, pp. 665–679, 2024.

Single-Layer Compact Wideband Antenna for 5 GHz WLAN Applications

Dieu Thi-Khanh Nguyen
Faculty of Electrical and Electronic Engineering
PHENIKAA University
Hanoi, Vietnam
21012907@st.phenikaa-uni.edu.vn

Noi Truong-Quang
Faculty of Electrical and Electronic Engineering
PHENIKAA University
Hanoi, Vietnam
noi.truongquang@phenikaa-uni.edu.vn

Tu Chu-Anh
Faculty of Electrical and Electronic Engineering
PHENIKAA University
Hanoi, Vietnam
tu.chuanh@phenikaa-uni.edu.vn

Hung Pham-Duy*
Faculty of Electrical and Electronic Engineering
PHENIKAA University
Ha Noi, Viet Nam
*Corres. author: hung.phamduy@phenikaa-uni.edu.vn

***Abstract*—This paper presents a single-layer antenna with wideband operation and compact size characteristics. The antenna is based on a metasurface with 2 × 2-unit cells as radiating element, which is capacitively fed by a crossed patch in the same layer. By excited different operating modes on the MS, wide impedance matching bandwidth can be achieved with a single-layer design. The proposed antenna with compact overall dimensions of 0.45 λ × 0.45 λ × 0.02 λ at 5.0 GHz exhibits wideband operation of 10.6% (4.9–5.45 GHz). Besides, despite having compact size, the proposed design can perform broadside gain of better than 4.5 dBi within the operating bandwidth.**

***Index Terms*—compact, patch antenna, WLAN, broadband**

I. Introduction

In modern wireless communication systems, microstrip patch antennas have emerged as a highly promising technology due to their low fabrication cost, simplicity, and ease of integration, outperforming structures like dipole/monopole, slot, horn, and Vivaldi antennas. In contrast, an issue that may deter the performance of conventional patch antennas is their narrow operating bandwidth of lower than 3% [1].

The methods of broadening the patch antennas' operating bandwidth could be divided into two main categories. The first approach is based on the use of additional layers of dielectric substrate to the primary radiating patches. It could be concluded that using stacked patches [2] could offer wideband operation as it introduces extra resonances to the primary radiator. Recent advancement in artificial material technology has brought state-of-the-art solutions for broadband antenna designs, which are artificial magnetic conductor (AMC) or metasurface (MS) structures. As reported in [3], [4], the antennas obtain not only broadband properties but also size reduction thanks to the use of AMCs. Additionally, using MS like in [5], [6] significantly enhances the antennas' operating band as the MSs are capacitively coupled through slots to excite extra operating modes. In fact, these designs with extra layers often yields the antennas with significant thickness. The remaining solution for wideband patch antennas is introduced within single layer configuration. Broad operating bands could be obtained by etching the antennas with slots on the primary radiators [7], [8] or slots with various shapes on the ground plane [9], [10]. Further bandwidth enhancement involves the integration of parasitic elements around the radiating patches [11], [12]. Despite the excellent performance in expanding the operating frequency range, the employment of slots may deteriorate the antennas' radiation performance, while embedding extra passive elements would cost the antennas' overall dimensions.

This paper introduces a novel design of microstrip patch antenna which could offer broad operational bandwidth. The primary radiating element of the antenna is introduced via an MS layer consisting of 4 unit cells in 2 × 2 arrangement. This radiator is excited by a crossed patch that is placed within the same layer with the MS, which results in a low-profile and compact configuration. Simulation has been implemented, which claims that the proposed method can help to significantly reduce the antenna's overall sizes whilst maintaining remarkable performance.

II. Antenna design

A. Radiating aperture

As the MS layer is the primary radiating aperture of the proposed antenna, its operating frequency range is first considered. There are several methods to determine the operating band of the MS such as dispersion diagram, characteristic mode analysis (CMA). In this paper, the approach of using CMA is applied. Fig. 1 shows the CMA of the utilized 2 × 2 MS layer for the first four modes. It is designed on a 1.52-mm-thick Taconic RF-35 substrate with a dielectric constant of 3.5 and the dimensions of P = 13.5 mm and W = 12.5 mm. According to the CMA, the dominant modes will have

979-8-3315-9431-2/25/$31.00 ©2025 IEEE

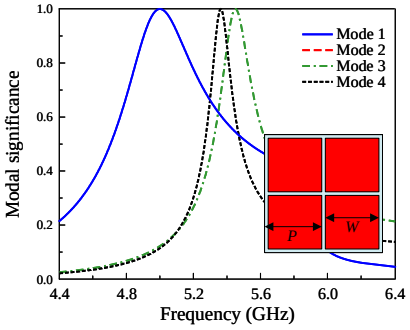


Fig. 1. Simulated modal significance of the array of 2 × 2 MS units.

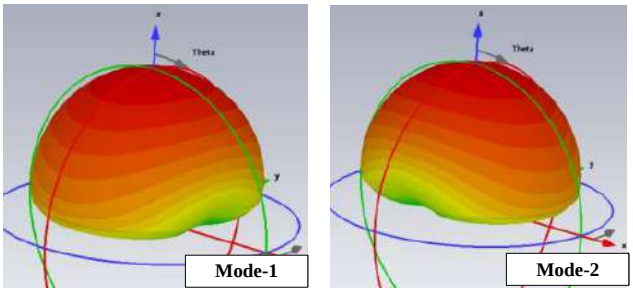


Fig. 2. 3-D radiation patterns of Mode 1 and Mode 2.

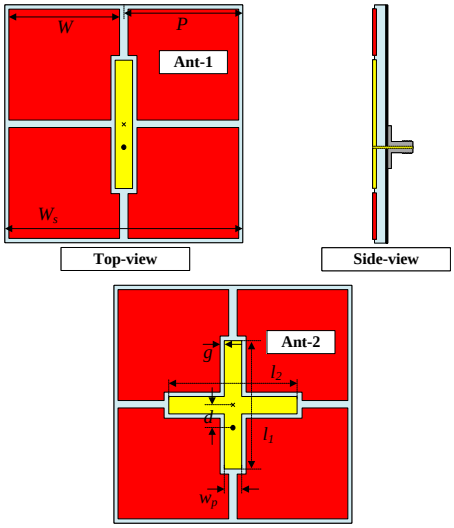


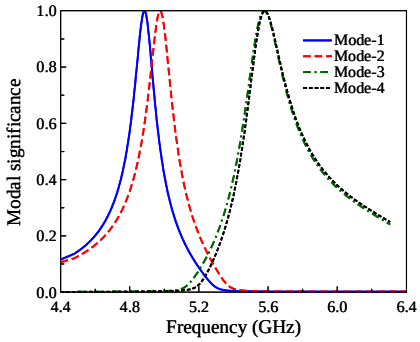
Fig. 3. Different feeding mechanisms for the MS layer.

a modal significance close to 1. Therefore, the used MS layer can resonate around 5.0 GHz.

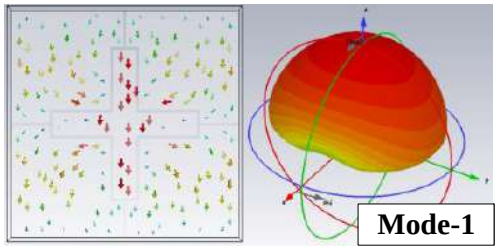
The 3-D radiation patterns of the first two modes (Mode 1 and 2) are presented in Fig. 2. These modes have the capability of radiating unidirectional beam with the strongest power in the broadside direction. Thus, they can be utilized to design directional antennas.

B. Radiating aperture with excitation source

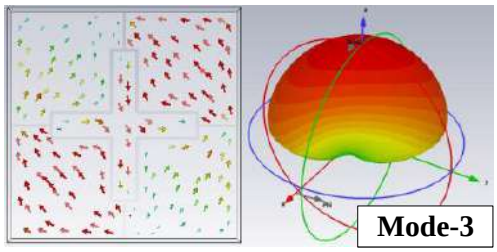
In order to excite the radiating aperture, the primary source of the antenna is considered. Fig. 3 shows two different con-



(a)



(b)

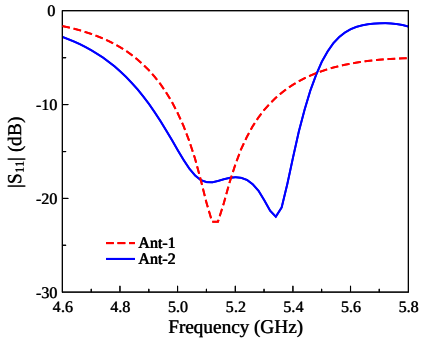


(c)

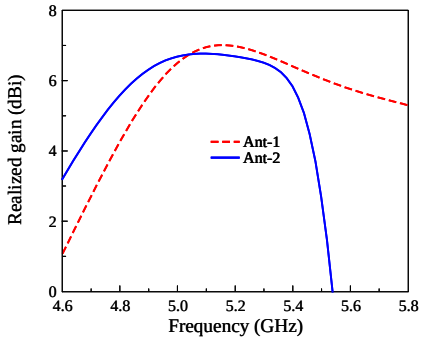
Fig. 4. CMA results of Ant-2. (a) Modal significance, (b) Current distribution and radiation pattern in Mode-1, and (c) Current distribution and radiation pattern in Mode-3.

figurations of the feeding approach. Ant-1 has a feeding part of a rectangular patch, while that for Ant-2 is a crossed patch with different lengths of l_1 and l_2. The optimal dimensions for Ant-2 are W_s = 27, P = 13.5, W = 13.3, g = 0.4, l_1 = 16.6, l_2 = 17.6, w_p = 2.8, d = 3.7 (unit: mm).

This paper shows two different feeding configurations. The radiating aperture of Ant-2 is excited by the crossed patch to achieve wideband operation. To demonstrate the potential of Ant-2 for wideband performance, CMA on this antenna is shown in Fig. 4. As seen, Ant-2 has a good broadside beam around 5.0 and 5.6 GHz. Note that the behavior of Mode-2 and Mode-4 is similar to Mode-1 and Mode-3. The difference is the direction of the vector current, which flows in an orthogonal direction. For Mode-1 and Mode-2, the current flowing in vertical and horizontal directions. Meanwhile, the currents of Mode-3 and -4 distribute along the diagonal direction. When these modes are utilized, wideband performance can be attained by Ant-2.



(a)



(b)

Fig. 5. Simulated performance of Ant-1 and Ant-2. (a) $|S_{11}|$ and (b) realized gain.

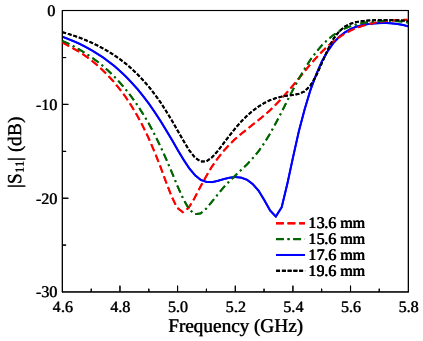
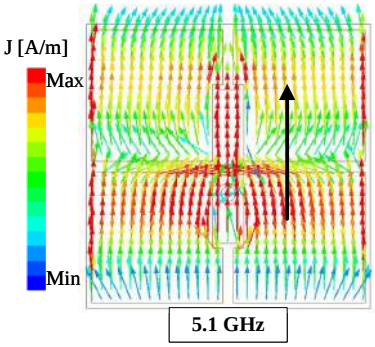


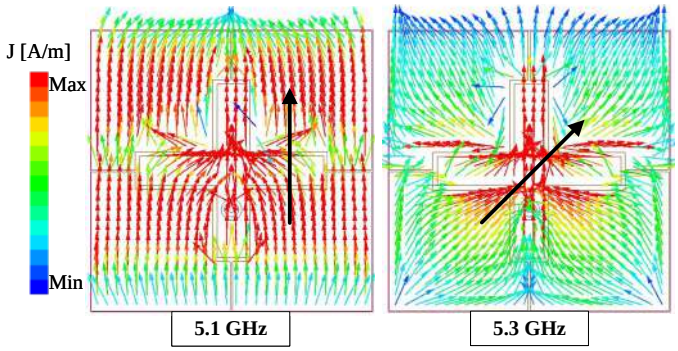
Fig. 6. Simulated $|S_{11}|$ of Ant-2 for different lengths of the horizontal patch, l_2.

III. Antenna Performance

The simulated results in terms of reflection coefficient ($|S_{11}|$) and broadside gain of Ant-1 and Ant-2 are compared and shown in Fig. 5. The simulated impedance matching bandwidth of Ant-1 is 5.8% (5.0–5.3 GHz), in which the reflection coefficient is less than –10 dB. Regarding the realized gain in the broadside direction, the gain across the operating bandwidth is in the range from 6.5 to 7.0 dBi. For Ant-2, additional resonances in the $|S_{11}|$ profile can be obtained, which significantly enhances the operating bandwidth of Ant-2 to 10.6% (4.9–5.45 GHz). Meanwhile, the realized gain within this band is from 4.5 to 6.7 dBi.



(a)



(b)

Fig. 7. Simulated current distributions of (a) Ant-1 and (b) Ant-2.

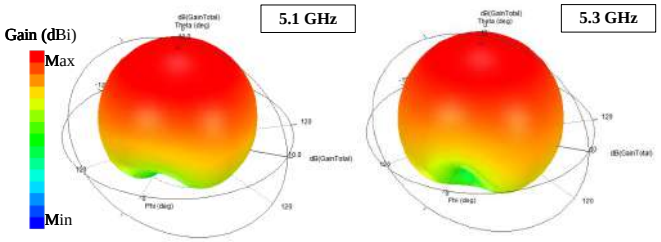
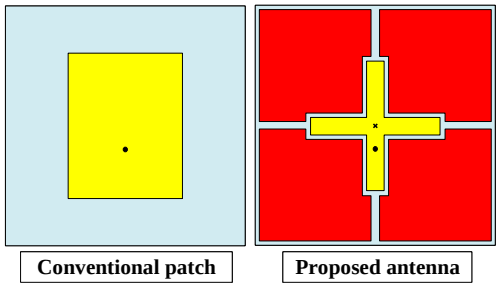


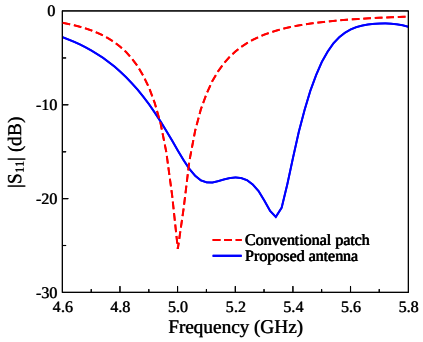
Fig. 8. Simulated 3-D radiation patterns of Ant-2 at different frequencies.

The wideband performance of Ant-2 is obtained due to the excitation of another mode on the MS layer. This is due to the presence of the horizontal patch with the length of l_2. Noted that this is chosen differently with the length of the vertical patch, l_1. Fig. 6 shows the simulated reflection coefficient for different values of l_2. As observed, changing l_2 has a significant impact on the higher resonance, while the lower one is almost stable around 5.1 GHz.

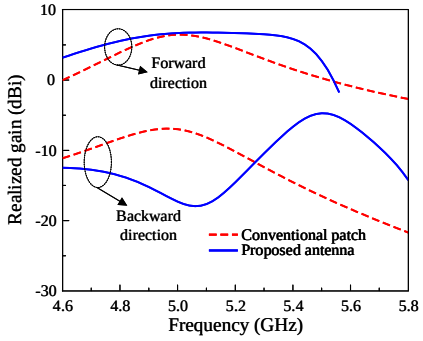
For verification, the simulated current distributions on Ant-1 and Ant-2 at different frequencies of 5.1 and 5.3 GHz are depicted in Fig. 7. As observed, the current flows along the vertical direction of the MS layer at the lower band of 5.1 GHz. Similar phenomena can be observed for both Ant-1 and Ant-2. Meanwhile, the current at 5.3 GHz of Ant-2 is distributed along the diagonal direction of the MS. This could be attributed to the presence of the horizontal patch.



(a)



(b)



(c)

Fig. 9. (a) Geometry, (b) simulated $|S_{11}|$, and (c) simulated realized gain.

The simulated current distribution on Ant-2 is consistent with the CMA discussed in Section 2, in which the lower resonance corresponds to Mode-1 or Mode-2 and the higher resonance corresponds to Mode-3 or Mode-4. The simulated radiation patterns at 5.1 and 5.3 GHz of Ant-2 depicted in Fig. 8 also confirms the directional beam of the antenna.

IV. Comparison with Conventional Design

Fig. 9 shows the comparison between the proposed single-layer antenna to the conventional rectangular patch antenna. Both antennas have similar overall dimensions for a fair comparison. It can be seen that the conventional patch exhibits narrow operating bandwidth from 4.92 to 5.08 GHz, equivalent to about 3.2%. The broadside gain is about 6.4 dBi. However, due to the small ground plane, high diffractions occurring at the edges of the ground plane cause high back radiation of about -7 dBi. For the proposed antenna, the bandwidth is extremely increased to about 10.6%, ranging from 4.9 to 5.45 GHz. The broadside gain within this band is better than 5.0 dBi. It is noted that the back radiation around 5.0 GHz is significantly suppressed to about – 18 dBi. For the proposed antenna, low back radiation of less than -10 dBi can be obtained in the frequency range from 4.9 to 5.3 GHz.

V. Conclusion

The single-layer MS antenna with compact size and wide-band operation characteristics has been presented and investigated in this paper. By using the crossed patch with different lengths as the feeding source for the MS layer, wideband operation can be achieved. Besides, this configuration also allows the proposed antenna to be designed with a single layer. The simulated results demonstrate that the proposed antenna has wide bandwidth of 10.6% (4.9–5.45 GHz) and broadside gain of better than 4.5 dBi with compact overall dimensions of $0.45\lambda \times 0.45\lambda \times 0.02\lambda$ at 5.0 GHz. Accordingly, the proposed design can be a potential candidate for 5 GHz WLAN applications.

References

[1] C. A. Balanis, "Microstrip and mobile communications antennas," *Antenna theory analysis and design*, pp. 788–823, 2016.

[2] V. Sarin, M. Nishamol, D. Tony, C. Aanandan, P. Mohanan, and K. Vasudevan, "A wideband stacked offset microstrip antenna with improved gain and low cross polarization," *IEEE Transactions on Antennas and Propagation*, vol. 59, no. 4, pp. 1376–1379, 2011.

[3] W. Yang, H. Wang, W. Che, and J. Wang, "A wideband and high-gain edge-fed patch antenna and array using artificial magnetic conductor structures," *IEEE Antennas and Wireless Propagation Letters*, vol. 12, pp. 769–772, 2013.

[4] H. Cheng, G. Xiao, and X. Wang, "A low-profile wideband patch antenna with modified parasitic mushroom structures on nonperiodic amc," *IEEE Antennas and Wireless Propagation Letters*, vol. 22, no. 4, pp. 719–723, 2022.

[5] W. E. Liu, Z. N. Chen, X. Qing, J. Shi, and F. H. Lin, "Miniaturized wideband metasurface antennas," *IEEE Transactions on Antennas and Propagation*, vol. 65, no. 12, pp. 7345–7349, 2017.

[6] D. Chen, W. Yang, Q. Xue, and W. Che, "Miniaturized wideband planar antenna using interembedded metasurface structure," *IEEE Transactions on Antennas and Propagation*, vol. 69, no. 5, pp. 3021–3026, 2020.

[7] S. K. Sharma and L. Shafai, "Performance of a novel ψ-shape microstrip patch antenna with wide bandwidth," *IEEE Antennas and Wireless Propagation Letters*, vol. 8, pp. 468–471, 2009.

[8] S. Radavaram and M. Pour, "Wideband radiation reconfigurable microstrip patch antenna loaded with two inverted u-slots," *IEEE Transactions on Antennas and Propagation*, vol. 67, no. 3, pp. 1501–1508, 2018.

[9] K. Mandal and P. P. Sarkar, "High gain wide-band u-shaped patch antennas with modified ground planes," *IEEE transactions on antennas and propagation*, vol. 61, no. 4, pp. 2279–2282, 2013.

[10] A. A. Deshmukh and V. A. Chavali, "Compact wideband microstrip antennas using c, h, and w-shape ground plane for gsm band applications," *Microwave and Optical Technology Letters*, vol. 65, no. 12, pp. 3235–3241, 2023.

[11] Y. Cao, Y. Cai, W. Cao, B. Xi, Z. Qian, T. Wu, and L. Zhu, "Broadband and high-gain microstrip patch antenna loaded with parasitic mushroom-type structure," *IEEE Antennas and Wireless Propagation Letters*, vol. 18, no. 7, pp. 1405–1409, 2019.

[12] D. Yang, H. Zhai, C. Guo, and H. Li, "A compact single-layer wideband microstrip antenna with filtering performance," *IEEE Antennas and Wireless Propagation Letters*, vol. 19, no. 5, pp. 801–805, 2020.

A Novel Compact Broadband Two-Section Branch-line Coupler with Circular-Coupled Lines

1st Xiaoyu Xie
School of Physics and Electronic Information Yunnan Normal University)
Yunnan, China
2224090013@ynnu.edu.cn

2nd Songyuan Yang*
School of Physics and Electronic Information Yunnan Normal Yunnan Key Laboratory of Optoelectronic Information Technology
Yunnan, China
romeoysy@ynnu.edu.cn

3rd Wei Wang
School of Physics and Electronic Information Yunnan Normal Yunnan Key Laboratory of Optoelectronic Information Technology
Yunnan, China
wangwvin@163.com

4th Zhihui Xin
School of Physics and Electronic Information Yunnan Normal Yunnan Key Laboratory of Optoelectronic Information Technology
Yunnan, China
xinzhihui.luncky@163.com

5th Jie Feng
School of Physics and Electronic Information Yunnan Normal Yunnan Key Laboratory of Optoelectronic Information Technology
Yunnan, China
Fengjie123@163.com

6th Hongxia Pu
School of Physics and Electronic Information Yunnan Normal University)
Yunnan, China
2324090005@ynnu.edu.cn

***Abstract*—This paper introduces a novel compact broadband two-section branch-line coupler. In traditional branch-line coupler designs, researchers frequently faced substantial hurdles. They grappled with large-scale dimensions that restricted their applications in space-constrained scenarios, and impedance matching problems led to sub optimal performance. The presented coupler features an ingeniously constructed external matching network from improved circular coupled lines. The use of these circular lines is a key innovation, reducing the overall size by 21.5% in contrast to their traditional counterparts. Additionally, a rectangular defected ground structure (DGS) is precisely etched into the ground plane beneath the three-branch line directional coupler structure. This process significantly improves the characteristic impedance of the branch lines. As a result, this innovative design impressively achieves a measured bandwidth of 46.8% at a center frequency of 2.2 GHz. By harnessing the distinct properties of circular coupled lines and DGS, our proposed coupler effectively overcomes these long-standing limitations, presenting a more efficient and compact solution for diverse applications such as 5G communication systems, wireless sensor networks, and satellite communication devices.**

Keywords—branch-line coupler, circular coupled transmission line, matching circuit, broadband.

I. INTRODUCTION

The branch-line directional coupler is a crucial microwave passive device utilized in various high-frequency systems. It can be used to realize balanced amplifiers, mixers, Butler matrices and array antennas [1] [2]. The directional coupler based on the traditional quarter-wavelength transmission lines exhibits a narrow bandwidth in operation. To increase the operating bandwidth and obtain better performance, the two-stage design and the principle of continuous resonators are adopted to enhance the performance of the coupler bandwidth, moreover, by using the zigzag line design, the compactness of the coupler is achieved, with an overall size of $0.37\lambda_0 \times 0.34\lambda_0 \times 0.0057\lambda_0$ [3]. By cascading the coaxial quarter-wavelength coupler and the band-pass filter [4], the branch-line coupler has a self-supporting structure and attains a bandwidth of 37.84% at a center frequency of 37GHz. Broadband characteristics can be achieved by applying a multi-node coupled transmission line equivalent to cascaded step impedance filters [5]. Through the introduction of the coupler port coupled transmission line as the matching network, also can improve the bandwidth performance [6]. By further adopting a two-stage coupled transmission line and impedance step design, its bandwidth characteristics in the X-band are improved [7]. Although the external matching network can obtain wider bandwidth and flat distribution characteristics [7] [8], these couplers occupy a larger circuit area.

In this paper, in order to achieve a wide bandwidth and compact size, a novel type of branch-line coupler with a circular-coupled transmission lines as a matching network is proposed. The circular-coupled line is integrated at the input/output ports, replacing the original coupled transmission line. In addition, a rectangular defected ground structure (DGS) structure is loaded to increase the width of the impedance-needed characteristic lines for low-cost fabrication. The simulation results and the measured results show that the bandwidth of the proposed design is 48.42%, and the isolation degree between ports is more than 14dB.

The remainder of this paper is organized as follows: Section II provides a comprehensive analysis of the factors influencing the input impedance. Section III details the design, fabrication, and compact implementation of the proposed coupler, along with its measurement results. Finally, Section IV concludes the paper by discussing the significance and potential applications of this work. These contributions aim to advance the state-of-

Yunnan Expert Workstation (No. 202305AF150012)
Foundation for Science Research of Office for Education,Yunnan Province(2022J0130)

the-art in coupler design and provide a foundation for future research in this field.

II. Circular Coupled Lines Structure

The circuit structure of a two-section branch-line coupler with an external matching network is presented in Figure 1. The circular-coupled transmission lines structure serving as a matching network is shown in Figure 2(a). Assuming that the size of the circular part is large enough, the cross coupling effect between the upper and lower halves of the coupled line is negligible, at which point it can be equated to two coupled lines connected in parallel, as shown in Figure 2(b).

The parallel coupled three-wires structure is composed of a transmission microstrip line in the middle and open branches on both sides, and it can be analyzed based on the lossless TEM transmission line model[9].

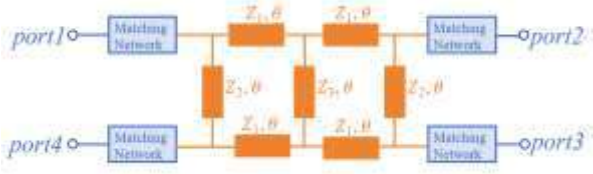


Fig. 1. Circuit configuration for the proposed branch-line coupler

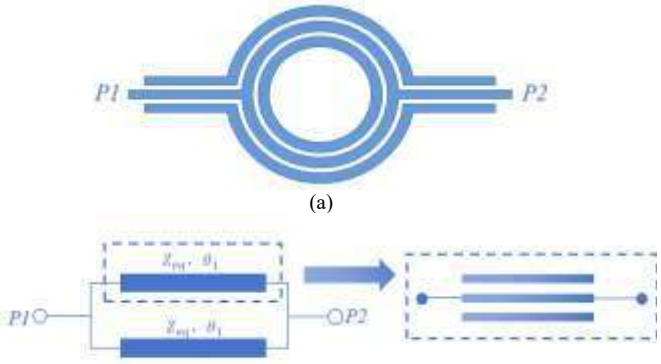


Fig. 2. (a) Circuit structure of circular-coupled lines (b) Equivalent circuit of circular-coupled lines

According to the cross section structure of the coupled line, an circuit equivalent of per unit length can be obtained in Figure 3. For analyzing characteristic impedance of coupled lines, the standard capacitor per unit length,C_{eq}, can be calculated as Eq. (1). Then, the characteristic impedance, Z_{eq}, is shown in Eq. (2), v_p is the phase volocity of the coupled lines.

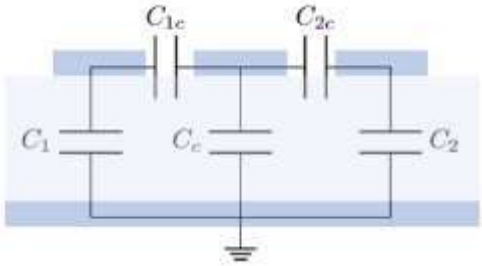


Fig. 3. Parallel coupled three-wires equivalent circuit

$$C_{eq} = C_c + \frac{C_1 C_{1c}}{C_1 + C_{1c}} + \frac{C_2 C_{2c}}{C_2 + C_{2c}} \quad (1)$$

$$Z_{eq} = \frac{1}{v_p C_{eq}} \quad (2)$$

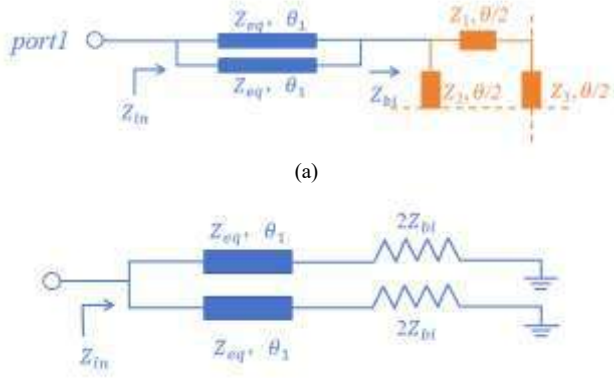


Fig. 4. (a) Quarter one-port circuit of proposed branch-line coupler (b) Schematic diagram of Z_{bi} decomposition into two $2Z_{bi}$ based on symmetry

Due to the highly symmetric structure of the coupler, the impedance matching relationships at the four ports can be simplified to the analysis of a single port using the even-odd mode symmetry analysis method[6]. Figure 4 shows the quarter-port circuit of the proposed branch-line coupler that includes port 1. Based on the symmetry, the impedance Z_{bi} in Figure 4(a) can be decomposed into two $2Z_{bi}$, as shown in Figure 4(b). Through the action of a pair of parallel transmission lines, the total input impedance Z_{in} of the equivalent circuit, which incorporates the circular-coupled line, can be derived, where θ_1 is the electrical length corresponding to the length of the parallel coupled line.

$$Z_{in} = \frac{1}{2} Z_{eq} \frac{2Z_{bi} + jZ_{eq} tan\theta_1}{Z_{eq} + 2jZ_{bi} tan\theta_1} \quad (3)$$

Based on the expression of Z_{in} , it is evident that the achievement of 50Ω port matching can be achieved by adjusting multiple degrees of freedom. Furthermore, since the equivalent capacitance of the coupled transmission lines remains constant over a wide frequency band, broadband matching can be achieved.

III. Coupler Design And Measurement Results

A. *Fabrication design of Two-Section Branch-Line Coupler using DGS*

The primary issue associated with the traditional two-section microstrip line directional coupler is that as the number of branch lines increases, the required characteristic impedance also rises, resulting in a narrower transmission line width[10].

Due to the impact of machining accuracy, it is difficult to realize a transmission line with an overly narrow line width. The high-characteristic-impedance microstrip line loaded with a rectangular DGS can substitute for the theoretical high-characteristic-impedance microstrip line to meet the impedance-matching design requirements[11].

The structure of the coupler with the added rectangular DGS is presented in Figure 5, and Figure 6 shows its simulated S-parameter response, which depicts a 27.27\% fractional bandwidth (FBW) at a 3 dB coupling level at the operating frequency of 2.2 GHz.

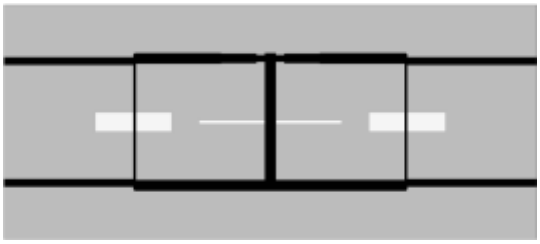


Fig. 5. Two-section branch-line coupler with defected ground structure

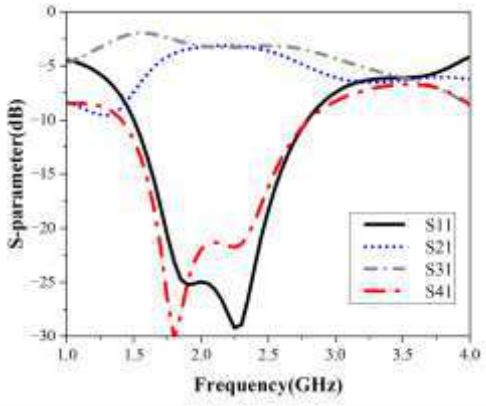


Fig. 6. S-parameters response of branch-line coupler with DGS

B. Compact equal power split coupler Branch-Line Coupler using coupled transmission lines

As mentioned in the former section, broadband characteristics can be achieved by employing coupled transmission lines as a matching network, and the bandwidth of the equal power split coupler is increased by introducing parallel coupled transmission lines at each port. The structure of the parallel coupled transmission line coupler is shown in Figure 7, with a coupled lines length of 19 mm, a line width of 0.2 mm and a pitch of 0.1 mm. Figure 8 shows the simulated S-parameter response of the parallel coupled transmission line coupler. Its relative bandwidth is 35.55%.

Since the branch-line coupler with parallel coupled transmission lines as the matching network occupies a large circuit area, on the premise of not affecting the transmission performance, we use the improved circular-coupled lines to reduce the overall size of the proposed coupler.

In Figure 9, the innermost circle of the circular-coupled line has a radius of 3.5 mm, the width of the circular-coupled line at all levels is 0.2 mm, the spacing is 0.2 mm, and the total length of the coupled lines is 14.9 mm.

The proposed coupler exhibits excellent transmission performance in the 1.8-2.9 GHz band, with the bandwidth further extended to 46.8%. Within this band, the return loss and isolation exceed -14 dB, and the average in-band insertion loss for the through and coupled ports is simulated as -4.28 dB and -4.24 dB, respectively, while the size is reduced by 21.5%. Figure 10 shows the simulated S-parameter response of the proposed coupler.

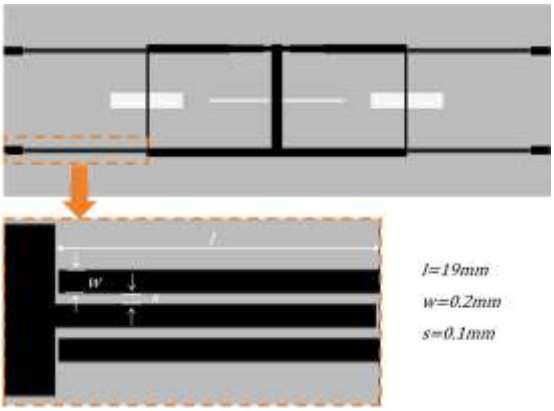


Fig. 7. Simulation pattern of equal power split coupler branch-line coupler with parallel coupled transmission lines

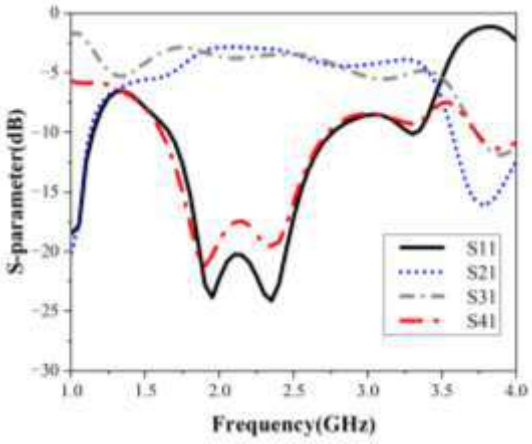


Fig. 8. Simulation results for equal power split coupler branch-line coupler with parallel coupled transmission lines

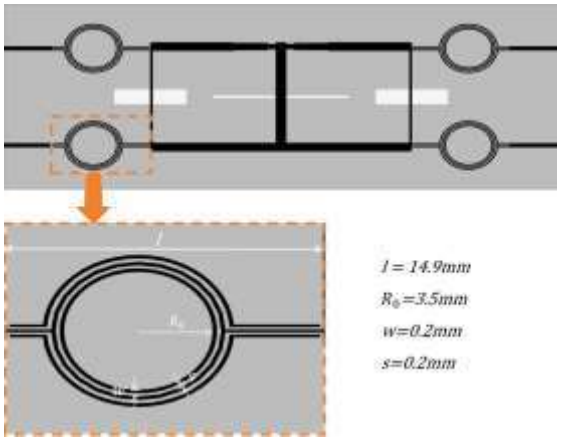


Fig. 9. Simulation pattern of equal power split coupler branch-line coupler with circular-coupled lines

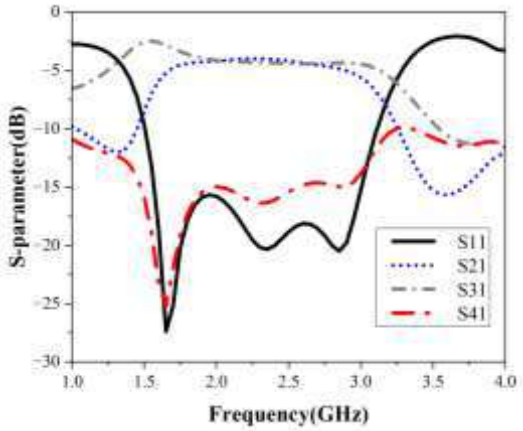


Fig. 10. Simulation results of equal power split coupler branch-line coupler with circular-coupled lines

C. Measurement Results

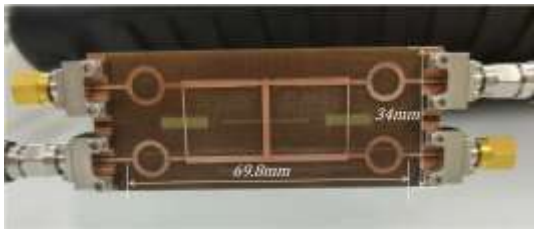
The physical circuit fabricated from the simulation model in Figure 6 is shown in Figure 11(a). The dielectric substrate adopts FR4 with a thickness of 0.6mm, the relative dielectric constant ε_r = 4.4, the tangent dielectric loss angle tangent tanδ=0.02, and the surface metal layer is copper with a thickness of 35um.

The actual circuit was measured using Ceyear RF & Microwave Comprehensive Testers 4957D, and the measurement setup and results are shown in Figure 11, Figure 12. The Thru-Reflect-Line (TRL) calibration method was employed to eliminate the parasitic effects caused by measurement connectors.

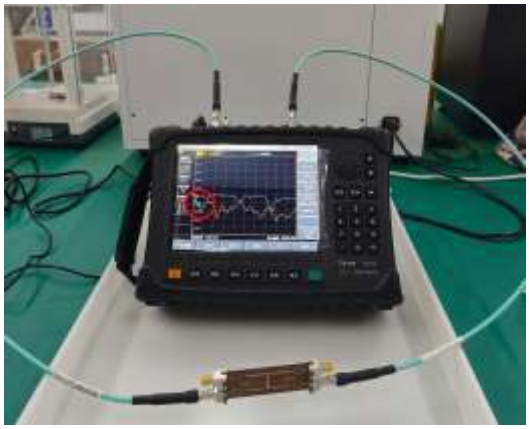
It can be observed that within the 1.8-2.9 GHz frequency band, the insertion losses (S21, S31) from port 1 to port 2 and port 3 respectively are relatively flat, less than -4.8 dB. The measured insertion losses are 0.4 dB worse compared with the simulation results. The possible reason is that the actual losses of the materials are greater than the estimated standard values. A 0.9 dB deviation is observed between the through- and coupled-port insertion losses compared to the simulation, likely due to the sequential two-port testing method with loaded matching components, which introduces minor

connection imbalances.

Furthermore, although the isolation and return loss curves show slight degradation compared to the simulation results, they remain below -12 dB. Taking into account the effects of processing errors, measurement errors, and circuit parasitic effects comprehensively, the circuit performance basically meets the design requirements.

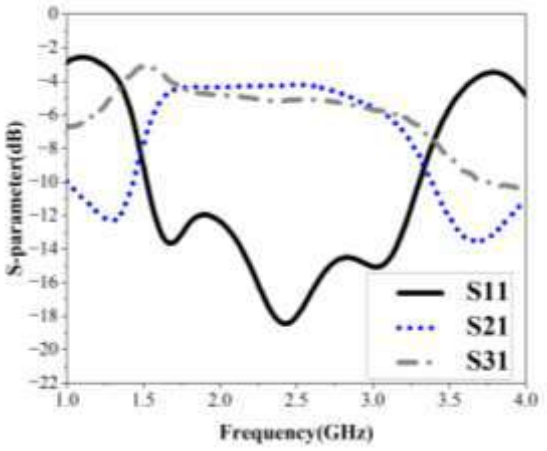


(a)

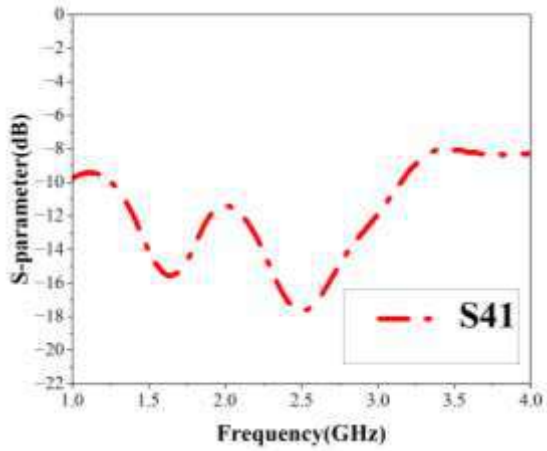


(b)

Fig. 11. (a) Photograph of manufactured compact branch-line coupler using circular-coupled transmission lines (b) Measurement setup



(a)



(b)

Fig. 12. Measured S-parameter results of proposed coupler

IV. Conclusion

This paper proposes a novel two-section broadband coupler utilizing circular-coupled transmission lines as matching networks. By employing circular-coupled lines, which offer a more compact structure compared to traditional parallel coupled lines, the proposed design achieves a broad relative bandwidth of 45%. The coupler is fabricated using a standard PCB process,

and a defective ground structure (DGS) is incorporated to optimize the characteristic impedance.

This innovative coupler exhibits several advantageous features, including a low profile, wide bandwidth, high isolation, and ease of fabrication. These characteristics make it highly suitable for a variety of applications, such as compact high-power combination amplifiers and high-isolation phase shifters, which are critical components in S-band communication and radar systems.

In summary, the proposed design has successfully surmounted the performance bottlenecks of traditional couplers and offered highly competitive solution for modern radio-frequency (RF) and microwave systems. Looking ahead, future research endeavors are projected to converge on two key aspects. The first is to further reduce the size of the device through optimized structural design and the use of novel materials. The second is to implement advanced integrated circuit processes to integrate the coupler with key functional components such as amplifiers and mixers, thereby realizing a high-performance on-chip front-end system integration solution. This envisioned technological trajectory is expected to augment the overall system performance and integration level while curtailing production costs, which will be conducive to 5G/6G front-end modules and Internet of Things (IoT) applications.

ACKNOWLEDGMENT

This work was supported by Yunnan Expert Workstation (No. 202305AF150012)and Foundation for Science Research of Office for Education,Yunnan Province(2022J0130). The author also thanks Yunnan Key Laboratory of Optoelectronic Information Technology for its help.

REFERENCES

[1] V. S. Sokolov and M. A. Stepanov, "Synthesis of a Dual-Band Circular Polarization Antenna for Global Navigation Satellite System GLONASS," 2021 XV International Scientific-Technical Conference on Actual Problems Of Electronic Instrument Engineering (APEIE), Novosibirsk, Russian Federation, 2021, pp. 299-302.

[2] B. Xie, L. -S. Wu and J. Mao, "Reconfigurable Branch-Line Coupler with Tunable Phase Difference," 2019 IEEE MTT-S International Wireless Symposium (IWS), Guangzhou, China, 2019, pp. 1-3.

[3] N-L. Nguyen, "A Compact Planar Branch-Line Hybrid Coupler for Wireless Applications," 2024 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia), Danang, Vietnam, 2024, pp. 1-4.

[4] J. Sheng, H. Yi, Y. Zhu, Y. Yang, W. Zhang and Q. Yang, "A wideband coaxial filtering branch-line coupler," 2024 International Applied Computational Electromagnetics Society Symposium (ACES-China), Xi'an, China, 2024, pp. 1-3.

[5] H. Yan, X. Wu, R. Wang and G. Wang, "Design of Microstrip Type Broadband Directional Coupler," 2023 International Applied Computational Electromagnetics Society Symposium (ACES-China), Hangzhou, China, 2023, pp. 1-3.

[6] T. Kawai, I. Ohta, and A. Enokihara,"Design methods for broadband 3dB branch-line and rat-race hybrids," IEICE Electronics Express, vol.10, no.12, p.1-14, June 2013.

[7] Y. Haoka, T. Kawai and A. Enokihara, "Study of X-Band Broadband Branch-Line Coupler Utilizing Two-Stage Coupled-Transmission Lines," 2019 IEEE Asia-Pacific Microwave Conference (APMC), Singapore, 2019, pp. 1113-1115.

[8] Y. Haoka, T. Kawai and A. Enokihara, "X-Band Broadband Branch-Line Coupler with Loose Coupling Utilizing Short-/Open-Circuited Coupled-Transmission Lines," 2018 Asia-Pacific Microwave Conference (APMC), Kyoto, Japan, 2018, pp. 1498-1500.

[9] David. M. Pozar, Microwave Engineering(4th), pp 48-59, 2012.

[10] S. Shafie, M. El-Sabagh, M. Dessouky, M. Shafee, S. Hammouda and H. Hegazy, "A simple model for on-chip microstrip transmission lines in millimeter wave circuits," 2016 28th International Conference on Microelectronics (ICM), Giza, Egypt, 2016, pp. 121-124.

[11] C. Shi and T. Ling, "Phase Adjustment Study of Rectangular DGS in Millimeter Wave Network," 2021 International Conference on Microwave and Millimeter Wave Technology (ICMMT), Nanjing, China, 2021, pp. 1-3.

Dual-Linearly Polarized and Circularly Polarized Antennas for UWB Applications

Thai Dinh Nguyen
Faculty of Electrical and Electronic Engineering
PHENIKAA University
Hanoi, Vietnam
thai.nguyendinh@phenikaa-uni.edu.vn

Hoang Nguyen-Huy
Faculty of Electrical and Electronic Engineering
PHENIKAA University
Hanoi, Vietnam
hoang.nguyenhuy@phenikaa-uni.edu.vn

Tan Dao-Duc
Faculty of Electrical and Electronic Engineering
PHENIKAA University
Hanoi, Vietnam
hoang.nguyenhuy@phenikaa-uni.edu.vn

Hung Tran-Huy*
Faculty of Electrical and Electronic Engineering
PHENIKAA University
Hanoi, Vietnam
*Corres. author: hung.tranhuy@phenikaa-uni.edu.vn

***Abstract*—This paper presents two different antennas for ultra-wideband (UWB) communication systems. It is worth noting that the conventional UWB antennas are generally designed using slotted or monopole structures, which suffer from low gain and unstable radiation patterns across the operating bandwidth. This paper presents another approach to design UWB antennas using tapered dipole structure. The utilized dipoles can work effectively with both dual-linearly polarized (LP) and circularly polarized (CP) radiation. The simulated operating bandwidths of the presented antennas are from 3.2 to 8.8 GHz, corresponding to about 93.4%. Besides, using a cavity reflector also helps to improve the antenna's broadside gain to about 12 dBi. Another advantage of the presented designs is stable radiation pattern across the operating bandwidth. Accordingly, these performances demonstrate the usefulness of the presented work in the UWB communication systems.**

***Index Terms*—UWB, circularly polarized, dual-linearly polarized, tapered dipole.**

I. Introduction

Ultra-wideband (UWB) technology, operating from 3.1–10.6 GHz, offers high-speed data transmission, low power consumption, and low spectral power density, making it ideal for mordern wireless communications [1]. To design antennas for UWB technology, multi-sense configurations, such as dual linear polarization or dual circular polarization, would be more suitable compared to single-sense ones since they enhances signal stability and mitigates polarization mismatch [2].

Numerous studies have explored the design of UWB antennas with both linear and circular polarization. For LP UWB antennas, monopole [3], [4], slot [5]–[7], and tapered dipole [8] structures are the most commonly used due to their simple and compact layouts. However, their bi-directional radiation patterns lead to lower gain, making them less suitable for practical implementations where antennas are mounted on metallic surfaces. To overcome the limitations of LP UWB antennas, CP UWB antennas have been investigated using different design methodologies. Multi-feed patch array [9], sequentially rotated excitation [10], [11], Archimedean spiral [12], planar spiral [13], modified monopole and ground plane [14], and tapered [15] structures have been proposed to achieve wideband CP performance with unidirectional radiation patterns. However, many of these designs suffer from low gain and unstable radiation patterns, as well as their performances are strongly affected by other circuits when integrating into devices.

This paper presents a solution to UWB antennas by employing tapered dipole configurations, which can provide both dual-linearly polarized and circularly polarized operation by applying proper feeding mechanism. Both designs cover an ultra-wide bandwidth of 93.4%, from 3.4 GHz to 8.8 GHz. To overcome the low-gain issue, the antennas are added with a cavity reflector, which help to enhance their gain up to 12 dBi. An additional benefit of the proposed designs is that they maintain a stable radation pattern within their operating frequency ranges.

II. UWB Dual-Linearly Polarized Antenna

The configuration of the proposed UWB dual-polarized antenna is illustrated in Fig. 1. The antenna consists of two crossed tapered dipoles, which are arranged perpendicularly to produce dual polarization. The antenna is printed on both sides of a Taconic RF-35 substrate (ε_r = 3.5). These dipoles are fed by two 50-Ω coaxial cables, whose outer conductors are connected to one arm of the dipole, and the inner conductors are connected to the other arms. Detailed design parameters are as follows: $L_d = 13.4$ mm, $W_d = 11.8$ mm, $l = 3.8$ mm, $d = 3.6$ mm, $r = 1.5$ mm, $w = 0.25$ mm, $W_s = 30.4$ mm, $h = 0.25$ mm.

The dipole's length is chosen about quarter-wavelength at the lowest operating frequency, and it is calculated based on

979-8-3315-9431-2/25/$31.00 ©2025 IEEE

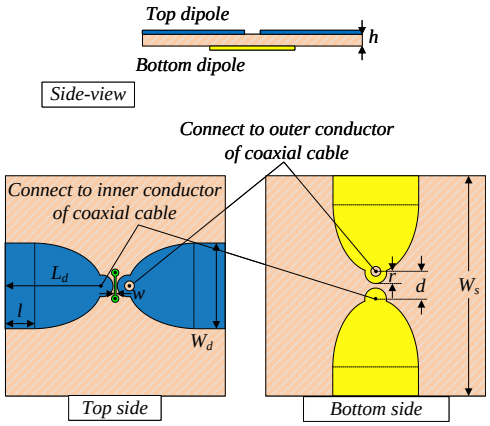


Fig. 1. Geometry of the proposed UWB dual-linearly polarized antenna.

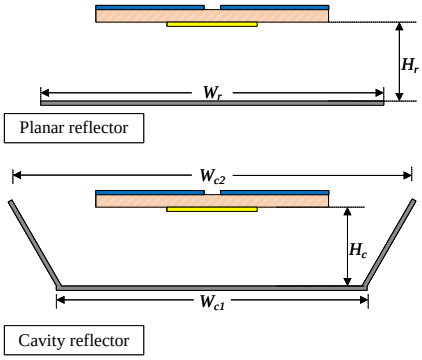


Fig. 2. Geometry of the proposed UWB dual-linearly polarized antenna over different reflectors. $W_r = 50$ mm, $H_r = 15$ mm, $W_{c1} = 50$ mm, $W_{c2} = 65$ mm, $H_c = 15$ mm.

the following equation:

$$L_d = \frac{\lambda_L}{4\sqrt{\varepsilon_{eff}}} \tag{1}$$

in which λ_L is the free-space wavelength, ε_{eff} is the effective permittivity of the used substrate respectively.

As the dipole radiates omni-directional radiation pattern, which features low-gain, a metallic reflector is employed to produce directional beam for high gain performance. Here, two different types of reflectors are investigated. Fig. 2 shows the geometry of the proposed dipole antenna over planar metallic and cavity reflectors. The performance comparison in terms of reflection coefficient, transmission coefficient, as well as broadside gain is depicted in Fig. 3. The data indicates that both antennas exhibit good matching and isolation performance in the frequency range from 3.2 to 8.8 GHz, in which the matching is lower than -10 dB and the isolation is better than 20 dB. However, there is a significant difference with respect to the broadside gain radiation. As observed, using the planar reflector has a drawback of high-power diffraction at the edges of the reflector. Consequently, the broadside gain is low of about 8.0 dBi. In contrast, using the cavity reflector has capability of prohibiting the edges' diffraction and focusing the

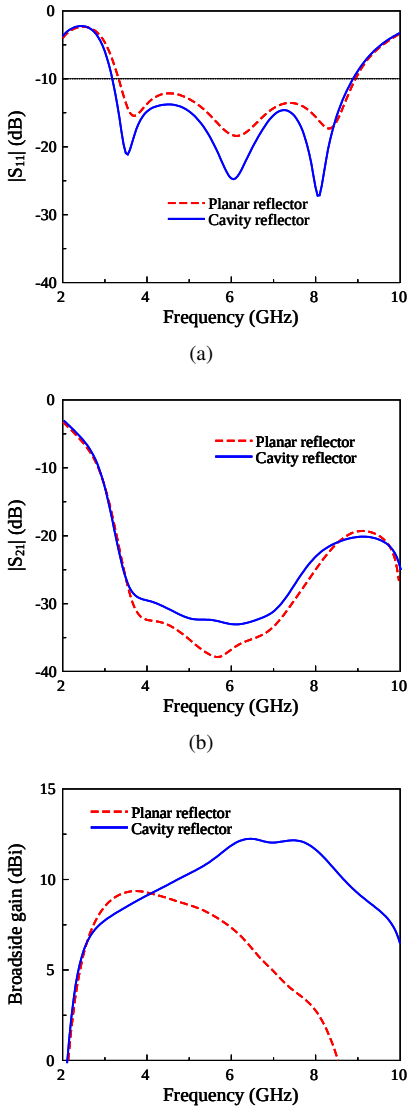


Fig. 3. Simulated (a) $|S_{11}|$, (b) $|S_{21}|$, and (c) broadside gain of the dipole over different reflectors.

radiated power to the forward direction. Accordingly, higher broadside gain can be achieved. Here, the maximum gain of about 12 dBi can be obtained around 7.0 GHz. Fig. 4 shows the simulated gain radiation patterns at 4.0 GHz to demonstrate the radiation features of the proposed antenna over cavity reflector. In general, the antenna radiates a good radiation pattern, which is quite symmetrical around the broadside direction. The cross-polarization level is significantly lower than the co-polarization level of about 30 dB. Additionally, a high front-to-back ratio of about 25 dB is also achieved.

III. UWB Circularly Polarized Antennas

A. UWB CP antenna

In this Section, an UWB CP antenna using the similar dipole discussed in Section 2 is considered. Fig. 5 shows

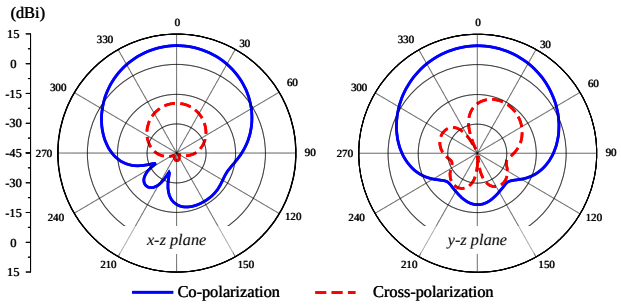


Fig. 4. Simulated gain radiation patterns of the dipole over cavity reflector at 4.0 GHz.

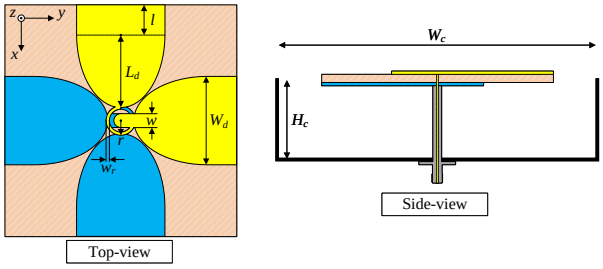


Fig. 5. Geometry of the UWB CP antenna over cavity reflector.

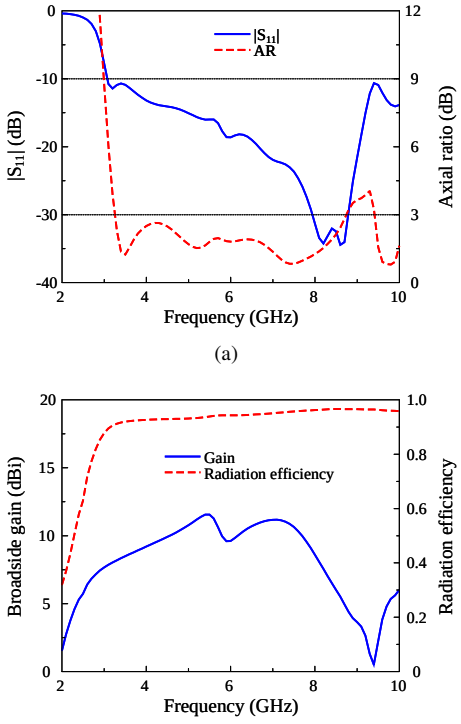


Fig. 6. Simulated (a) $|S_{11}|$, AR and (b) broadside gain and radiation efficiency of the UWB CP antenna.

the geometry of the UWB CP antenna, which consists of two orthogonal dipoles and a metallic cavity. Noted that the previous section demonstrated that using cavity reflectors is better than planar reflectors. Therefore, the performance of the CP antenna over the planar reflector is not shown here for brevity. The antenna is printed on both sides of a Taconic RF-35 substrate with a dielectric constant of ε_r = 3.5. The dipoles are directly fed by a single 50-Ω coaxial cable, in which the inner conductor is linked to the top dipole (yellow color), while the other is connected to the bottom dipole (blue color). The optimized parameters are as follows: $r = 1.8$ mm, $w_r = 0.4$ mm, $w = 1.7$ mm, $L_d = 9$ mm, $W_d = 11.1$ mm, $l = 3.6$ mm, $H_c = 16$ mm, $W_c = 65$ mm.

Fig. 6 illustrates the simulated performance of the presented UWB CP antenna in terms of reflection coefficient, $|S_{11}|$, AR, broadside gain, and radiation efficiency as well. It is obvious that in the frequency band starting from 3.2 to 8.8 GHz, the antenna achieves good matching performance with the reflection coefficient below –10 dB. Besides, the antenna also exhibits well CP radiation with AR of lower than 3 dB. Therefore, it can be concluded that the antenna can work effectively at an extremely large bandwidth of 93.4%. Within this band, the antenna's broadside gain is always better than 6 dBi with a peak gain value of 12 dBi around 5.5 GHz. In terms of radiation efficiency, the simulated data demonstrates that the antenna has an efficiency of greater than 90% across the operating bandwidth.

Fig. 7 shows the simulated gain radiation patterns at 4.0 GHz in two principal planes of $x - z$ and $y - z$. It can be concluded that the antenna exhibits directional beam with the peak radiation towards the broadside direction. Due to the use of the cavity reflector, the antenna achieves high front-to-back ratio of better than 20 dB.

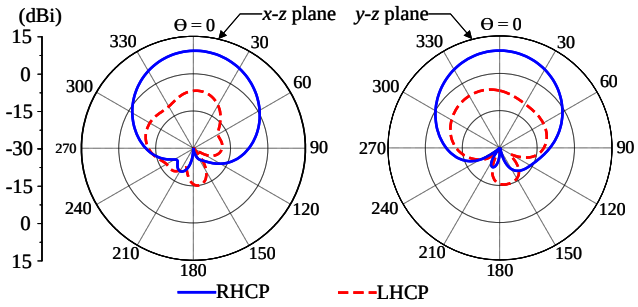


Fig. 7. Simulated gain radiation patterns at 4.0 GHz of the UWB CP antenna.

B. UWB CP antenna with band rejection

To avoid interference with different wireless application systems, a UWB antenna with several band rejections is necessary. Here, two different slots are inserted into each dipole's arm. Fig. 8 shows the geometry of the UWB antenna with band-notched characteristics. The slot's length is chosen about quarter-wavelength at the desired frequencies. The optimal design parameters are as follows: $r = 1.8$, $w_r = 0.4$, $w = 1.7$, $R_m = 9$, $W_e = 11.1$, $L = 3.6$, $s_1 = 1.5$, $L_1 = 10.2$, $w_1 = 0.3$ (unit: mm).

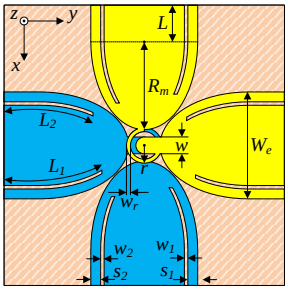


Fig. 8. Top-view of the UWB CP antenna with dual band-notched characteristics.

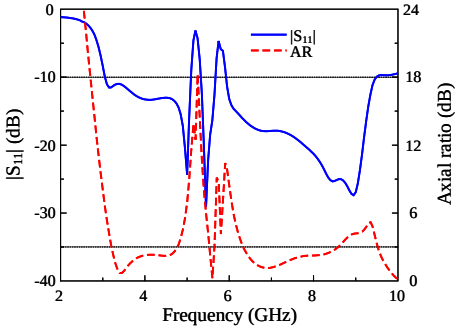


Fig. 9. Simulated $|S_{11}|$ and AR of the proposed notched-band antenna.

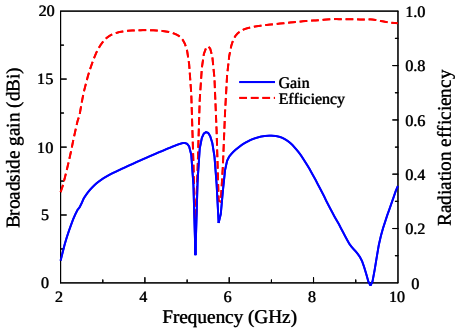


Fig. 10. Simulated gain of the proposed notched-band antenna.

As illustrated in Fig. 9, the antenna exhibits good impedance matching bandwidth in extremely wide frequency range, except the notched bands from 5.1 to 5.4 GHz and 5.6 to 5.9 GHz. For the WLAN application bands of 5.15–5.35 GHz and 5.725–5.825 GHz, the reflection coefficient values are higher than –6.6 dB. With respect to AR, the antenna provides poor CP performance in the notched bands of 4.8–5.4 GHz and 5.6–6.2 GHz, which fully cover the 5.2/5.8 GHz WLAN. An observation from Fig. 10 claims that the antenna exhibits low broadside gain as well as poor radiation efficiency around the rejected bands. Out of the notched bands, the radiating bands observe higher gain of better than 6.2 dBi and radiation efficiency of greater than 82%.

IV. Conclusion

This paper introduces tapered dipole antennas for ultra-wideband systems, addressing conventional limitations of low gain and unstable radiation. By employing the tapered dipole configuration, the proposed antennas achieve dual-linear and circular polarization capabilities, covering an impressive bandwidth of 93.4% from 3.2 to 8.8 GHz. The incorporation of a cavity reflector further enhances the gain to approximately 12 dBi, while ensuring consistent radiation performance across the operational frequency range. These advancements highlight the practical significance of the proposed designs, offering a robust and efficient solution for modern UWB applications.

References

[1] W. Zhuang, X. Shen, and Q. Bi, "Ultra-wideband wireless communications," *Wireless communications and mobile computing*, vol. 3, no. 6, pp. 663–685, 2003.

[2] C. Dietrich, K. Dietze, J. Nealy, and W. Stutzman, "Spatial, polarization, and pattern diversity for wireless handheld terminals," *IEEE Transactions on Antennas and Propagation*, vol. 49, no. 9, pp. 1271–1281, 2001.

[3] K. H. Alharbi, M. Moniruzzaman, R. W. Aldhaheri, A. J. Aljohani, S. Singh, M. Samsuzzaman, and M. T. Islam, "Ultra-wideband monopole antenna with u and l shaped slotted patch for applications in 5g and short distance wireless communications," *International Journal of Applied Electromagnetics and Mechanics*, vol. 66, no. 1, pp. 159–180, May 2021.

[4] S. Baudha, A. Basak, M. Manocha, and M. V. Yadav, "A compact planar antenna with extended patch and truncated ground plane for ultra wide band application," *Microwave and Optical Technology Letters*, vol. 62, no. 1, pp. 200–209, Aug. 2019.

[5] S. Jabeen and G. Hemalatha, "Microstrip fed pi-slot patch antenna with t-slot dgs for uwb applications," *Progress In Electromagnetics Research C*, vol. 129, pp. 63–72, 2023.

[6] B. Hammache, A. Messai, I. Messaoudene, and T. A. Denidni, "Compact stepped slot antenna for ultra-wideband applications," *International Journal of Microwave and Wireless Technologies*, vol. 14, no. 5, pp. 609–615, May 2021.

[7] Y. Zhu, K. Chen, S.-Y. Tang, C. Yu, and W. Hong, "Ultrawideband strip-loaded slotted circular patch antenna array for millimeter-wave applications," *IEEE Antennas and Wireless Propagation Letters*, vol. 22, no. 9, pp. 2230–2234, Sep. 2023.

[8] L. Xiang, F. Wu, K. Chen, R. Zhao, S. Ma, Y. Zhu, C. Yu, Z. H. Jiang, Y. Yao, and W. Hong, "Wideband single and dual linearly polarized magneto-electric dipole array antennas for 5g/6g millimeter-wave applications," *IEEE Open Journal of Antennas and Propagation*, vol. 5, no. 2, pp. 525–539, 2024.

[9] Q.-S. Wu, X. Zhang, and L. Zhu, "Co-design of a wideband circularly polarized filtering patch antenna with three minima in axial ratio response," *IEEE Transactions on Antennas and Propagation*, vol. 66, no. 10, pp. 5022–5030, Oct. 2018.

[10] T. Lou, Z. Shen, and X.-X. Yang, "Circularly polarized uwb antenna based on a single-folded substrate," *IEEE Antennas and Wireless Propagation Letters*, vol. 23, no. 7, pp. 2195–2199, 2024.

[11] R. Xu, Z. Shen, and S. S. Gao, "Compact-size ultra-wideband circularly polarized antenna with stable gain and radiation pattern," *IEEE Transactions on Antennas and Propagation*, vol. 70, no. 2, pp. 943–952, 2022.

[12] Y.-W. Zhong, G.-M. Yang, J.-Y. Mo, and L.-R. Zheng, "Compact circularly polarized archimedean spiral antenna for ultrawideband communication applications," *IEEE Antennas and Wireless Propagation Letters*, vol. 16, pp. 129–132, 2017.

[13] X. Gao and Z. Shen, "Uhf/uwb tag antenna of circular polarization," *IEEE Transactions on Antennas and Propagation*, vol. 64, no. 9, pp. 3794–3802, 2016.

[14] E. Thakur, N. Jaglan, and S. D. Gupta, "Ultra-wideband compact circularly polarized antenna," *Wireless Personal Communications*, vol. 123, no. 1, pp. 407–420, Sep. 2021.

[15] W. Yang, C. Huang, X. Zhu, W. Huang, and L. Xu, "A wideband circularly polarized antenna with high impulse signal fidelity for ir-uwb positioning applications," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 71, no. 4, pp. 1944–1948, 2024.

Design of a Temperature Sensor Based on Auto-zero Technology

Enrui Zhuo
School of Integrated circuits and Electronics
Beijing Institute of Technology
Beijing, China
tassel0@126.com

Xiaoran Li
School of Integrated circuits and Electronics
Beijing Institute of Technology
Beijing, China
xiaoran.li@bit.edu.cn

Lei Zhang*
School of Integrated circuits and Electronics
Beijing Institute of Technology
Beijing, China
zhl666@bit.edu.cn

Jiale Bao
School of Integrated circuits and Electronics
Beijing Institute of Technology
Beijing, China
3220242169@bit.edu.cn

***Abstract*—A Complementary Metal Oxide Semiconductor (CMOS) temperature sensor using auto-zero technology was designed based on the CMOS 55nm process. The sensor consists of a temperature detection module and an analog-to-digital converter (ADC) module. The temperature detection module uses bipolar transistors (BJT) as temperature sensing devices, and the ADC adopts a Sigma-Delta/Cyclic hybrid architecture. In order to achieve accurate quantification of the output signal, auto-zero technology is used to minimize the measurement accuracy degradation caused by the offset voltage of the operational amplifier (op-amp). The simulation results show that under a 2.5V power supply and a 50kHz clock input, the temperature sensor achieved 16 bit temperature measurement results within 30ms conversion time, with a maximum inaccuracy of ± 0.25℃, and achieved a temperature resolution of 4m℃ in the temperature range of -45 to 125℃.**

Keywords—temperature sensor, bipolar transistor, Sigma-Delta/Cyclic ADC, auto-zero technology

I. INTRODUCTION

Temperature is one of the most common physical quantities in nature and has a significant impact on life. Temperature sensors can convert the physical signal of temperature into an electrical signal, so that people can observe the temperature. Currently the mainstream temperature sensor is divided into: thermocouple type [1], resistance type [2], CMOS type [3] and so on. Among them, CMOS-type temperature sensors have low power consumption, small size, high accuracy, wide temperature range and other advantages, so CMOS-type temperature sensors are widely used in a variety of electronic equipment and industrial applications [4]-[6].

The CMOS temperature sensor mainly consists of a temperature sensing module, an ADC module, and some digital modules. The temperature sensing module utilizes the temperature characteristics of the BJT to generate a temperature-dependent signal; the ADC module typically employs a high-order Sigma-Delta ADC or a Sigma-Delta/Cyclic hybrid architecture ADC; and the digital module is responsible for generating the clock signals and controlling the overall feedback logic of the circuit. Operational amplifier (op-amp) is an indispensable unit circuit of this architecture's ADC, because the offset voltage of the op-amp affects the linearity, stability, and temperature drift of the system, significantly limiting the ADC's accuracy. Therefore, this paper designs a high-precision temperature sensor with the application of auto-zero technology for Sigma-Delta/Cyclic hybrid architecture ADC. The auto-zero technology can greatly reduce the op-amp offset voltage, thereby improving the performance of the temperature sensor. Additionally, the combination of the Sigma-Delta ADC and Cyclic ADC using circuit multiplexing technology can significantly reduce conversion time and overall power consumption while ensuring the ADC's accuracy.

II. TEMPERATURE SENSING MODULE

The temperature sensing module adopts a PNP-type BJT as the temperature sensing device. Utilizing the principle that the base-emitter voltage difference of a BJT with different current densities is positively correlated with temperature, the module employs this characteristic. The base-emitter voltage difference of a BJT is expressed as follows:

$$V_{BE} = \frac{kT}{q}\ln\left(\frac{I_C}{I_S}\right) \tag{1}$$

where k is the Boltzmann constant, T is the temperature in Kelvin, q is the unit charge of the electron, I_C is the collector current, and I_S is the saturation current. The saturation current I_S is related to the carrier mobility and intrinsic carrier concentration of the device with the expression:

$$I_S = bT^{4+m}\exp(\frac{-E_g}{kT}) \tag{2}$$

where E_g is the bandgap energy of semiconductor silicon, b is the scaling factor, $m \approx 1.5$. From equations (1) and (2) , we can conclude that V_{BE} is a negatively correlated quantity with temperature.

979-8-3315-9431-2/25/$31.00 ©2025 IEEE

Assuming that the current ratio of the two BJTs is p:1, due to the different current densities, the two BJTs will generate V_{BE1} and V_{BE2} signals, respectively, and the difference between their base-emitter voltages, ΔV_{BE}, is expressed as:

$$\Delta V_{BE} = V_{BE1} - V_{BE2} = \frac{kT}{q}\ln(p) \tag{3}$$

Combining V_{BE} and ΔV_{BE} yields a temperature independent reference voltage V_{REF}, and we define the ratio of $\alpha\Delta V_{BE}$ to V_{REF} as μ.

$$\mu = \frac{\alpha\Delta V_{BE}}{V_{REF}} = \frac{\alpha\Delta V_{BE}}{\alpha\Delta V_{BE} + V_{BE}} \tag{4}$$

μ has a good positive temperature coefficient, and a linear fit of temperature T to μ can be realized from Equation.(5) to obtain the final temperature result.

$$T = A \cdot \mu - B \tag{5}$$

Through simulation, the μ-T curve can first be obtained, where A is the reciprocal of the slope of the curve and B is the difference between the fitted temperature and the actual temperature. After obtaining the values of A and B, the fitting of Equation.(5) can be achieved. The values of A and B are related to the process parameters. [7].

The schematic diagram of the temperature sensing module is shown in Fig. 1. In order to improve the accuracy of V_{BE} and ΔV_{BE}, the circuit adopts Dynamic Elements Matching (DEM) for the current mirrors in the temperature sensing circuit, and six unit current sources are designed to average the mismatch error between the unit current sources.

The temperature error of the temperature sensing module is shown in Fig. 2, and the maximum temperature error is within ± 0.1 ℃.

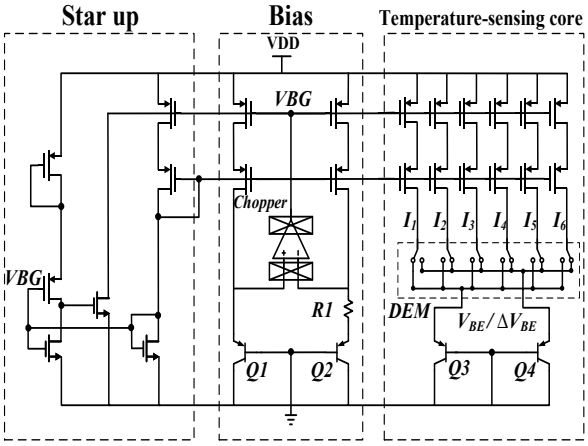


Fig. 1. Structure diagram of temperature sensor

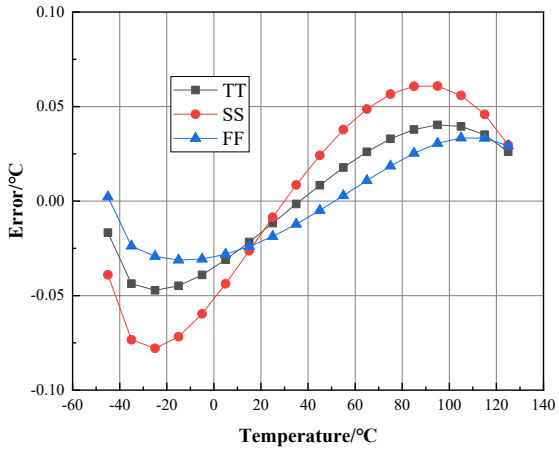


Fig. 2. Simulation results of temperature sensing module

III. Sigma-Delta/Cyclic ADC Module Using Auto-zero Technology

A. *Auto-zero Integrator Circuit Design*

In the integrated circuit fabrication process, due to the uncertainty of each process, the same devices will be mismatched. The effect on the traditional switching capacitor integrator is manifested in the offset voltage, which seriously deteriorates the performance of the circuit. Switching capacitor integrators usually have two working states: the sampling stage and φ_1 the integration stage φ_2. For the traditional integrator, its structure is shown in Fig. 3. In the stage φ_1, the charge stored Q_{S1} in the capacitor C_{S1} is:

$$Q_{S1} = (VIN - 0)C_{S1} \tag{6}$$

After the transition to the φ_2 stage, from the principle of charge conservation, the charge change of the sampling capacitor is equal to the charge change of the integrating capacitor, and the C_{S1} and C_{I1} charge changes by the amount:

$$\Delta Q_{S1} = (VIN - V_{os})C_{S1} = \Delta Q_{I1} \tag{7}$$

It can be found from the conservation of charge $\Delta Q_{I1} \neq Q_{S1}$.

Equation (6) and (7) shows that the output result of the integrator is disturbed by the offset voltage of the op-amp. This defect can be greatly attenuated by the auto-zero technology.

The simplified structure of the auto-zero integrator is shown in Fig. 4. During the φ_1 stage, the op-amp is connected to the unit negative feedback mode, Thus, the voltage at the negative input of the integrator is equal to the offset voltage V_{OS}, the capacitors C_{S2} and C_Z store the charge as:

$$Q_{S2} = (VIN - 0)C_{S2} \tag{8}$$

$$Q_Z = (V_{OS} - 0)C_Z \tag{9}$$

After transition to the φ_2 stage, the voltage on the left plate of C_{S2} remains zero. Due to the continuity of voltage across the capacitors, the voltage on the left plate of C_Z also remains zero. According to the principle of charge conservation, all the charge in C_{S2} is transferred to C_{I2} . At this point, the C_{S2} and C_{I2} charge changes by the amount::

$$\Delta Q_{S2} = (VIN - 0)C_{S2} = \Delta Q_{I2} \tag{10}$$

Equation(8) and (10) show that $\Delta Q_{I2} = Q_{S2}$, we can consider that the charge is completely transferred from the sampling capacitor to the integrating capacitor.

It can be seen from the above equation that under ideal conditions, the auto-zero technology eliminates the offset voltage. In practical situations, compared to the traditional integrator's millivolt-level offset voltage, the effect of the auto-zero technology is significant, typically reducing the input offset to the microvolt level. In Fig. 5, we can observe that the offset voltage of traditional integrator can reach over 2mV(3σ), while after adopting auto-zero technology, the offset voltage can be reduced by about 70% .

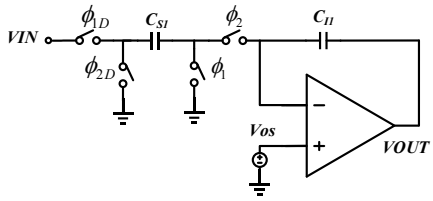


Fig. 3. Traditional integrator structure

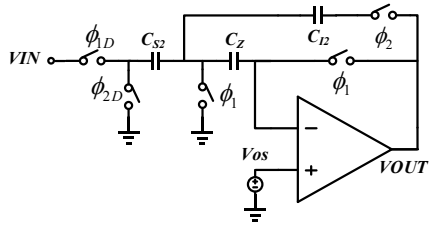


Fig. 4. Auto-zero integrator structure

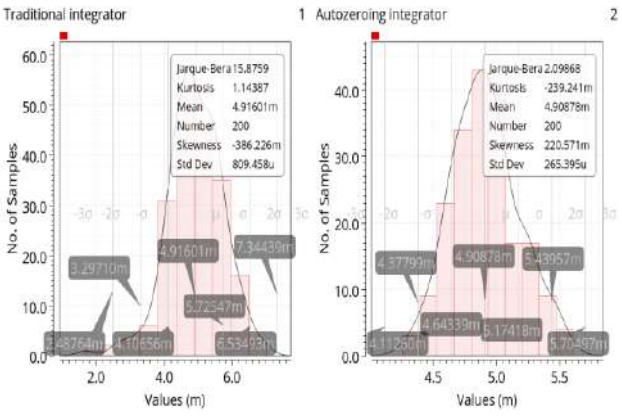


Fig. 5. Integrator simulation results

B. *Sigma-Delta/Cyclic ADC*

After the temperature signal is generated by the temperature sensing module, the ADC module needs to quantize the analog voltage signal output from the temperature sensing module into a digital signal related to the temperature.

Theoretically, after a finite number of conversion cycles M of the Sigma-Delta ADC, a residual voltage ΔV remains at the output of the integrator. In order to achieve an overall higher resolution with fewer conversion cycles, a Cyclic ADC can be used to directly quantize this residual voltage ΔV . Therefore, this paper adopts the Sigma-Delta/Cyclic hybrid architecture ADC, in which the first-order Sigma-Delta ADC quantizes the high-order 9-bit digital code, and the Cyclic ADC quantizes the low-order 7-bit digital code to realize the 16-bit resolution quantization. Compared with the result of using only the first-order Sigma-Delta ADC for quantization, it saves at least 50% of the conversion time. The circuit structure is shown in Fig. 6.

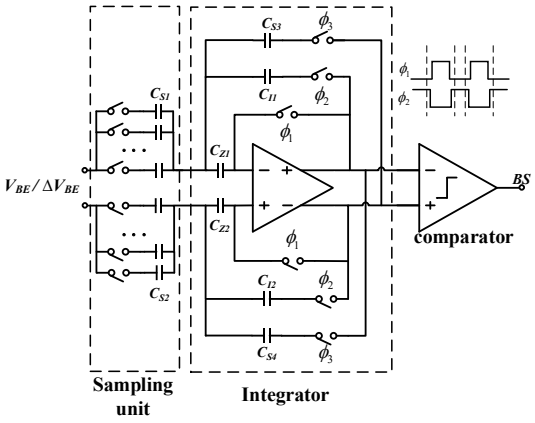


Fig. 6. Hybrid ADC circuit structure

The entire ADC module is divided into three parts: sampling unit, integrator, and comparator. In the sampling unit, this paper uses DEM technology to further eliminate capacitor mismatch; In the integrator circuit, this paper uses a gain-boost operational amplifier to reduce the charge leakage effect caused by limited gain. The operational amplifier has a gain greater than 130 dB in typical (TT), slow (SS), and fast (FF) process corners; In the comparator circuit, this paper uses a two-stage open-loop comparator to provide greater gain and less noise. In TT, SS, and FF process corners, the comparator can distinguish voltages below 1μV, and the maximum noise root mean square voltage is less than $5\,\mu V/\sqrt{Hz}$.

IV. Experimental Results

The layout of the overall temperature sensor is shown in Fig. 7, with a total chip area of 810 μm *600 μm and a core area of 430 μm*345 μm. The circuit designed in this paper is based on a CMOS 55nm process, with a circuit supply voltage of 2.5V and an input clock frequency of 50 kHz. Within the temperature measurement range from -45°C to 125°C, the error curves output by the temperature sensor under TT, SS, and FF process corners are shown in Fig. 8. From the simulation results, the inaccuracy of the sensor is ±0.25°C. Additionally, with a conversion time of 30 ms, the temperature measurement

resolution is 4m°C. A performance comparison of several published temperature sensors is provided in Table I. According to the table, the design presented in this paper exhibits superior accuracy and resolution.

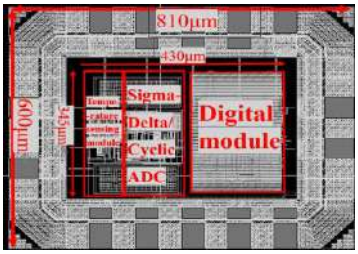


Fig. 7. Layout of temperature sensor

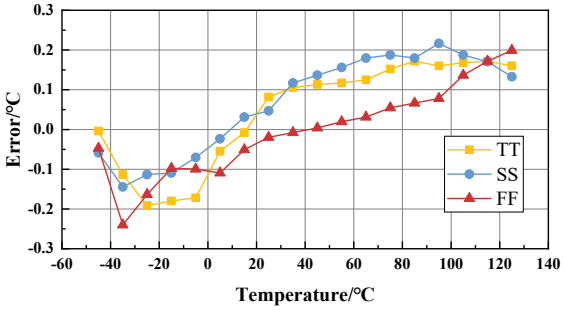


Fig. 8. Error simulation results of temperature sensor

TABLE I. PERFORMANCE COMPARISON OF SEVERAL PUBLISHIED TEMPERATURE SENSOR

Reference	[8]	[9]	[10]	[11]	**This work**
Technology	180nm	40nm	55nm	110nm	**55nm**
Conversion Time	20μs	241μs	10.68ms	0.8ms	**30ms**
Inaccuracy	±1.8°C	±0.4°C	-0.24/+0.28°C	±1°C	**±0.25**°C
Temperature range	-45°C to 125°C	-20°C to 90°C	-40°C to 125°C	-40°C to 140°C	**-45°C to 125°C**
Resolution	N/A	250m°C	10m°C	144m°C	**4m**°C

V. CONCLUSION

This paper proposes a temperature sensor using auto-zero technology based on the CMOS 55nm process. The temperature sensor is mainly divided into a temperature sensing module and an ADC module. The temperature sensing module generates a voltage signal related to temperature, and the ADC module converts the signal into a digital signal. In the design, an auto-zero integrator was used, which reduced the offset voltage by 70% compared to traditional integrators. Within a conversion time of 30ms, the temperature sensor can achieve a maximum inaccuracy of ±0.25°C within the temperature measurement range of -45-125°C, with a temperature measurement resolution of 4m °C. In the future, we will continue to research calibration technology to improve the practicality of our designs.

REFERENCES

[1] U. Sönmez, R. Quan, F. Sebastiano and K. A. A. Makinwa, "A 0.008-mm2 area-optimized thermal-diffusivity-based temperature sensor in 160-nm CMOS for SoC thermal monitoring," ESSCIRC 2014 - 40th European Solid State Circuits Conference (ESSCIRC), Venice Lido, Italy, 2014, pp. 395-398

[2] S. Pan, Y. Luo, S. H. Shalmany and K. A. A. Makinwa, "9.1 A resistor-based temperature sensor with a 0.13pJ·K2 resolution FOM," 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2017, pp. 158-159

[3] . Z. Huang, Z. Tang, X. -P. Yu, Z. Shi, L. Lin and N. N. Tan, "A BJT-Based CMOS Temperature Sensor With Duty-Cycle-Modulated Output and ±0.5°C (3σ) Inaccuracy From −40 °C to 125 °C," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 68, no. 8, pp. 2780-2784, Aug. 2021

[4] K. Souri, K. Souri and K. Makinwa, "A 40μW CMOS temperature sensor with an inaccuracy of ±0.4°C (3σ) from −55°C to 200°C," 2013 Proceedings of the ESSCIRC (ESSCIRC), Bucharest, Romania, 2013, pp. 221-224

[5] Y. -C. Hsu, C. -L. Tai, M. -C. Chuang, A. Roth and E. Soenen, "5.9 An 18.75μW dynamic-distributing-bias temperature sensor with 0.87°C(3σ) untrimmed inaccuracy and 0.00946mm2 area," 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, USA, 2017, pp. 102-103

[6] Z. Tang, Y. Fang, Z. Huang, X. -P. Yu, Z. Shi and N. N. Tan, "An Untrimmed BJT-Based Temperature Sensor With Dynamic Current-Gain Compensation in 55-nm CMOS Process," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 66, no. 10, pp. 1613-1617, Oct. 2019

[7] M. A. P. Pertijs, A. Niederkorn, X. Ma, B. McKillop, A. Bakker, and J.H. Huijsing, "A CMOS smart temperature sensor with a 3σ inaccuracy of ±0.5°C from 50°C to 120°C," in IEEE Journal of Solid-State Circuits, vol. 40, no. 2, pp. 454-461, Feb. 2005

[8] B. Gadogbe and R. Geiger, "A 15μW Low Cost CMOS Smart Temperature Sensor," 2023 IEEE 66th International Midwest Symposium on Circuits and Systems (MWSCAS), Tempe, AZ, USA, 2023, pp. 599-603

[9] Y. -S. Ahn, J. -M. Park, J. -K. Kang and J. Jun, "A ±0.48°C (3σ) Inaccuracy BJT-Based Temperature Sensor With 241 μs Conversion Time for Display Driver IC in 40 nm CMOS," in IEEE Access, vol. 11, pp. 132843-132851, 2023

[10] X. Zhang, F. Qian, J. Xi and L. He, "A BJT-Based Fully Integrated 16-bit ZOOM Temperature Sensor with an Inaccuracy of 0.28°C (3σ) from -40°C to 125°C using improved 1-point Calibration," 2024 IEEE International Symposium on Circuits and Systems (ISCAS), Singapore, Singapore, 2024, pp. 1-5

[11] J. -H. Park, J. -H. Hwang, C. Shin and S. -J. Kim, "A 3.1-μW BJT-Based CMOS Temperature-to-Frequency Converter with Untrimmed Inaccuracy of ±1°C (3σ) from -40°C to 140°C," 2021 IEEE Asian Solid-State Circuits Conference (A-SSCC), Busan, Korea, Republic of, 2021, pp. 1-3

A Comparative Analysis of Machine Learning Models for Lung Cancer Detection

James Alvis R. Azarcon*
Mapua University
Manila, Philippines
jarazarcon@mymail.mapua.edu.ph

Jefferson A. Costales
Mapua University
Manila, Philippines
jacostales@mapua.edu.ph

Shikhar Shiromani
Georgia Institute of Technology
San Francisco, USA
sshiromani3@gatech.edu

***Abstract*—This study compares eight machine learning models for lung cancer detection developed between 2021 and 2024, evaluating their architectures, methodologies, and performance metrics. Results show significant progress, with precision and recall rates exceeding 97%. Region-based models like Faster R-CNN with FBOA achieve top performance (99% precision, 98% recall), while YOLO v8 balances speed and accuracy, and contrastive learning excels in data-scarce settings. Challenges remain in interpretability, clinical integration, and regulation. Future research should explore multimodal integration, explainable AI, federated learning, and lightweight architectures to enhance clinical deployment. These findings provide insights for researchers and clinicians aiming to improve AI-assisted lung cancer diagnostics.**

Keywords—Lung Cancer Detection, Machine Learning, Convolutional Neural Networks, Deep Learning, Medical Image Analysis

I. INTRODUCTION

Lung cancer remains one of the leading causes of cancer-related deaths worldwide, with early detection being crucial for improving survival rates [1]. Traditional diagnostic methods such as low-dose computed tomography (LDCT) scans are effective but often rely on expert radiologists for interpretation, introducing potential for human error and interpretation variability [2]. The integration of artificial intelligence, particularly machine learning techniques, has shown remarkable potential in enhancing detection accuracy and consistency.

Recent years have witnessed rapid evolution in the application of AI to medical imaging, with numerous approaches being developed specifically for lung cancer detection [3]. These range from conventional neural networks to sophisticated architectures that leverage contrastive learning and region-based detection strategies. The diversity of approaches presents both opportunities and challenges for clinical implementation.

This paper aims to systematically analyze and compare different machine learning approaches for lung cancer detection that have emerged between 2021 and 2024. By examining their architectures, methodologies, and performance metrics, we seek to identify trends, strengths, and limitations in current research. The findings will provide valuable guidance for researchers and healthcare professionals interested in developing or implementing AI-assisted lung cancer detection systems.

II. THEORETICAL BACKGROUND

This section outlines the core concepts underpinning AI models for lung cancer detection, spanning computer vision, artificial intelligence, and medical imaging.

A. Machine Learning in Medical Image Analysis

Machine learning, particularly deep learning, has revolutionized medical image analysis by automating feature extraction [3]. Convolutional Neural Networks (CNNs) excel at learning hierarchical representations from raw image data [5]. Medical images like CT scans pose challenges due to subtle abnormalities and their three-dimensional nature, requiring high sensitivity and specificity [13, 19].

B. Neural Network Architectures for Lung Cancer Detection

Various architectures support lung cancer detection. Traditional Artificial Neural Networks (ANNs) rely on preprocessing and feature engineering [7]. CNNs use shared weights to detect spatial patterns [3], while Region-based CNNs, like Faster R-CNN, focus on cancerous regions with high precision [12]. Transformer architectures, adapted via self-attention, enhance performance by focusing on relevant image regions [6, 9].

Transformer architectures, originally developed for natural language processing by [26], have been adapted for medical imaging through mechanisms like self-attention, which allows the model to focus on relevant image regions while maintaining global context [6]. Wang et al. [9] demonstrated this approach with a hybrid CNN-transformer model that achieved significant performance improvements over traditional methods.

C. Clinical Integration Theory

Clinical integration involves technical, workflow, and regulatory considerations [20, 21]. AI serves as a "second reader," requiring interpretability for clinician trust [20, 23]. Federated learning addresses privacy by enabling collaborative training without data sharing [24].

These foundations guide the evaluation of machine learning approaches for lung cancer detection in this study..

D. Object Detection Framework

The Object detection frameworks adapted for medical imaging represent a significant theoretical advancement in lung cancer detection. Single Shot Multibox Detector (SSD) approaches, as explored by Lin et al. [15], detect potential cancer

979-8-3315-9431-2/25/$31.00 ©2025 IEEE

regions in a single forward pass of the network, offering computational efficiency advantages.

You Only Look Once (YOLO) frameworks, particularly the recent V8 variants implemented by Wehbe et al. [10] and Wang et al. [11], process entire images holistically rather than relying on region proposals. This approach enables real-time detection capabilities while maintaining high accuracy, with reported mean Average Precision (mAP) values exceeding 97% [10].

E. Conceptual Framework

Fig 1 is a conceptual framework for AI-based lung cancer detection, illustrating the pipeline from image acquisition and preprocessing. The framework culminates in performance evaluation using metrics like accuracy, precision, recall, IoU, Dice coefficient, and mAP [2, 14, 19], aligning with the paper's emphasis on rigorous assessment and clinical integration [20].

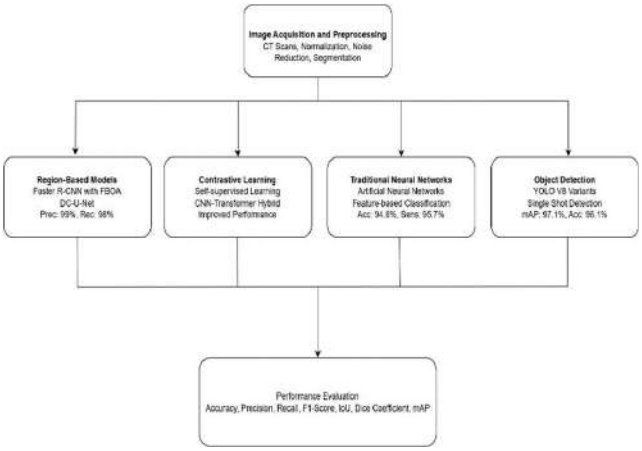


Fig 1. Conceptual Framework for AI-Based Lung Cancer Detection

The study's framework integrates these theories to evaluate AI approaches for lung cancer detection, emphasizing:

1. Traditional Neural Networks: Baseline performance with limited feature extraction [7].
2. Contrastive Learning Approaches: Enhances performance with limited data [8, 9].
3. Object Detection Frameworks like YOLO v8 variants and SSD: Balances speed and accuracy with high mAP [10, 11, 15].
4. Region-Based Approaches such as Faster R-CNN with FBOA: Achieve top precision and recall [12]. Models require rigorous evaluation and clinical integration for practical value [12].

The framework emphasizes that regardless of architectural approach, all models must be evaluated through rigorous performance metrics [13, 14] and can only deliver clinical value when successfully integrated into healthcare workflows [20]. This conceptual understanding provides the theoretical foundation for comparing and evaluating the various AI approaches analyzed in this study.

F. Performance Evaluation Theory

The theoretical basis for performance evaluation in lung cancer detection systems extends beyond simple accuracy measurements [13]. Sensitivity (recall) and specificity represent fundamental metrics that quantify a model's ability to correctly identify positive and negative cases, respectively. For lung cancer detection, high sensitivity is particularly crucial to minimize missed diagnoses, while specificity prevents unnecessary follow-up procedures [2].

Evaluation extends beyond accuracy to include sensitivity, specificity, Intersection over Union (IoU), Dice coefficient, and mAP [2, 14]. High sensitivity minimizes missed diagnoses, while mAP assesses object detection performance [10, 19]..

G. Clinical Integration Theory

The theoretical foundation for clinical integration of AI-based lung cancer detection systems encompasses technical, workflow, and regulatory considerations [20, 21]. The concept of AI as a "second reader" rather than a replacement for radiologists represents a key theoretical framework for clinical implementation [20].

Clinical integration of AI-based lung cancer detection involves technical, workflow, and regulatory challenges [20, 21]. AI acts as a "second reader," requiring interpretability for clinician trust [20, 23]. Federated learning enables privacy-preserving collaborative training across institutions, enhancing model generalizability [24].

H. Advantages and Disadvantages of Methods

1. The Traditional Neural Networks (ANNs):
    - Advantages: Simple, effective baseline [7].
    - Disadvantages: Needs manual feature engineering, limiting accuracy (94.6%) [7, 19].Contrastive Learning (Self-Supervised,
2. CNN-Transformer Hybrids):
    - Advantages: Handles limited data well [8, 9].
    - Disadvantages: High computational cost, augmentation-dependent [9].
3. Object Detection (YOLO v8 Variants):
    - Advantages: Fast, high mAP (97.1%) [10, 11].
    - Disadvantages: Misses small nodules due to low feature resolution [19].
4. Region-Based Approaches (Faster R-CNN with FBOA):
    - Advantages: Top precision (99%), recall (98%) via region focus [12].
    - Disadvantages: Computationally heavy, not real-time [12, 19].

III. Literature Review

A. Evolution of AI in Medical Imaging

The application of artificial intelligence in medical imaging has evolved significantly over the past decade [4]. Early approaches focused on traditional machine learning techniques that required extensive feature engineering. The advent of deep learning, particularly convolutional neural networks (CNNs), marked a significant turning point, enabling automated feature extraction directly from raw image data [5]. More recently, transformer-based architectures and self-supervised learning strategies have further pushed the boundaries of what's possible in medical image analysis [6].

B. Current Machine Learning Approaches for Lung Cancer Detection

Recent literature showcases diverse approaches to lung cancer detection using machine learning. He et al. [7] demonstrated the effectiveness of artificial neural networks in achieving high accuracy rates. Contrastive learning methods, as explored by Ciga & Martel [8] and Wang et al. [9], have shown promise in improving model performance with limited labeled data. Object detection frameworks like YOLO, implemented by Wehbe et al. [10] and Wang et al. [11], have been adapted for medical imaging with impressive results. Additionally, region-based approaches such as those developed by Sinthia et al. [12] have achieved remarkable precision and recall rates.

C. Performance Metrics in Medical AI

Evaluating AI models for medical applications requires careful consideration of performance metrics beyond simple accuracy [13]. Sensitivity (recall) measures the model's ability to correctly identify positive cases, while specificity evaluates its ability to correctly identify negative cases. Other important metrics include precision, F1-score, and area under the ROC curve (AUC). For segmentation tasks, metrics like Intersection over Union (IoU) and Dice coefficient provide insights into the model's spatial accuracy [14].

IV. Methodology

A. Selection Criteria for Models

For this comparative analysis, we selected eight prominent machine learning approaches for lung cancer detection published between 2021 and 2024. The selection criteria included: 1) relevance to lung cancer detection, 2) publication in peer-reviewed journals or conferences, 3) availability of performance metrics, and 4) representation of diverse methodological approaches.

B. Classification of AI Approaches

The selected models were classified into four categories:

- Traditional Neural Networks: Artificial Neural Networks (ANNs) as implemented by He et al. [7]
- Contrastive Learning Approaches: Including self-supervised contrastive learning [8] and semantically-relevant contrastive learning [9].
- Object Detection Frameworks: Single Shot Multibox Detector [15] and YOLO variants [10, 11]
- Region-Based Approaches: Faster R-CNN with fuzzy butterfly optimization [12] and DC-U-Net with dilated convolution [16]

C. Performance Metrics Analysis

We evaluated each model based on the performance metrics reported in their respective publications. These metrics included accuracy, sensitivity, specificity, precision, recall, F1-score, mean Average Precision (mAP), Intersection over Union (IoU), and Dice coefficient. When comparing models, we normalized the metrics where possible to ensure fair comparison.

V. Results

A. Performance Comparison

TABLE I. PERFORMANCE COMPARISON OF AI MODELS FOR LUNG CANCER DETECTION

Author	***Algorithm/Framework***	***Performance***
He et al. [7]	Artificial Neural Network (ANN)	94.6% accuracy, 95.7% sensitivity, 93.5% specificity
Wang et al. [9]	Hybrid CNN-transformer backbone (CTransPath) and semantically-relevant contrastive learning (SRCL)	Significantly improved performance compared to other SSL methods
Ciga & Martel [8]	Self-supervised contrastive learning	Classification outperforms similar high performing self-supervised methods by up to 10 points on average
Lin et al. [15]	Single shot multibox object detector (SSD)	Mean score of 0.7911 for average precision
Sinthia et al. [12]	Faster region-based convolutional neural network (RCNN) aided by fuzzy butterfly optimization algorithm (FBOA)	99% precision, 98% recall, 99% F-measure, 97% accuracy
Wehbe et al. [10]	You Only Look Once V8 (YOLO v8) and TNMClassifier	Mean Average Precision (mAP) of 97.1%
Wang et al. [11]	Enhanced You Only Look Once V8 (YOLO v8-LCM)	Outperforms accuracy of other tested models such as ResNet (ACC = 0.81) with an accuracy of 0.961
Chen et al. [16]	DC-U-Net with dilated convolution	Intersection over Union (IoU) of 0.9627 and Dice coefficient of 0.9743

B. Temporal Trends in Model Performance

Analysis of the temporal progression from 2021 to 2024 reveals a clear trend toward higher performance metrics [17]. Models published in 2022-2024 generally demonstrate improved accuracy, precision, and recall compared to earlier approaches. This trend is particularly notable in the evolution of

object detection frameworks, where YOLO v8 variants [10, 11] show substantial improvements over earlier detectors.

C. Architectural Innovations

Recent architectural innovations have contributed significantly to performance improvements [18]. The integration of transformer architectures with CNNs, as demonstrated by Wang et al. [9], has enabled better feature extraction and representation learning. Similarly, the adaptation of object detection frameworks specifically for medical imaging has yielded impressive results, with YOLO v8 variants achieving mAP above 97% [10].

VI. Discussion

A. Strengths and Limitations of Different Approaches

Traditional Neural Networks: Offer baseline performance but lack advanced feature extraction [7].

Contrastive Learning Approaches: Excels with limited labeled data, learning useful representations [8, 9].

Object Detection Frameworks: Balances speed and accuracy for real-time use, but struggles with small nodules [10, 11, 19].

Region-Based Approaches: Achieves top precision (99%) and recall (98%), yet computationally intensive [12].

B. Clinical Implications

AI models with 97-99% accuracy, precision, and recall [10, 12] can enhance clinical practice as second readers, reducing missed diagnoses [20]. Challenges include workflow integration, interpretability, and regulatory hurdles [21]..

C. Future Research

Several promising avenues for future research emerge from this analysis:

1. Multimodal Integration: Combine imaging with clinical data for better accuracy [22].
2. Explainable AI: Use attention mechanisms for transparency [23].
3. Federated Learning: Enable collaborative training while preserving privacy [24].
4. Lightweight Architectures: Develop efficient models for resource-limited settings [25].

VII. Conclusion

From 2021-2024, region-based models like Faster R-CNN with FBOA lead with 99% precision and 98% recall [12], while YOLO v8 offers speed and strong performance [10, 11]. Contrastive learning aids data-scarce settings [8, 9]. Challenges in interpretability, integration, and regulation persist, with future work focusing on explainable AI, multimodal data, and lightweight models for clinical deployment.

Despite promising results, challenges remain in ensuring clinical interpretability, AI integration and addressing regulatory requirements. As evidenced by prior research on technology integration in Philippine healthcare, these challenges prevent the seamless adoption of technology in the health industry [26.27,28]. Future research should prioritize explainable AI, multimodal data integration, and lightweight architectures for deployment in diverse clinical settings. Ultimately, the collaboration between AI researchers and healthcare professionals will be pivotal in translating high-performance detection models into reliable, patient-centered diagnostic tools.

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," CA Cancer J. Clin., vol. 68, no. 6, pp. 394–424, 2018.

[2] National Lung Screening Trial Research Team, "Reduced lung-cancer mortality with low-dose computed tomographic screening," N. Engl. J. Med., vol. 365, no. 5, pp. 395–409, 2011.

[3] G. Litjens et al., "A survey on deep learning in medical image analysis," Med. Image Anal., vol. 42, pp. 60–88, 2017.

[4] A. Esteva et al., "A guide to deep learning in healthcare," Nat. Med., vol. 25, no. 1, pp. 24–29, 2019.

[5] H. Shen, "Deep learning gives cancer research a boost," Nature, vol. 555, no. 7697, pp. 436–437, 2018.

[6] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learn. Represent., 2021.

[7] X. He et al., "Artificial neural network for lung cancer detection using computed tomography scan imaging," J. Med. Imaging Health Inf., vol. 11, no. 4, pp. 1113–1121, 2021.

[8] O. Ciga and A. Martel, "Self-supervised contrastive learning for digital histopathology," Mach. Learn. with Appl., vol. 6, p. 100198, 2021.

[9] Z. Wang et al., "TransPath: Transformer-based self-supervised learning for histopathological image classification," Med. Image Anal., vol. 76, p. 102305, 2022.

[10] A. Wehbe et al., "Enhanced lung cancer detection using YOLO v8 and a novel TNM classification approach," IEEE J. Biomed. Health Inform., vol. 28, no. 2, pp. 754–763, 2024.

[11] J. Wang et al., "YOLO v8-LCM: A large context modeling approach for improved lung cancer detection," IEEE Trans. Med. Imaging, vol. 43, no. 1, pp. 217–228, 2024.

[12] K. Sinthia et al., "Faster RCNN aided by fuzzy butterfly optimization for enhanced lung cancer detection," Int. J. Imaging Syst. Technol., vol. 32, no. 4, pp. 1524–1537, 2022.

[13] R. Smith-Bindman et al., "Metrics for assessing the quality of value and utility of patient-centered outcomes for diagnostic tests," Acad. Radiol., vol. 27, no. 2, pp. 178–186, 2020.

[14] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," BMC Med. Imaging, vol. 15, no. 1, p. 29, 2015.

[15] Y. Lin et al., "Single shot object detection for lung cancer screening," Comput. Methods Programs Biomed., vol. 213, p. 106523, 2022.

[16] J. Chen et al., "DC-U-Net with dilated convolution for medical image segmentation," IEEE Trans. Med. Imaging, vol. 40, no. 10, pp. 2688–2699, 2021.

[17] W. H. Dunn, T. M. Dehkordi, and F. Ghazvinian, "Comparing the performance of machine learning methods for lung cancer detection (2016-2022)," J. Digit. Imaging, vol. 35, no. 6, pp. 1575–1592, 2022.

[18] P. Meyer et al., "Survey on deep learning for pulmonary medical imaging," Front. Med., vol. 7, p. 670, 2020.

[19] Z. Zhu, Y. Fu, and Q. Li, "A comprehensive review of small object detection: Benchmarks, challenges, and solutions," Int. J. Comput. Vis., vol. 130, no. 9, pp. 2337–2361, 2022.

[20] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nat. Med., vol. 25, no. 1, pp. 44–56, 2019.

[21] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," N. Engl. J. Med., vol. 380, no. 14, pp. 1347–1358, 2019.

[22] N. C. Fernandez et al., "Multimodal deep learning for lung cancer detection and classification," Med. Image Anal., vol. 72, p. 102120, 2021.

[23] S. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," Nat. Mach. Intell., vol. 2, no. 1, pp. 56–67, 2020.

[24] T. S. Brisimi et al., "Federated learning of predictive models from federated electronic health records," Int. J. Med. Inf., vol. 112, pp. 59–67, 2018.

[25] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 6848–6856, 2018.

[26] M. O. Gallardo, J. Dela Torre, and R. Ebardo, "The role of initial trust in the behavioral intention to use telemedicine among Filipino older adults," *Gerontology and Geriatric Medicine*, vol. 10, 2024.

[27] R. Ebardo and J. B. Tuazon, "Identifying healthcare information systems enablers in a developing economy," in *2019 IEEE 6th International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, Dec. 2019, pp. 1–6.

[28] M. Bernabe and R. Ebardo, "Barriers leading to the discontinuance of telemedicine among healthcare providers: A systematic review," *Data and Metadata*, vol. 4, p. 440, 2025. doi: 10.56294/dm2025440

Identification And Diagnosis Of Eye Diseases Using Deep Learning And Yolo

Vy Thi Thanh Huong[1]
[1]*Faculty of Electrical Engineering Technology*
Industrial University of Ho Chi Minh City (IUH)
No 12 Nguyen Van Bao, Ho Chi Minh, 70000, Southern, Vietnam
E-mail: vythithanhhuong@iuh.edu.vn
ORCID ID: https://orcid.org/0000-0001-7636-4378

Tran Tham Hoang Long[2]
[2]*Faculty of Electronics Technology*
Industrial University of Ho Chi Minh City (IUH)
No 12 Nguyen Van Bao, Ho Chi Minh, 70000, Southern, Vietnam
E-mail: tranthamhoanglong01@gmail.com

Huu Q. Tran [2,*]
[2]*Faculty of Electronics Technology*
Industrial University of Ho Chi Minh City (IUH)
No 12 Nguyen Van Bao, Ho Chi Minh, 70000, Southern, Vietnam.
*Corresponding author
Email: tranquyhuu@iuh.edu.vn

***Abstract*—This paper explores the identification and diagnosis of eye diseases using YOLO technology and convolutional neural networks. Initially, we conduct an extensive review of convolutional neural networks, YOLO technology, and related studies from both domestic and international sources. We then create a database by capturing images of normal and diseased eyes using cameras and mobile phones from patients at the hospital, resulting in approximately 1,500 images of diseased eyes and 500 images of healthy eyes. These images are labeled and tested on Colab to evaluate the database's reliability. If the database is found lacking in size or reliability, we augment it with additional images. Once the database is finalized, we write the code, load the database, and test the program's ability to recognize eyes captured by cameras. Any errors encountered are addressed and corrected. Ultimately, we develop a stable identification system with high reliability and write an embedded program for Raspberry Pi 4. We conduct experiments and rectify any errors discovered during this process. As a result, the recognition efficiency of eye diseases using the constructed database exceeds 94%.**

Keywords— Crossed eyes, identification, python, raspberry, YOLO

I. INTRODUCTION

The accurate identification and diagnosis of eye diseases are essential for preventing vision impairment and improving patient outcomes. Traditional diagnostic methods, while effective, often require significant time, expertise, and resources, limiting their accessibility in resource-constrained settings. To address these challenges, advanced technologies such as convolutional neural networks (CNNs) and object detection frameworks like YOLO (You Only Look Once) have emerged as powerful tools for automating and enhancing the diagnostic process. These technologies not only offer faster and more accurate identification of eye diseases but also pave the way for cost-effective and scalable solutions in ophthalmology [1, 2]. This study leverages the strengths of CNNs and YOLO technology to develop a robust system for identifying and diagnosing eye diseases. A comprehensive analysis of existing CNN architectures and YOLO frameworks is conducted, incorporating insights from both domestic and international research [3, 4]. To support the development and evaluation of the system, we create a specialized database comprising approximately 2,000 labeled images, including 1,500 images of diseased eyes and 500 images of healthy eyes, collected from hospital patients using cameras and mobile phones [5, 6]. Data augmentation techniques are employed to enhance the size and reliability of the database as needed [7, 8]. The system development process includes training and testing the CNN and YOLO models on the constructed database, optimizing the models for high recognition accuracy, and embedding the final program onto a Raspberry Pi 4 platform. Rigorous testing and debugging ensure the stability and reliability of the system [1, 9]. Experimental results demonstrate that the system achieves a recognition efficiency exceeding 90%, highlighting its potential as a practical solution for real-world applications [5, 10].

This paper provides a detailed account of the design, implementation, and evaluation of the proposed eye disease identification system, showcasing the effectiveness of integrating YOLO technology and CNNs in addressing critical challenges in ophthalmology. By contributing to the advancement of automated diagnostic tools, this research aims to improve accessibility to early detection and treatment of eye diseases, particularly in underserved regions [11-13].

II. THEORETICAL BASIS

A. Deep Learning YOLO

YOLO (You Only Look Once) is a CNN model used for object detection in computer vision. It is a deep learning-based object detection method that allows for the quick and efficient identification and classification of objects in images and videos. YOLO was developed to address the need for efficient and rapid object detection. The YOLO algorithm operates using three main techniques:

- **Residual Block:** A fundamental block in CNN architectures used in deep learning models like ResNet.

Figure 1 depicts how the input image, captured by the camera, is divided into a grid structure for object detection. Each grid cell predicts the presence of an object within its region, including the object's location and class. This grid-based division, a core component of YOLO models, enables efficient, real-time object localization and classification across the entire image.

979-8-3315-9431-2/25/$31.00 ©2025 IEEE

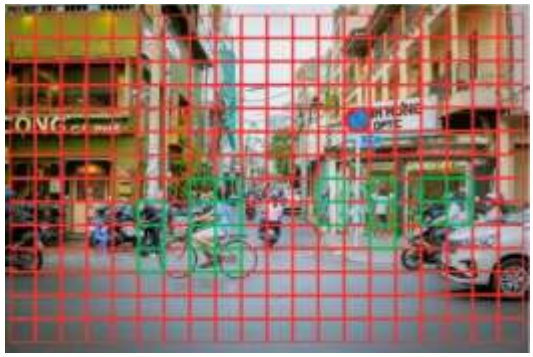


Fig. 1. The image captured by the camera is divided into a grid.

• Bounding Box Regression:

Fig. 2. The image is highlighted.

Figure 2 illustrates the concept of **"Bounding Box Regression"**, where predicted bounding boxes are adjusted to more precisely align with the actual location of the object in the image. The highlighted areas demonstrate how the model refines the bounding box coordinates—center (x, y), width, and height—enhancing object localization accuracy, a critical step in object detection frameworks like YOLO.

• Intersection Over Union (IOU): Measures bounding box overlap to ensure precise object detection, depicted in Figure 3 (Blue border for predicted image, green border for actual image).

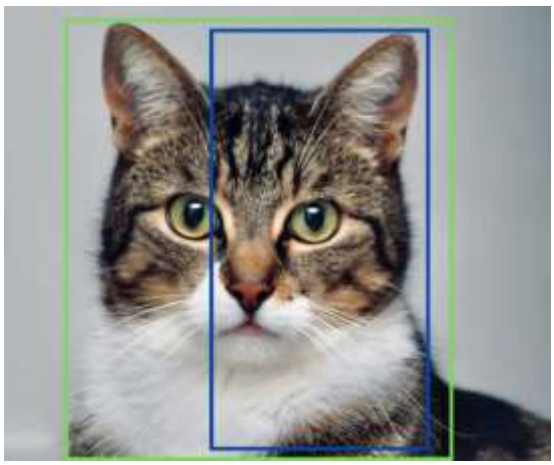


Fig. 3. Blue border (predicted image), green border (actual image).

Figure 3 illustrates the concept of IOU used in object detection. The image displays a cat's face with two overlapping bounding boxes: the blue border represents the predicted bounding box generated by the detection model, while the green border denotes the ground truth (actual) bounding box. The overlapping area between the two boxes visualizes how IOU is calculated—as the ratio of the intersection area to the union area. This metric plays a crucial role in evaluating and refining the accuracy of object localization in models such as YOLO.

B. Convolutional Neural Networks

CNNs are widely used methods for processing images and videos, particularly in computer vision tasks. As technology advances, CNNs are increasingly utilized in various aspects of daily life due to their fast and accurate recognition capabilities. They are particularly powerful in handling image processing tasks related to object identification and classification.

1) Convolution

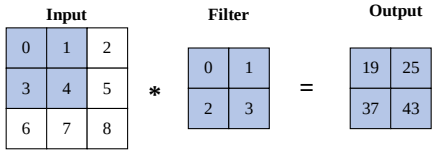


Fig. 4. Illustration of convolution.

Figure 4 demonstrates the convolution process in image processing. It shows an input matrix (4x4) on the left, labeled "Input," with values ranging from 0 to 8. A 2x2 filter matrix, labeled "Filter," with values [0, 1, 2, 3], is applied to the input. The convolution operation slides the filter over the input matrix, performing element-wise multiplication and summing the results to produce a 3x3 output matrix, labeled "Output," with values [19, 25, 37, 43]. This process extracts prominent features while discarding redundant details, as described in the convolution definition, highlighting its role in feature extraction using a kernel in convolutional neural networks (CNNs).

- **Kernel:** A small matrix that performs the convolution operation on the input image.
- **Padding:**

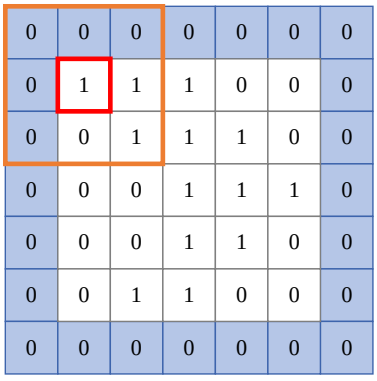


Fig. 5. Padding.

Figure 5 illustrates the concept of padding in image processing. The figure presents a 5×5 matrix, where the central 3×3 region—filled with values of 1—represents the original input data. The surrounding cells contain zeros, indicating the added padding. An orange border highlights a 4×4 region, and a red border outlines a 2×2 subsection within it, emphasizing how

padding extends the input matrix. This added padding (zero-padding in this case) is essential in convolutional neural networks (CNNs) to maintain spatial dimensions after convolution, ensuring that the output retains the same size as the input.

- **Stride:**

0	0	0	0	0	0	0
0	1	1	1	0	0	0
0	0	1	1	1	0	0
0	0	0	1	1	1	0
0	0	0	1	1	0	0
0	0	1	1	0	0	0
0	0	0	0	0	0	0

Fig. 6. Stride.

Figure 6 illustrates the concept of stride in convolution operations. The image presents a 7×7 matrix, where the central 5×5 yellow-highlighted region contains values of 1, representing the input data. The outer cells are padded with zeros. The stride is demonstrated by the movement of the convolutional kernel across the input matrix—specifically, shifting one cell at a time in this example. This movement pattern, or stride, directly influences the output size and computational efficiency of convolutional neural networks (CNNs), with larger strides resulting in reduced output dimensions and fewer computations.

2) Neural Network

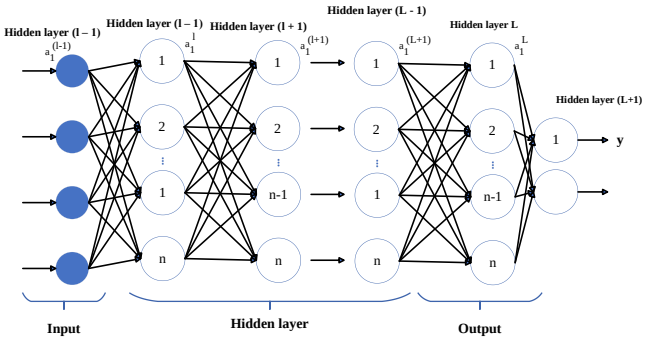


Fig. 7. Illustration of a neural network.

Figure 7 illustrates the architecture of a neural network, modeled after the structure of the human brain. The diagram begins with an input layer on the left, consisting of multiple nodes representing input features. These nodes are connected to a series of hidden layers labeled "Hidden Layer 1" through "Hidden Layer L," where each layer contains several interconnected neurons responsible for processing and transforming data. The connections between nodes signify the flow of information through the network. On the right, the output layer delivers the final classification result. This fully connected, layered design demonstrates how neural networks learn complex, non-linear patterns to perform classification tasks—functionally analogous to logistic regression but with greater capacity for abstraction and decision-making.

- *Convolutional Neural Network Architecture Model Architecture*

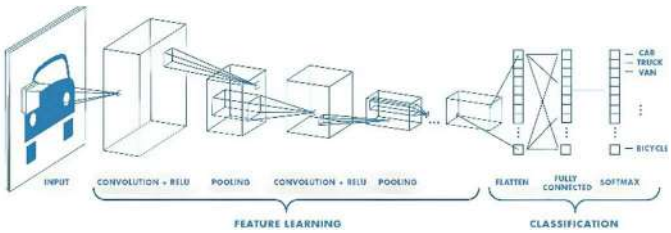


Fig. 8. Convolutional Neural Network Architecture.

Figure 8 illustrates the structure of a Convolutional Neural Network (CNN) designed for image classification. The diagram begins with an input layer on the left, represented by an image of a car, which feeds into the feature learning phase. This phase includes multiple stages of convolutional (CONV) layers, followed by ReLU activation layers to introduce non-linearity, and pooling (POOL) layers (either Max or Average pooling) to reduce the dimensionality of the input matrix. These layers extract and refine features from the input image. The process then transitions to the classification phase, where the data is flattened, passed through fully connected (FC) layers, and finally processed by a softmax layer to output probabilities for different classes (e.g., "Car," "Van," "Bicycle"). As described, the CNN comprises input, CONV, ReLU, POOL, and FC layers, showcasing the architecture's ability to perform feature extraction and classification.

III. System Analysis And Design

A. System Block Diagram

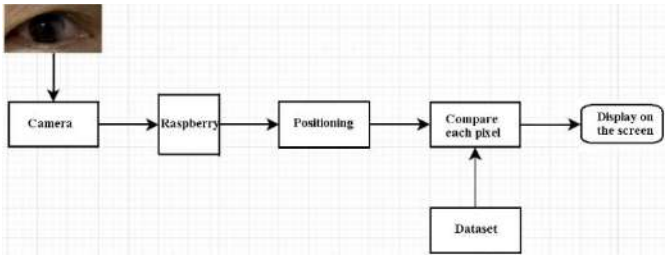


Fig. 9. System Block Diagram.

Figure 9 illustrates the functional block diagram of an eye disease detection system. The Camera Block captures images from the external environment, which are then sent to the Raspberry Block for initial processing and analysis. The Localization Block employs the YOLO V7 algorithm to precisely detect and adjust the pupil's position within the image. Subsequently, the Comparison Block uses convolution methods to enhance pixel clarity and compares these pixels with those from reference images in the Data File, which stores a dataset of various eye diseases. Finally, the Display Block visualizes the detection results on the screen.

B. Building a Predefined Recognition Dataset

1) Database Construction

- *System Recognition Model*

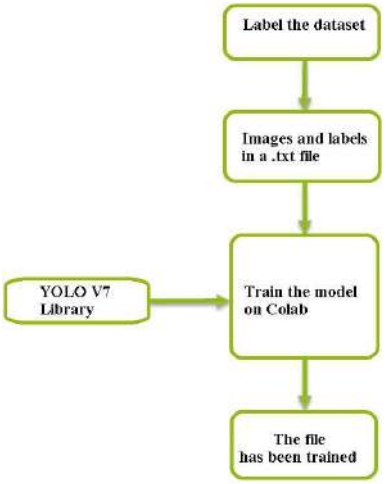


Fig. 10. System Recognition Model.

Figure 10 depicts the workflow for training a model to recognize eye diseases using the YOLO V7 framework. The process is presented as a flowchart comprising the following steps: First, the dataset is labeled by annotating the images to prepare them for training. These annotated images, along with their corresponding labels, are then saved in .txt file format. Next, the labeled dataset is uploaded and used to train the model on Google Colab, where the YOLO V7 library is integrated into the training pipeline. The final step confirms the successful completion of the training process, resulting in a trained model capable of identifying eye diseases. This flowchart provides a clear and systematic overview of the steps involved in preparing and training the recognition system.

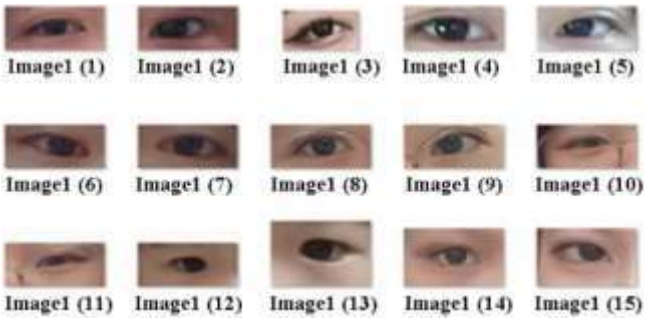


Fig. 11. Normal Eye Database.

Figure 11 displays a collection of 15 sample images of normal, healthy human eyes, used as reference data in the eye disease detection system. These images help the model distinguish between normal and abnormal eye conditions. The database captures diversity in eye appearance—such as variations in shape, eyelid position, and lighting conditions—enhancing the model's recognition accuracy and robustness.

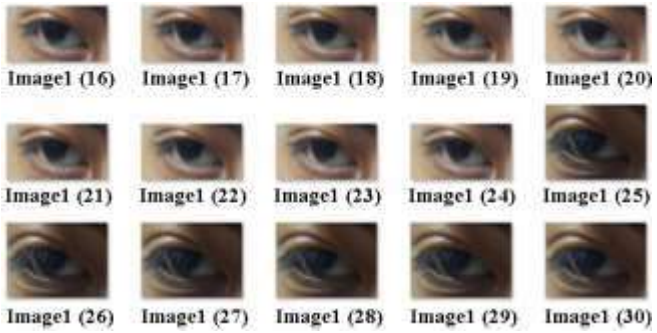


Fig. 12. Cross-Eye Database.

Figure 12 shows 15 sample images of individuals with strabismus (cross-eye condition), highlighting noticeable deviations in eye alignment that serve as key indicators for the detection model. The dataset includes images captured from various angles and under different lighting conditions, enhancing the model's ability to accurately detect and classify cross-eye symptoms across diverse real-world scenarios.

Fig. 13. Pterygium Eye Database.

Figure 13 presents 21 sample images from the pterygium eye database, showcasing eyes affected by pterygium—a benign growth of the conjunctiva that may extend onto the cornea. The images capture variations in the size, shape, and progression of the condition under different lighting conditions and perspectives. This diverse dataset enhances the robustness of the detection model in identifying and classifying pterygium across a wide range of real-world clinical scenarios.

The dataset was collected by taking pictures of patients' eyes at hospitals, as well as from relatives and friends. The database primarily consists of images of Vietnamese people's eyes. It targets various groups in society: patients, students, children, teachers, etc., and satisfies different conditions and angles of photography. The database includes: 500 images of normal eyes, 1000 images of eyes with pterygium, and 300 images of cross-eyes. All three datasets will be stored in a single file, and labels will be drawn to tag each type separately.

a) Labeling Data

Fig. 14. Label Drawing Web.

Figure 14 illustrates the interface of MakeSense, a widely used web-based image annotation tool designed for developing deep learning models. This open-source platform requires no complex installation—users can simply open it in their browser. MakeSense prioritizes data privacy by not uploading images to external servers, ensuring user data remains secure. It supports various annotation types, including rectangles, lines, points, and polygons, and provides flexible output formats such as YOLO, VOC XML, VGG JSON, and CSV. These features make it an efficient and reliable tool for preparing labeled datasets in AI training workflows.

Fig. 15. Label Drawing Interface.

Figure 15 shows the label drawing interface of the MakeSense annotation tool in use. The interface enables users to draw bounding boxes around regions of interest—in this case, the eye area in medical images. On the left panel, multiple sample images are displayed for batch annotation. The central panel provides zoom and alignment features for precision, while the right panel offers labeling options, including class selection and attribute editing. This streamlined interface enhances labeling efficiency and accuracy, which is crucial for training effective deep learning models for eye disease detection.

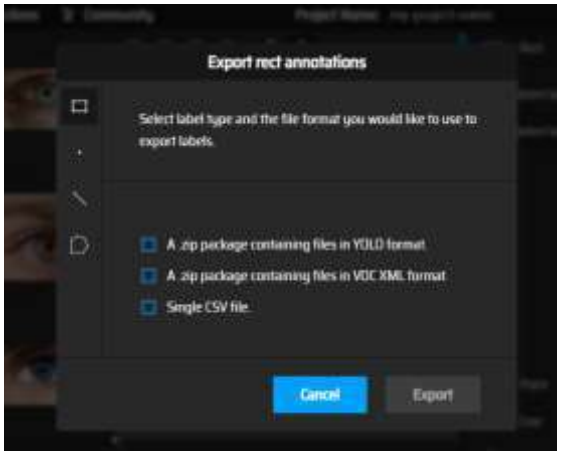


Fig. 16. File Export Interface.

Figure 16 illustrates the file export interface of the MakeSense annotation tool. After completing the labeling process, users are prompted to select the desired output format for the annotated data. Supported export options include YOLO format, VOC XML, and CSV file types, allowing seamless integration with various machine learning frameworks. This functionality facilitates the direct application of labeled datasets in model training, evaluation, and deployment phases, enhancing workflow efficiency in computer vision tasks such as eye disease classification.

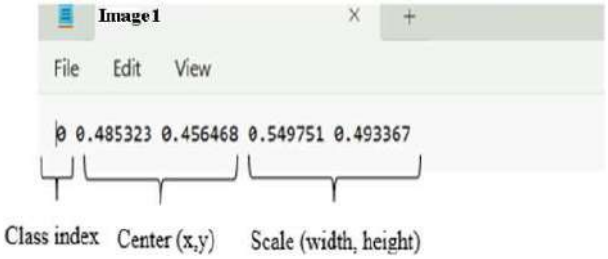


Fig. 17. Parameter Meanings.

Figure 17 displays the YOLO annotation format, where each line corresponds to a labeled object in an image. The structure consists of five key parameters:

- ✓ **Class index:** Denotes the object's category as an integer, indicating its position in the predefined class list.
- ✓ **Center (x, y):** Specifies the normalized coordinates of the bounding box's center, relative to the image dimensions.
- ✓ **Scale (width, height):** Represents the normalized width and height of the bounding box, providing the size of the labeled region.

This format is widely used in object detection models for its compactness and computational efficiency.

b) Training on Google Colab

Google Colab supports users in running simulations, training YOLO, and executing code in Python. This platform provides

an environment for users to analyze data, research AI, and develop Machine Learning models.

c) Post-training parameters

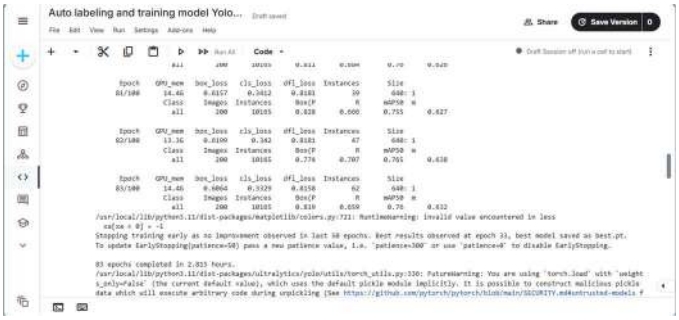


Fig. 18. Parameters After Training Completion.

Figure 18 displays the training log output from Google Colab during the training phase of a YOLO-based object detection model. The log includes metrics such as loss, objectness score, class accuracy, and more, with confidence levels ranging from 0.7 to 0.9, indicating high model reliability. This performance suggests the model has effectively learned to detect and classify objects in the dataset. Consequently, when deployed for recognition, the system is expected to provide fast inference speed and high accuracy, making it suitable for real-time applications.

d) Results of Disease Detection Using Images on COLAB

Fig. 19. Cross-Eye Detection.

Figure 19 shows a screenshot of a Google Colab interface demonstrating the detection of strabismus (cross-eye condition) using a YOLO-based model. The image displays a close-up of a person's eyes with yellow bounding boxes accurately highlighting the misaligned pupils, indicating the model's successful identification of the cross-eye condition. The interface includes a file explorer on the left, showing various dataset files, and a browser window displaying the image with detection results, captured at 02:28 PM +07 on Saturday, May 24, 2025. This visual evidence supports the model's ability to detect and localize eye abnormalities in real-time.

Fig. 20. Pterygium Detection.

Figure 20 displays a screenshot from a Google Colab interface, demonstrating pterygium detection using a YOLO-based model. The image features a close-up of a person's eyes, with pink bounding boxes marking pterygium—a conjunctival growth extending onto the cornea—in one eye. The interface shows a file explorer on the left with dataset files and a browser window with the detection results, captured at 02:34 PM +07 on Saturday, May 24, 2025. This output highlights the model's ability to accurately detect and localize pterygium in real-world eye images.

Fig. 21. Normal Eye or Other Disease Detection.

Figure 21 displays a screenshot from a Google Colab interface showcasing the detection of normal eyes or other eye conditions using a YOLO V7-based model. The image features a close-up of a person's eyes, with purple bounding boxes indicating the eyes are classified as normal or potentially exhibiting other diseases, distinct from cross-eye or pterygium. The interface includes a file explorer on the left, listing dataset files, and a browser window presenting the detection results, captured at 02:40 PM +07 on Saturday, May 24, 2025. After testing images with the newly trained database, the results demonstrate that YOLO V7 can accurately detect and distinguish between three eye types—cross-eye, pterygium, and normal or other diseases—with high accuracy. This output underscores the model's effectiveness in identifying normal eyes or other conditions in real-world scenarios.

2). Model Configuration

- Learning rate: This value determines the learning speed of the model. If the learning rate is too high, the model may easily get stuck in local minima and lose its learning ability. Conversely, if the learning rate is too low, the model will learn slowly and take more time.

- Batch size: This value affects the computation speed. If the batch size is too large, the model may require a lot of GPU memory and other hardware resources to process. Conversely, if

the batch size is too small, the model will learn slowly and inefficiently.

- Number of iterations: The more iterations, the better the model will be trained. However, it is important to avoid overfitting and outdated training.

- Objectness score threshold: This value determines the threshold for identifying an object in the image. If this value is too low, the model may make many errors. Conversely, if this value is too high, the model may miss many objects and not perform effectively.

- IOU threshold: This value determines the threshold for cutting objects in the image. If this value is too low, the model may have many overlapping objects. Conversely, if this value is too high, the model may miss many objects and not perform effectively.

C. Hardware design

1) Raspberry pi 4

Fig. 22. Raspberry pi 4.

Figure 22 displays the Raspberry Pi 4 single-board computer, a compact and versatile device renowned for its powerful performance and diverse applications. It features a Broadcom BCM2711 processor, RAM options from 2GB to 8GB, USB 3.0 and USB 2.0 ports, Gigabit Ethernet, and dual micro-HDMI outputs supporting high-resolution displays. Additionally, it offers built-in Wi-Fi and Bluetooth for seamless wireless connectivity. Thanks to its processing power and compatibility with various operating systems, particularly Raspberry Pi OS, the Raspberry Pi 4 is widely utilized in artificial intelligence, image processing, embedded systems, and IoT. In this project, it functions as the central processing unit for the eye recognition and disease detection system.

2) 32GB Memory Card

Fig. 23. 32GB Memory Card.

3) Webcam dahua z2 plus 1080p

Fig. 24. Webcam dahua z2 plus 1080p.

D. System model

Fig. 25. System Model Completion.

Figure 25 illustrates the fully completed eye disease detection system model. The system operates stably and demonstrates high detection reliability, achieving confidence scores ranging from 0.7 to 0.9. It is capable of accurately identifying normal eyes, strabismus (cross-eye condition), and pterygium. The integration of components such as image acquisition, processing, deep learning-based classification, and result display ensures that the system functions effectively in real-world applications, delivering both speed and accuracy in medical diagnostics.

IV. Experimental Results

Fig. 26. Experimental Results (Detection of Normal Eyes or Other Diseases).

Figure 26 presents the experimental results of the trained model in detecting normal eyes or other non-targeted eye conditions. The image output features a bounding box with an associated confidence score, showcasing the system's ability to accurately identify cases without strabismus or pterygium symptoms. These results demonstrate the model's robustness in distinguishing healthy eyes from various unclassified abnormalities, ensuring reliable classification across diverse inputs.

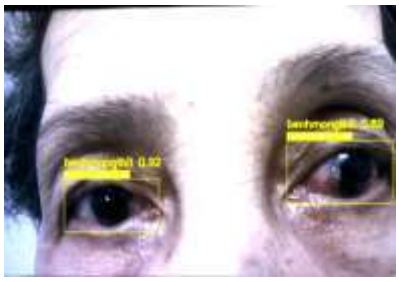


Fig. 27. Experimental Results (Detection of Pterygium).

Figure 27 displays the output of the trained YOLO-based model in detecting pterygium, a fibrovascular growth extending from the conjunctiva onto the cornea. The image features a bounding box highlighting the affected eye region, accompanied

by a confidence score reflecting the model's certainty. These results highlight the system's effectiveness in accurately identifying pterygium, even under varied lighting and image conditions, confirming its practical applicability for real-world eye disease screening.

Fig. 28. Experimental Results (Detection of Cross-Eye).

Figure 28 presents the outcome of the trained YOLO-based model in detecting cross-eye (strabismus). The image features a bounding box surrounding the misaligned pupils, clearly indicating the presence of strabismus. The model provides a high confidence score for the detection, demonstrating its ability to accurately identify and localize eye misalignment. This result underscores the system's potential to support early diagnosis of cross-eye conditions through automated image analysis.

V. Conclusion

This study explores the recognition and diagnosis of eye diseases. The Raspberry Pi and webcam function reliably when powered. Once the system is embedded onto the Raspberry Pi, it can detect human eyes and differentiate between normal eyes (or other conditions), cross-eye, and pterygium. The detection speed is swift in a single scan, and the model's reliability is high (0.7 – 0.9). Future work involves gathering more databases to identify a broader range of eye diseases, enhancing the model's accuracy, improving the system's performance under various lighting conditions, integrating additional sensors for more comprehensive diagnostics, and developing a user-friendly interface for easier operation and interpretation of results.

Data Availability

The data used to support the findings of this study are included in the paper.

Conflicts of Interest

The authors declare there is no conflict of interest in this manuscript.

Acknowledgment

Huu Q. Tran acknowledges the support of time and facilities from Industrial University of Ho Chi Minh City (IUH) for this study.

References

[1] Du, Fanyu & Zhao, Lishuai & Luo, Hui & Xing, Qijia & Wu, Jun & Zhu, Yuanzhong & Xu, Wansong & He, Wenjing & Wu, Jianfang. (2024). Recognition of eye diseases based on deep neural networks for transfer learning and improved D-S evidence theory. BMC Medical Imaging. 24. 10.1186/s12880-023-01176-2.

[2] Nuzzi, Raffaele & Boscia, Giacomo & Marolo, Paola & Ricardi, Federico. (2021). The Impact of Artificial Intelligence and Deep Learning in Eye Diseases: A Review. Frontiers in Medicine. 8. 10.3389/fmed.2021.710329.

[3] Bali, Akanksha & Mansotra, Vibhakar. (2023). Analysis of Deep Learning Techniques for Prediction of Eye Diseases: A Systematic Review. Archives of Computational Methods in Engineering. 31. 10.1007/s11831-023-09989-8.

[4] Alayón, Silvia & Hernández, Jorge & Fumero, Francisco & Sigut, Jose & Díaz-Alemán, Tinguaro. (2023). Comparison of the Performance of Convolutional Neural Networks and Vision Transformer-Based Systems for Automated Glaucoma Detection with Eye Fundus Images. Applied Sciences. 13. 12722. 10.3390/app132312722.

[5] Juneja, Mamta & Singh, Shaswat & Agarwal, Naman & Bali, Shivank & Gupta, Shubham & Thakur, Niharika & Jindal, Prashant. (2020). Automated detection of Glaucoma using deep learning convolution network (G-net). Multimedia Tools and Applications. 79. 10.1007/s11042-019-7460-4.

[6] Khalaf, Noor & Najim, Mohammed & Cankaya, I.. (2024). Simplified Convolutional Neural Network Model for Automatic Classification of Retinal Diseases from Optical Coherence Tomography Images. Al-Nahrain Journal for Engineering Sciences. 26. 314-319. 10.29194/NJES.26040314.

[7] Ting, Daniel & Pasquale, Louis & Peng, Lily & Campbell, John & Lee, Aaron & Raman, Rajiv & Tan, Gavin & Schmetterer, Leopold & Keane, Pearse & Wong, Tien Yin. (2018). Artificial intelligence and deep learning in ophthalmology. British Journal of Ophthalmology. 103. bjophthalmol-2018. 10.1136/bjophthalmol-2018-313173.

[8] Moraru, Andreea & Costin, Danut & Moraru, Radu & Brănișteanu, Daniel. (2020). Artificial intelligence and deep learning in ophthalmology - present and future (Review). Experimental and Therapeutic Medicine. 20. 10.3892/etm.2020.9118. Penteado, R. C., et al. "Glaucoma Detection and Analysis Using Deep Learning and Portable Imaging Devices." Springer Lecture Notes in Computer Science (LNCS), vol. 12261, 2020, pp. 93-104, doi:10.1007/978-3-030-58621-8_9.

[9] Ueno, Yuta & Oda, Masahiro & Yamaguchi, Takefumi & Hideki, Fukuoka & Nejima, Ryohei & Kitaguchi, Yoshiyuki & Miyake, Masahiro & Akiyama, Masato & Miyata, Kazunori & Kashiwagi, Kenji & Maeda, Naoyuki & Shimazaki, Jun & Noma, Hisashi & Mori, Kensaku & Oshika, Tetsuro. (2024). Deep learning model for extensive smartphone-based diagnosis and triage of cataracts and multiple corneal diseases. British Journal of Ophthalmology. 108. bjo-2023. 10.1136/bjo-2023-324488.

[10] Quellec, Gwenole & Hajj, Hassan & Lamard, Mathieu & Conze, Pierre-Henri & Massin, Pascale & Cochener, Béatrice. (2021). ExplAIn: Explanatory Artificial Intelligence for Diabetic Retinopathy Diagnosis. Medical Image Analysis. 72. 102118. 10.1016/j.media.2021.102118.

[11] Shojol, Md. Shojeb & Siddique, Md & Haque, Fariha. (2023). Enhanced Convolutional Neural Networks for Early Detection and Classification of Ophthalmic Diseases. 209-213. 10.1109/ICICT4SD59951.2023.10303558.

[12] Tong, Yan & Lu, Wei & Yu, Yue & Shen, Yin. (2020). Application of machine learning in ophthalmic imaging modalities. Eye and Vision. 7. 10.1186/s40662-020-00183-6.

[13] Rajpurkar, P., Chen, E., Banerjee, O. et al. AI in health and medicine. Nat Med 28, 31–38 (2022). https://doi.org/10.1038/s41591-021-01614-0.

Implementation and Verification of the Improved SpaceCAN

Men-Shen Tsai
Graduate Institute of Automation Tech. Research Center of Energy Conservation for New Generation of Residential Commercial, and Industrial Sectors, National Taipei University of Technology,
Taipei, Taiwan
mstsai@mail.ntut.edu.tw

Ya-Wen Wu*
Graduate Institute of Automation Tech.
National Taipei University of Technology,
Taipei, Taiwan
t112618006@ntut.edu.tw

***Abstract*—As CubeSat missions become more diverse and complex, the communication demands on the On-Board Computer (OBC) under limited resources are also increasing. An OBC system architecture based on the CANopen protocol was proposed in this study, improving upon the existing SpaceCAN design to achieve stable and efficient communication among subsystems. By extending the CANopen protocol, this research designs each subsystem to have multiple Process Data Objects (PDOs) for data transmission, rather than relying on a single PDO to transmit all data. This design reduces the software execution burden, and enhances data transmission efficiency and priority control capability. The verification results show that this design effectively addresses data congestion issues, improving the real-time and reliability of CubeSat OBC applications.**

Keywords—CANopen, SpaceCAN, CubeSat, OBC.

I. INTRODUCTION

The concept of CubeSat was originated in 1999 by Prof. Jordi Puig-Suari[1] and Prof. Bob Twiggs[2]. Their standardized size and modular design quickly made them essential tools for space exploration. In the system architecture of CubeSat, OBC is the command and data handling subsystem, responsible for executing control commands from the ground station, monitoring the operation of each subsystem, processing data from various subsystems, and managing the satellite's mission plans. Due to the small size of CubeSats and limited space resources, the OBC design must achieve efficient mission processing and reliable communication capabilities with limited hardware resources, ensuring real-time performance and accuracy during the satellite's operation in orbit.

Currently, various communication technologies are used in CubeSat OBC typically to implement the information transmission between different subsystems, including I2C, SPI, and UART[3][4][5]. However, with the increase in mission complexity, these communication technologies tend to be insufficient in terms of parallel task processing, real-time performance, and scalability. The Controller Area Network (CAN), which supports multi-node communication and has the features of high reliability, real-time performance, and strong anti-interference capabilities, is gradually becoming an ideal choice for OBC communication[6]. However, the CAN protocol only defines the physical layer and data link layer in the Open System Interconnection (OSI) model, without specifying the application layer. This limits its ability to effectively address certain specialized application challenges, such as the incompatibility issues when products are designed by different manufacturers, leading to interoperability problems.

CANopen is a high-level application layer protocol of CAN, designed for multi-node distributed systems. It provides standardized device configuration and communication services, enhancing interoperability between different systems and devices. With built-in node management, error handling mechanisms, and real-time support, CANopen is particularly suitable for CubeSat applications, which integrate multiple subsystems. SpaceCAN[7] is a similar design. But, in SpaceCAN's design, although each subsystem has its own device node ID, all data from the same node is transmitted using the same PDO, and the data is transmitted independently, relying on program definitions to distinguish the content of the messages. While this design simplifies the management and configuration of PDOs, it also increases the software design burden, requiring additional definitions to parse different data formats and content within the program. When data transmission is frequent and varied, using the same PDO may lead to data congestion, affecting transmission efficiency, and it also lacks the ability to prioritize data based on its importance.

To address the aforementioned issues, this study proposes a design where each subsystem uses not just one PDO for data transmission, rather each data has its own PDO. Additionally, smaller data are combined into a single PDO message to reduce the number of PDOs and avoid bandwidth limitations. The remaining structure of this paper is as follows: Section II covers the architecture of CubeSat system and explaining the functions of the OBC. Section III discusses SpaceCAN communication, analyzing the design of SpaceCAN and exploring its advantages and disadvantages, with improvements suggested for the drawbacks. Section IV implements the improved SpaceCAN, explaining the design of the system's hardware and software in detail. Section V verifies the improved SpaceCAN by testing the communication capabilities of the system. Finally, Section VI presents the conclusion, summarizing the outcomes of this study and discussing future developments.

979-8-3315-9431-2/25/$31.00 ©2025 IEEE

II. ARCHITECTURE OF CUBESAT SYSTEM

CubeSat is a type of small satellite built on the concept of standardization, low cost, and ease of design. Its smallest basic unit is 1U, which represents a mass of no more than 1.33 kg and dimensions of a 10 cm cube. Depending on the mission requirements, there are six different specifications (Fig.1). Because the system architecture adopts a modular design, different functions can be allocated to individual subsystems, which are then integrated to form the satellite. A typical CubeSat system architecture(Fig.2) includes the following main subsystems:

- Electrical Power Subsystem (EPS): Responsible for providing and managing the satellite's electrical power, including power generation, storage, and distribution.
- Attitude Determination and Control Subsystem (ADCS): Controls and adjusts the satellite's attitude and orientation to ensure that the satellite can stably move towards with its intended target.
- Telemetry Tracking and Communication Subsystem (TT&C): Responsible for communication between the satellite and the ground station, usually supporting UHF, VHF, or S-band communications.
- Payload: The core part of the satellite's mission, equipped with scientific instruments or technical equipment based on specific mission requirements, such as cameras, sensors, experimental modules, etc.
- Command and Data Handling Subsystem (C&DH): The satellite's central control unit, responsible for managing subsystems, processing data, and handling the transmission and reception of commands. It typically integrates microprocessors, memory, and interface modules.
- Structure and Mechanism Subsystem: Provides structural support for the satellite, protecting internal components from the harsh space environment. It is usually made of aluminum alloys or composite materials.
- Thermal Control Subsystem: Manages the internal temperature of the satellite to ensure that all components operate within their designed temperature ranges.

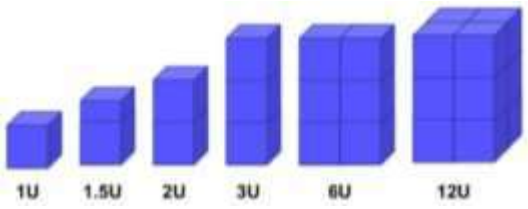


Fig.1 The six common specifications of CubeSat[8].

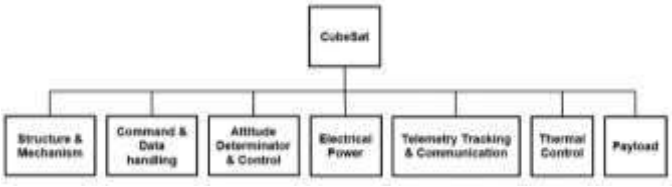


Fig.2 The architecture of CubeSat subsystem.

III. SPACECAN COMMUNICATION

SpaceCAN is a communication protocol specifically designed for space missions. It leverages the physical and data link layers of the CAN protocol to support reliable communication between multiple nodes while offering strong resistance to interference, making it ideal for use in the complex electromagnetic environments of space. By integrating the CANopen application protocol, SpaceCAN standardizes communication methods and device node management, ensuring interoperability and stability in data exchange between various subsystems.

A. CANopen Communication Protocol

CANopen is defined by CAN-in Automation (CiA). It is primarily used for network control applications in distributed control and embedded systems, allowing standardized communication between different systems, and enabling interoperability between various CAN devices.

In a CANopen network system, each device has its own unique object dictionary, which consists of a series of objects that define all types of object attributes, and describe the device and its network behavior. Each object has a unique 16-bit index and an 8-bit subindex. The index can be divided into several sections as shown in 0.

TABLE I. OBJECT DICTIONARY SECTION TABLE

Index	Section definition
0x0000~0x01FF	Communication specifications reserved for CANopen.
0x1000~0x1FFF	Standardized equipment parameters.
0x2000~0x5FFF	The application parameters are defined by the device manufacturer and configured according to the device capabilities.
0x6000~0x9FFF	Variables that can be mapped to PDOs, typically data during real-time runs.
0xA000~0xFFFF	Reserved area for manufacturer free use.

CANopen network communication is primarily consisting of four communication objects: Service Data Object (SDO), Process Data Object (PDO), Network Management (NMT), and special function objects. Special function objects include SYNC, TIME STAMP, EMCY, Heartbeat, and Boot-up. Among these, SDO accesses the values of a device's object dictionary via index and subindex, requires a response, and is mainly used for transmitting parameters during device configuration. PDO is used for transmitting real-time data, with a maximum size of 8 bytes, and does not require a response. NMT is used to manage device states, control devices, and promptly detect device failures.

B. The Traditional SpaceCAN Design

The design of SpaceCAN mainly utilizes CANopen's PDO, NMT, and some special function objects. Each subsystem is assigned a unique node ID, and the nodes are managed using CANopen's NMT service to start, stop, and reset them, facilitating system configuration and maintenance. For error detection and handling, the system leverages CAN protocol's error reporting and CANopen's EMCY to manage fault states, ensuring a quick response and appropriate handling when errors occur.

Real-time data transmission is carried out through PDOs and supports event-driven communication. PDOs are only triggered for transmission when data changes or specific events occur, allowing the system to update data based on actual needs, and avoiding unnecessary data traffic. In this design, all data from a single node is transmitted using the same PDO, simplifying PDO management and bus configuration, and maintaining bus communication clarity. However, when data transmission is frequent, using only one PDO for transmission may lead to congestion, and the application layer has to parse the format and content of each piece of data, increasing the complexity of application layer software design, especially when handling multiple data types and priorities. Considering the above issues, the contributions of this research are as follows:

1) Allocating a subsystem's data to multiple PDOs for transmission helps to avoid data congestion. Additionally, combining multiple small data points into a single PDO reduces the number of PDOs, thereby lowering the bandwidth burden.

2) The design of multiple PDOs can be configured using different Communication Object ID (COB-ID), mapping more important data to PDOs with smaller COB-ID for priority transmission, thus achieving data transmission prioritization.

IV. The Implementation of Improved SpaceCAN

A. Hardware Design

In this study, the STM32L432KC microcontroller is selected as the OBC, due to its ARM Cortex-M4 core, which operates at a frequency of up to 80MHz, ensuring high performance when handling the complex CANopen protocol. Additionally, its low-power mode extends the CubeSat's battery life. CAN communication requires both a CAN controller and a CAN transceiver. The STM32L432KC has integrated CAN controller, simplifying this part of the circuit design. The CAN transceiver chosen is the SN65HVD230, which complies with the ISO11898 standard, offering good noise immunity and overheating protection while consuming low power, making it suitable for CubeSat's energy-constrained and compact design. The hardware architecture is shown in Fig.3, where two STM32L432KC microcontrollers are used, with one acting as the OBC and the other as a CubeSat subsystem.

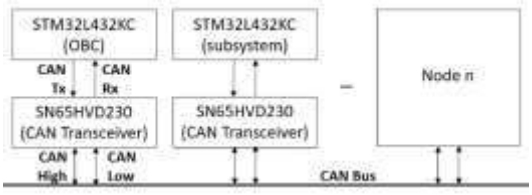


Fig.3 The implementation of hardware architecture.

B. Software Design

This paper configures the pins of STM32L432KC using STM32CubeIDE, a tool that provides an extensive libraries and sample programs. Initialize the CAN communication interface, set the system clock to 8MHz, and configure the bit timing as shown in Fig.4. The timer is set to enable a 1ms overflow interrupt timer, and the interrupt is activated. Then, the initialization code was generated.

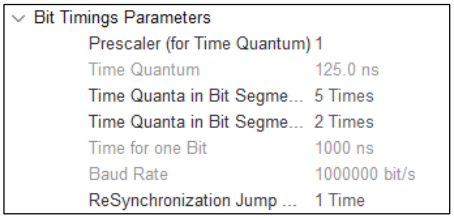


Fig.4 STM32CubeIDE bit time configuration.

Considering cost, scalability, and flexibility, the free and open-source CANopenNode protocol stack[9] was selected in this paper, which can efficiently run on resource-constrained systems. It is easy to extend the object dictionary and add custom application layer functions, and it supports multiple STM32 series microcontrollers.

The Electronic Data Sheet (EDS) is a document that describes the content of a CANopen node's object dictionary, defining the node's functions and data structure according to a standard syntax. It provides the basis for interoperability between devices, ensuring that equipment from different manufacturers can understand each other and use the node's functions. The object dictionary is an internal data structure of the device that implements the device's parameters and data, while the EDS is an external description file of the object dictionary, detailing these data structures. EDSEditor is a tool specifically designed to edit and manage CANopen EDS, helping to generate, modify, and validate EDS files to ensure that the object dictionary of a CANopen node complies with CANopen specifications. In this study, libedssharp[10], developed by Robin Cornelius and written in C sharp, is used as the EDSEditor, providing an easy-to-use graphical interface.

In this design, one of the STM32L432KCs is configured as the OBC, with its EDS edited to enable synchronous messages and define the synchronization interval. It sends a sync message to the CAN bus at regular intervals, which other subsystems use for time synchronization and to trigger specific operations. Additionally, the OBC needs to monitor the operating status of each subsystem by defining a heartbeat consumer and specifying the time within which it should receive a heartbeat from a certain node. If no heartbeat is received, an EMCY message is sent. Each subsystem is configured to send a heartbeat at regular intervals, informing the OBC of its operational status.

Each subsystem should send its health status or sensor data periodically, so it is configured to transmit synchronous Transmit PDO (TPDO), which is received by the OBC for monitoring. The OBC is configured with event-triggered Receive PDO (RPDO), which stores mapped data in the corresponding mapping object when a change in the received PDO is detected, and then sends the RPDO at the next synchronization.

To simulate the transmission of control commands from a ground station to the CubeSat, the OBC controls each subsystem by triggering an event that activates TPDO. The event timer is set to zero, meaning the TPDO is triggered by the program only once to simulate a control command situation. The subsystem is also configured with event-triggered RPDO, which only receives data when the mapped value in the TPDO changes.

The object dictionary and program are designed based on the hypothetical scenarios described above, and four PDOs are configured. Each PDO maps multiple objects to simulate the scenario of multiple PDO transmissions. The PDO configuration of the OBC is shown in 0, and the PDO configuration of the subsystem is shown in 0. By observing the message transmission, the feasibility of the design can be validated.

TABLE II. THE PDO CONFIGURATION OF OBC

PDO	COB-ID	Trans type	Mapping parameter
TPDO1	195h	254	60000020h/60030108h/60030208h/ 60030308h/60030408h
TPDO2	295h	254	60030510h/60030620h/ 60030708h/60030808h
TPDO3	395h	254	60080140h
TPDO4	495h	254	60080208h/60080310h/60080420h
RPDO1	194h	0	60010020h/60020020h
RPDO2	294h	0	60040020h
RPDO3	394h	0	60050020h
RPDO4	494h	0	60060020h/60070010h/60080508h

TABLE III. THE PDO CONFIGURATION OF SUBSYSTEM

PDO	COB-ID	Trans type	Mapping parameter
TPDO1	194h	3	60000020h/60020108h/60020208h/ 60020308h/60020408h
TPDO2	294h	1	60040820h
TPDO3	394h	2	60040920h
TPDO4	494h	2	60040A20h/60040B10h/60040C08h
RPDO1	195h	254	60010020h/60030020h
RPDO2	295h	254	60040110h/60040220h/60040310h
RPDO3	395h	254	60040440h
RPDO4	495h	254	60040508h/60040610h/60040720h

V. THE VERIFICATION OF THE IMPROVED SPACECAN

The traditional SpaceCAN design process, as shown in Fig.5, uses a single PDO to transmit each piece of data. This approach requires writing code to parse the total length and service type of the data, leading to higher programming complexity. In contrast, the improved SpaceCAN design process proposed in this study, as shown in Fig.6, significantly simplifies programming by allocating data to multiple PDOs for transmission. This not only effectively enhances bandwidth utilization but also allows priority configuration via COB-ID, ensuring that critical data can be transmitted with higher priority.

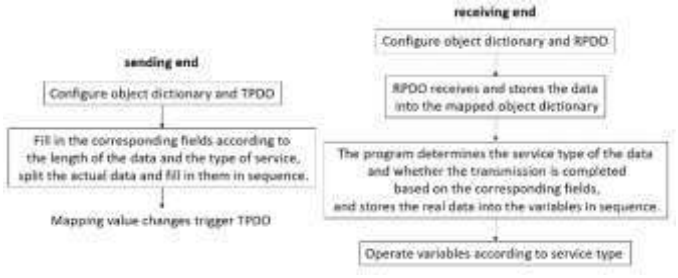


Fig.5 The design process diagram of traditional SpaceCAN.

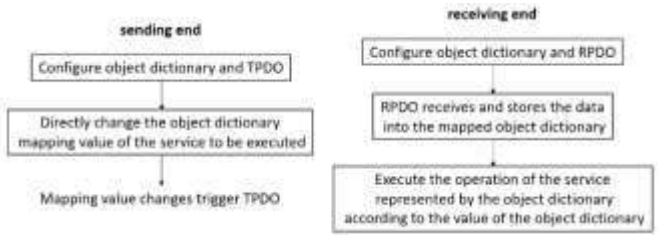


Fig.6 The design process diagram of the improved SpaceCAN.

To validate the study, a Raspberry Pi 3 Model B++ was connected to a 2-CH CAN HAT module, enabling CAN communication. The operating system was Raspbian, and the CANopenLinux protocol stack was implemented, allowing it to join the CAN network and receive messages. The hardware architecture is shown in Fig.7. Additionally, the tool candump, from the can-utils package in Linux, was used to display CAN messages. This command was used to observe communication between STM32 devices, verifying the feasibility of this design.

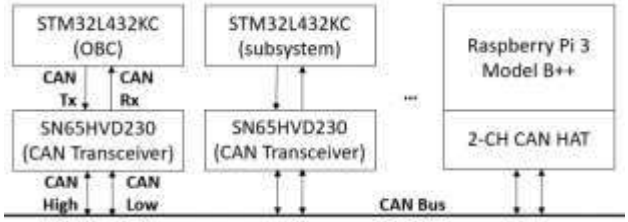


Fig.7 The hardware architecture diagram of verified.

A. *The Object Dictionary of the Device Configured by SDO*

The OBC device node 21 (0x15) is programmed to write into the object dictionary of the subsystem node 20 (0x14), modifying the value at index 0x6000 to 0xAB00CDEF. Additionally, the value at subindex 8 of index 0x6004 is changed to 0x12345678, the value at subindex 9 is changed to 0x13572468, the value at subindex 11 is changed to 0x9999, and the value at subindex 12 is changed to 0x88. As shown in Fig.8, the observed result met expectations.

```
can0  714   [1]  00
can0  701   [1]  7F
can0  701   [1]  7F
can0  614   [8]  23 00 60 00 EF CD 00 AB
can0  594   [8]  60 00 60 00 00 00 00 00
can0  614   [8]  23 04 60 08 78 56 34 12
can0  594   [8]  60 04 60 08 00 00 00 00
can0  614   [8]  23 04 60 09 68 24 57 13
can0  594   [8]  60 04 60 09 00 00 00 00
can0  614   [8]  2B 04 60 0B 99 99 00 00
can0  594   [8]  60 04 60 0B 00 00 00 00
can0  614   [8]  2F 04 60 0C 88 00 00 00
can0  594   [8]  60 04 60 0C 00 00 00 00
can0  715   [1]  00
```

Fig.8 The message of Boot-up & SDO.

B. *The Real-Time Data Transferred from PDO*

The OBC object dictionary configures the TPDO as asynchronous, triggered by an event in the program, simulating the OBC receiving a command from the ground station to communicate with subsystems in real time to retrieve data or execute control actions. The result observed in candump is shown in Fig.9, where the configuration matches the values

modified, and the PDO with a smaller COB-ID is transmitted with higher priority.

```
can0  195   [8]  34 12 34 12 00 00 00 00
can0  295   [8]  68 24 00 00 00 00 00 00
can0  395   [8]  21 43 65 87 78 56 34 12
can0  495   [7]  00 57 13 00 00 00 00
```

Fig.9 PDO message.

The subsystem object dictionary configures TPDO to be triggered synchronously. TPDO1 is set to transmit once every 3 SYNC signals and TPDO2 is set to transmit once per SYNC signal, while TPDO3 and TPDO4 are set to transmit once every 2 SYNC signals. The results observed in candump are shown in Fig.10, Fig.11, and Fig.12, respectively. The mapping objects were modified via the OBC program, and the content matches the updates.

```
can0  080   [0]
can0  294   [4]  78 56 34 12
```

Fig.10 TPDO2 message.

```
can0  080   [0]
can0  294   [4]  78 56 34 12
can0  394   [4]  68 24 57 13
can0  494   [7]  00 00 00 00 99 99 88
```

Fig.11 TPDO3&TPDO4 message.

```
can0  080   [0]
can0  194   [8]  EF CD 00 AB 00 00 00 00
can0  294   [4]  78 56 34 12
```

Fig.12 TPDO1 message.

C. *The Object of Heartbeat and Synchronisation*

The OBC object dictionary is configured to transmit SYNC every 1,000,000 μs, and the subsystem is configured to send a Heartbeat every 1000 ms. The results observed in candump are shown in Fig.13.

```
can0  714   [1]  05
can0  080   [0]
```

Fig.13 SYNC of OBC & Heartbeat of subsystem.

D. *Life-guarding*

The OBC object dictionary configures a heartbeat consumer, and if no heartbeat message from the subsystem is received within 1300ms, an EMCY indicating a heartbeat timeout is sent. The program is set to send an NMT control command to reset the node when a heartbeat timeout occurs. In this case, the SDO is configured to write and disable the heartbeat to simulate a timeout scenario. As shown in Fig.14, after executing the candump command, the observed results met expectations.

```
can0  000   [2]  81 14
can0  095   [8]  30 81 10 1B 00 00 00 00
can0  714   [1]  00
can0  714   [1]  05
can0  095   [8]  00 00 00 1B 00 00 00 00
```

Fig.14 EMCY message & NMT control message.

VI. CONCLUSION

This study focuses on improving the SpaceCAN design for the OBC system of a CubeSat. To address the data congestion and transmission efficiency issues in the traditional SpaceCAN design, this research proposes using multiple PDOs to transmit different types of data, and combining small data into a single PDO message to reduce bandwidth usage and enhance system transmission efficiency. Additionally, based on the COB-ID priority mechanism of the CANopen protocol, it ensures the real-time transmission of critical data, addressing the problem of data priority differentiation.

The experimental results show that the improved system effectively resolves data congestion issues, enabling the communication of multiple data in a very short time, significantly improving the real-time performance and stability of data transmission. This design is suitable for resource-constrained CubeSat systems and offers good scalability, making it applicable to other small satellites and distributed control systems in the future.

ACKNOWLEDGMENT

The authors acknowledge the financial support by the "Research Center of Energy Conservation for New Generation of Residential, Commercial, and Industrial Sectors" from The Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan under contract No. T7141101-19.

REFERENCES

[1] J. Puig-Suari, C. Turner and W. Ahlgren, "Development of the standard CubeSat deployer and a CubeSat class PicoSatellite," 2001 IEEE Aerospace Conference Proceedings (Cat. No.01TH8542), Big Sky, MT, USA, 2001, pp. 1/347-1/353 vol.1, doi: 10.1109/AERO.2001.931726.

[2] B. Twiggs and S. Kuroki, "BioExplorer bus - Low cost approach [satellite design]," Proceedings, IEEE Aerospace Conference, Big Sky, MT, USA, 2002, pp. 1-1, doi: 10.1109/AERO.2002.1036862.

[3] Tzu-Ya Tai （2021）. Flight Software and Firmware Design of IDEASSat/INSPIRESSat-2. National Digital Library of Theses and Dissertations in Taiwan. https://hdl.handle.net/11296/pxk6mk

[4] Chang-Hung Tsai （2023）. Development of Flight Software for a Cubesat. National Digital Library of Theses and Dissertations in Taiwan. https://hdl.handle.net/11296/26buhx

[5] Laizans, K., Sünter, I., Zalite, K., et al. (2014). Design of the fault tolerant command and data handling subsystem for ESTCube-1. Proceedings of the Estonian Academy of Sciences, 63(2S), 222-231.

[6] Sajjad, W., Shafique, A., & Mahmood, R. (2023). Designing of Reliable, Low-Power, and Performance-Efficient Onboard Computer Architecture for CubeSats. IEEE Journal on Miniaturization for Air and Space Systems, 5(2), 59-72.

[7] SpaceCAN. LibreCube Documentation. https://librecube.gitlab.io/standards/spacecan/

[8] Sonja Caldwell (2024, August 15).What Are SmallSats and CubeSats. NASA. https://www.nasa.gov/what-are-smallsats-and-cubesats/

[9] CANopenNode. (2024, August 14). GitHub. https://github.com/CANopenNode/CANopenNode/tree/master

[10] Robin Cornelius (2023, March 2). Libedssharp. GitHub. https://github.com/robincornelius/libedssharp

Latency Optimization in Clustering NOMA-Aided Cell-Free Massive MIMO with Mobile Edge Computing

Tien V. Thai
Faculty of Electronics and Telecommunication Engineering, The University of Danang, University of Sci. and Tech.
Da Nang, Vietnam
thaivantien@dut.udn.vn

Mai T. P. Le
Faculty of Electronics and Telecommunication Engineering, The University of Danang, University of Sci. and Tech.
Da Nang, Vietnam
lpmai@dut.udn.vn

Hieu V. Nguyen
Faculty of Electronics and Telecommunication Engineering, The University of Danang, University of Sci. and Tech.
Da Nang, Vietnam
nvhieu@dut.udn.vn

Huu Q. Tran
Faculty of Electronics Technology, Industrial University of Ho Chi Minh City,
Ho Chi Minh, Vietnam
tranquyhuu@iuh.edu.vn

***Abstract*—Reducing network latency while maintaining energy efficiency is a key challenge in next-generation wireless networks, particularly in a clustering NOMA-aided cell-free massive multiple-input multiple-output (CF-mMIMO) system with mobile edge computing (MEC). Unlike conventional pairing-based NOMA, we propose a novel clustering NOMA approach using K-means algorithm, which efficiently groups user equipments (UEs) into clusters based on their channel conditions and proximity to access points (APs). The objective is to optimize power allocation and offloading ratios while minimizing the overall system latency and maintaining energy efficiency. To tackle this non-convex optimization problem, we introduce a two-step solution framework, where the first step applies K-means clustering for optimal UE grouping, and the second step employs successive convex approximation (SCA) to jointly optimize power allocation and offloading decisions. Simulation results demonstrate that our proposed clustering-based NOMA approach for MEC-enabled CF-mMIMO networks achieves faster convergence, lower latency, and higher energy efficiency compared to traditional pairing-based NOMA, CF-mMIMO, and conventional massive MIMO schemes.**

***Index Terms*—Cell-free massive MIMO, clustering NOMA, K-means, mobile edge computing (MEC), latency optimization, and successive convex approximation (SCA).**

I. Introduction

The explosive growth of data-intensive applications in 5G and beyond networks has driven the need for advanced wireless architectures to provide high throughput, low latency, and efficient energy utilization [1]. Recent studies have highlighted that next-generation wireless networks will require revolutionary technologies to meet these demanding requirements [2]. Among the promising technologies, mobile edge computing (MEC), non-orthogonal multiple access (NOMA) and cell-free massive MIMO (CF-mMIMO) have recently emerged as potential candidates as shown in [3], [4], [5]. As a key enabler for 6G networks, MEC marks a fundamental shift from traditional cloud computing by bringing computational resources closer to end-users at the network edge [6], [7]. In particular, MEC significantly reduces latency in data transmission and processing by enabling distributed computation [2], thereby enhancing the performance of latency-sensitive applications [8], [9].

On the other hand, NOMA improves wireless network capacity by allowing multiple users to share frequency resources [10], [11]. By using power-domain multiplexing, NOMA assigns lower power to strong users and higher power to weaker ones [12]. Traditionally, NOMA pairs user equipments (UEs) based on channel disparity, where UEs with stronger channels decode and subtract interference from weaker-channel UEs using successive interference cancellation (SIC) [13]. However, pairing-based NOMA suffers from inefficiencies in dynamic environments, as optimal user pairing is computationally demanding and requires continuous real-time adjustments [8].

To address these limitations, this paper introduces a clustering-based NOMA approach utilizing K-means clustering, which offers several key advantages as follows. First, instead of manually pairing UEs, K-means clustering autonomously groups UEs based on channel similarity and spatial proximity [14]. Compared to heuristic pairing strategies, K-means clustering significantly reduces computational complexity while enhancing scalability [15]. Furthermore, by effectively clustering UEs, interference management is optimized, leading to higher spectral efficiency and lower latency.

Aiming to minimize system latency, this work proposes a low-complexity optimization framework for NOMA-aided

This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.04-2023.40.
Corresponding author: nvhieu@dut.udn.vn

979-8-3315-9431-2/25/$31.00 ©2025 IEEE

MEC CF-mMIMO networks, where we employ successive convex approximation (SCA) method to jointly optimize power allocation and computational offloading decisions while maintaining energy efficiency constraints. By leveraging K-means clustering for user grouping and SCA for resource allocation, the proposed solution may achieve significant latency reduction compared to conventional approaches. Finally, extensive numerical results are used to demonstrate that the proposed framework may reduce median latency while maintaining comparable energy efficiency over the traditional mMIMO systems. The main contributions can be summarized as follows:

- We first introduce a clustering NOMA approach for CF-mMIMO MEC networks based on K-means algorithm, aiming to replace the heuristic pairing NOMA for optimal user grouping.
- We further formulate a joint optimization problem to address power allocation and computational offloading tasks, wherein the objective is to minimize the system latency while maintaining energy efficiency.
- To solve the formulated problem, we propose a two-step algorithm that combines K-means clustering with SCA, enabling an efficient resolution of the non-convex optimization problem.
- Lastly, extensive simulation results are used to demonstrate the outperformance of the proposed method in comparison with the pairing NOMA-based CF-mMIMO, the CF-mMIMO, and traditional mMIMO schemes.

The structure of this paper is further organized as follows: Section II presents the system model and problem formulation. Section III details the clustering NOMA approach using K-means and the SCA-based optimization method. Section IV provides simulation results and comparisons. Section 5 concludes the paper.

***Notations*:**

The main notations are listed as illustrated in Table I.

TABLE I
KEY NOTATIONS

Symbol	Definition
System Parameters	
K, M, N	Number of UEs, APs, and clusters
$\mathcal{C}_n$	Set of UEs in cluster n
B	System bandwidth
Channel Parameters	
$h_{m,k}$	Channel coefficient between AP m and UE k
$g_{m,k}$	Small-scale Rayleigh fading coefficient
$\beta_{m,k}$	Large-scale fading coefficient
α	Path loss exponent
Resource Variables	
P_k	Transmit power of UE k
P_{max}	Maximum transmit power constraint
ρ_k	Computation offloading ratio of UE k
Task-Related Parameters	
D_k	Data size of UE k's task (bits)
C_k	CPU cycles required per bit for UE k
f_{serv}	MEC server computing frequency
Performance Metrics	
L_k	Total latency for UE k
R_k	Achievable rate for UE k
λ	Energy-latency trade-off coefficient
Mathematical Operators	
$\|\cdot\|$	Euclidean norm
$(\cdot)^T$	Matrix/vector transpose
$\mathbb{E}[\cdot]$	Expectation operator
$\min(\cdot)$	Minimum value
$\max(\cdot)$	Maximum value
$\arg\min$	Argument of minimum
$\mathcal{CN}(0,1)$	Complex normal distribution

II. SYSTEM MODEL AND PROBLEM FORMULATION

A. Network Architecture

We consider a clustering NOMA-based CF-mMIMO system integrated with MEC, where multiple APs are deployed over a wide area to serve a set of UEs. Each AP is equipped with a single antenna and connected to a centralized processor via backhaul links, facilitating coordinated signal processing. Operating in a NOMA framework, multiple UEs share the same time-frequency resources through power-domain multiplexing, while the cell-free architecture eliminates cell boundaries to enhance spectral efficiency and reduce interference.

Let *M* denote the number of APs, and *K* denote the total number of UEs in the system. Each AP *m* is equipped with a single antenna and serves multiple UEs through the clustering NOMA technique. The UEs are randomly distributed across the coverage area and grouped into *N* clusters based on their channel conditions and spatial proximity to APs using the K-means algorithm. This clustering approach enables efficient power-domain multiplexing by grouping UEs with similar channel characteristics, thereby optimizing resource allocation and reducing interference between clusters.

B. Channel Model

The wireless channel between AP m, $m \in \mathcal{M} \triangleq \{1, 2, \cdots, M\}$, and UE k, $k \in \mathcal{K} \triangleq \{1, 2, \cdots, K\}$, is modeled as:

$$h_{m,k} = g_{m,k}\sqrt{\beta_{m,k}}, \tag{1}$$

where $g_{m,k} \sim \mathcal{CN}(0,1)$ represents small-scale Rayleigh fading. $\beta_{m,k}$ accounting for large-scale fading is defined as:

$$\beta_{m,k} = \frac{1}{d_{m,k}^{\alpha}}, \tag{2}$$

where $d_{m,k}$ is the distance between AP m and UE k, α is the path loss exponent.

C. SINR and Data Rate in Clustering NOMA

We suppose that the NOMA is employed using user clustering approach, and then UEs are assigned to one of N clusters. Without loss of generality, the set including the indexes of UEs in the cluster n, $n \in \mathcal{N} \triangleq \{1, 2, \cdots, N\}$ is denoted as $\mathcal{C}_n$. Considering cluster n and applying the power-domain NOMA principles, we obtain the signal-to-interference-plus-noise ratio (SINR) and achievable data rates as follows.

By utilizing maximum ratio combining (MRC) receiver, the SINR for decoding the message at a certain UE k in $\mathcal{C}_n$ is given by

$$\text{SINR}_k = \frac{P_k \|\mathbf{h}_k^H \mathbf{h}_k\|_2^2}{\sum_{i \in \mathcal{K} \setminus \mathcal{C}_n} P_i \|\mathbf{h}_k^H \mathbf{h}_i\|_2^2 + IN + \sigma^2}, \tag{3}$$

where $IN = \sum_{i \in \mathcal{C}_n, \Pi_{\mathcal{C}_n}(i) > \Pi_{\mathcal{C}_n}(k)} P_i \|\mathbf{h}_k^H \mathbf{h}_i\|_2^2$, P_k and $\mathbf{h}_k \triangleq [h_{1,k},\ h_{2,k},\ \cdots,\ h_{M,k}]^T$ represent the power allocated for UE k and the channel vector of UE k to M APs. $\Pi_{\mathcal{C}_n}(k)$ stands for the decoding order of UE k in cluster n. The noise power is specified as the additional white Gaussian noise, with the noise variance σ^2. It is emphasized that by employing the SIC technique in a cluster, the messages of UEs with stronger signals are decoded and removed from the received signal before decoding the other messages. Thus, the achievable data rate for UE k in cluster C_n is:

$$R_k = B \log_2(1 + \text{SINR}_k), \qquad \text{[bps/Hz]} \tag{4}$$

where B is the system bandwidth.

D. Latency and Energy Consumption Model

The total latency for user k consists of local computing and edge computing components:

$$L_k = \max\{L_{\text{local},k}, L_{\text{edge},k}\}. \tag{5}$$

Local Computing Latency:

$$L_{\text{local},k} = \frac{(1-\rho_k) D_k C_k}{f_k^{\text{local}}}, \tag{6}$$

where f_k^{local} is the local CPU frequency in cycles/second.

Edge Computing Latency:

$$L_{\text{edge},k} = L_{\text{trans},k} + L_{\text{comp},k} + L_{\text{queue},k}, \tag{7}$$

where $L_{\text{trans},k} = \frac{\rho_k D_k}{R_k}$, $L_{\text{comp},k} = \frac{\rho_k D_k C_k}{f_k^{\text{serv}}}$ and $L_{\text{queue},k}$ illustrate the transmission latency, computation latency at the MEC server, and queuing latency, respectively. D_k is the task size (in bits) and R_k is the achievable data rate. C_k is the number of CPU cycles per bit and f_{serv} is the MEC server frequency.

The total energy consumption of UE k consists of both transmission and computation energy [8], [16]:

$$E_k = E_{\text{trans},k} + E_{\text{comp},k}. \tag{8}$$

The transmission energy is

$$E_{\text{trans},k} = \rho_k \frac{D_k}{R_k} p_k. \tag{9}$$

The local computation energy consumption is:

$$E_{\text{comp},k} = \kappa_k (f_k^{\text{local}})^2 (1-\rho_k) D_k C_k, \tag{10}$$

where κ_k represents the effective switched capacitance coefficient of user k's processing unit. It is worth noting that the total energy consumption is considered a penalty in the objective function for optimal energy usage, *i.e.*:

$$E_{penalty} = \sum_{k=1}^{K} E_k. \tag{11}$$

E. Optimization Problem Formulation

Given that, the objective is to optimize the power allocation and offloading ratio, aiming to minimize latency while considering energy consumption constraints. The optimization problem is formulated as follows:

$$\min_{\mathbf{p},\boldsymbol{\rho}} \quad \max_{k \in \mathcal{K}} \{L_k\} + \lambda E_{\text{penalty}} \tag{12a}$$

$$\text{s.t.} \quad 0 \le P_k \le P_{\max}, \quad \forall k, \tag{12b}$$

$$0 \le \rho_k \le 1, \quad \forall k, \tag{12c}$$

$$R_k \ge \bar{R}_k, \quad \forall k, \tag{12d}$$

where (12a) represents the trade-off between latency and energy efficiency. (12b) ensures that power allocation does not exceed the maximum limit. (12c) enforces valid offloading ratios. (12d) guarantees that data rate remains within an acceptable range.

III. Proposed Solution: K-means Clustering and SCA Optimization

Due to the coupled relationship between power allocation, offloading decisions, latency, and energy consumption, the formulated problem is inherently a non-convex optimization problem. To efficiently solve this problem, we propose a two-step approach as follows:

A. K-means Clustering Algorithm

We use the K-means algorithm to form clusters based on UE spatial distribution and channel conditions. The clustering process is formalized in **Algorithm** 1.

1) Feature Selection for Clustering: For each UE k, we construct a feature vector that captures both the channel quality and spatial information:

$$\mathbf{f}_k = [\beta_{1,k}, \beta_{2,k}, ..., \beta_{M,k}, d_{1,k}, d_{2,k}, ..., d_{M,k}]^T, \tag{13}$$

where $\beta_{m,k}$ represents the large-scale fading coefficient between AP m and UE k, and $d_{m,k}$ denotes their physical distance. This feature selection ensures that UEs with similar channel conditions and proximity to APs are grouped together, facilitating efficient power allocation and interference management.

B. SCA-based Optimization method

After obtaining the cluster assignments from K-means, we proceed to solve the non-convex optimization problem in (12a) for power allocation and computation offloading. Due to the coupled relationship between variables and non-linear constraints, we employ a SCA-based algorithm to transform the original problem into a sequence of tractable convex subproblems.

1) Problem Analysis: First, we rewrite the original optimization problem (12a) using the explicit latency expression:

$$\min_{\mathbf{p},\boldsymbol{\rho}} \quad L_k + \lambda E_{penalty} \tag{14a}$$

$$\text{s.t.} \quad 0 \le P_k \le P_{\max}, \quad \forall k, \tag{14b}$$

$$0 \le \rho_k \le 1, \quad \forall k. \tag{14c}$$

$$R_k \ge \bar{R}_k, \quad \forall k, \tag{14d}$$

$$L_k \le t, \quad \forall k, \tag{14e}$$

The non-convexity in this problem arises from two sources:

Algorithm 1 K-means Clustering for NOMA User Grouping

Require:
Set of UEs $\mathcal{K} = \{1, 2, ..., K\}$
Number of clusters N
Feature vectors $\{\mathbf{f}_k\}_{k=1}^{K}$
Ensure:
Cluster assignments $\{\mathcal{C}_n\}_{n=1}^{N}$
Initialize: Randomly select N initial centroids $\{\boldsymbol{\mu}_n^{(0)}\}_{n=1}^{N}$
Set iteration index $t = 0$
repeat
  // Cluster Assignment
  for $k \in \mathcal{K}$ **do**
    $n_k^* = \arg\min_n \|\mathbf{f}_k - \boldsymbol{\mu}_n^{(t)}\|^2$
    Assign UE k to cluster $\mathcal{C}_{n_k^*}$
  end for
  // Centroid Update
  for $n = 1$ to N **do**
    $\boldsymbol{\mu}_n^{(t+1)} = \frac{1}{|\mathcal{C}_n|} \sum_{k \in \mathcal{C}_n} \mathbf{f}_k$
  end for
  $t = t + 1$
until Centroids converge or maximum iterations reached
return Cluster assignments $\{\mathcal{C}_n\}_{n=1}^{N}$

- The coupling between power allocation p_k and achievable rate R_k in the transmission latency term.
- The product of optimization variables ρ_k and R_k in the denominator.

2) Convex Approximation: To handle these non-convexities, we apply first-order Taylor approximation around the current point $(\mathbf{p}^{(n)}, \boldsymbol{\rho}^{(n)})$. Let's focus on the transmission latency term:

$$\frac{\rho_k D_k}{R_k} \leq \frac{\rho_k^{(n)} D_k}{R_k^{(n)}} + \frac{D_k}{R_k^{(n)}}(\rho_k - \rho_k^{(n)}) - \frac{\rho_k^{(n)} D_k}{(R_k^{(n)})^2} \nabla_{\mathbf{p}} R_k^{(n)} (\mathbf{p} - \mathbf{p}^{(n)}). \tag{15}$$

The computation latency term $\frac{\rho_k D_k C_k}{f_{serv}}$ is already convex in ρ_k, requiring no approximation.

3) SCA Algorithm: We propose an iterative algorithm to solve the transformed problem:

4) Convergence Analysis: The convergence of Algorithm 2 is guaranteed by the following theorem:

Theorem 1. *The sequence of solutions generated by Algorithm 2 converges to a stationary point of the original problem.*

Proof. The key properties ensuring convergence are:

1) The Taylor approximation provides a locally tight upper bound.
2) The objective value decreases monotonically.
3) The feasible set is compact and convex.

At each iteration, we have:

$$f(\mathbf{x}^{(n+1)}) \leq \tilde{f}(\mathbf{x}^{(n+1)}; \mathbf{x}^{(n)}) \leq \tilde{f}(\mathbf{x}^{(n)}; \mathbf{x}^{(n)}) = f(\mathbf{x}^{(n)}), \tag{16}$$

where f is the original objective and $\tilde{f}$ is its convex approximation. This monotone decrease, combined with the bounded feasible set, ensures convergence to a local optimum. □

Algorithm 2 SCA for Joint Power and Offloading Optimization

Initialize: $\mathbf{p}^{(0)}$, $\boldsymbol{\rho}^{(0)}$, convergence threshold ϵ
Set iteration index $n = 0$
repeat
  Compute $R_k^{(n)}$ and $\nabla_{\mathbf{p}} R_k^{(n)}$ for all k
  Solve the convex subproblem:

$$\min_{\mathbf{p}, \boldsymbol{\rho}} \sum_{k=1}^{K} \Big(\text{Approximated transmission latency} + \frac{\rho_k D_k C_k}{f_{serv}} + L_{queue,k} \Big) + \lambda E_{penalty}$$

  s.t. Original linear constraints
  Update: $(\mathbf{p}^{(n+1)}, \boldsymbol{\rho}^{(n+1)})$
  $n = n + 1$
until $||\mathbf{x}^{(n+1)} - \mathbf{x}^{(n)}|| < \epsilon$
return Optimal solution $(\mathbf{p}^*, \boldsymbol{\rho}^*)$

5) Computational Complexity: The convex subproblem in each iteration can be efficiently solved using standard convex optimization tools. The computational complexity per iteration is determined as $\mathcal{O}(\mathrm{v}^2\mathrm{c}^{2.5} + \mathrm{c}^{3.5})$, where v and c are the number of variables and constraints, respectively. For a given K UEs, we have $\mathrm{v} = 2K$ and $\mathrm{c} = 3K$, and thus the per-iteration complexity is approximated as $\mathcal{O}(K^{4.5})$. Consequently, the total complexity is estimated as $\mathcal{O}(K^{4.5} \log(1/\epsilon))$, with the typical convergence reaching at 5-7 iterations. Remarkably, the separable structure of the objective function allows for potential parallel implementation, where each AP can optimize its local variables while coordinating through the centralized processor. This makes the algorithm suitable for practical deployment in CF-mMIMO systems.

IV. Performance Evaluation

A. Simulation Setup

To evaluate the effectiveness of the proposed clustering NOMA (K-means) CF-mMIMO architecture, we conduct extensive Monte Carlo simulations. The system parameters are based on practical configurations for 5G and beyond networks, as shown in Table II.

TABLE II
Simulation Parameters

Parameter	Value
Network area ($D_m \times D_m$)	$1km \times 1km$
Number of APs	64
Number of UEs	10
Carrier frequency	5 GHz
Noise power	-174 dBm/Hz
Bandwidth	20 MHz
Path loss exponent	3.7
Maximum UE transmission power	10 dBm
MEC server frequency (f_{serv})	$\{5, 10\}$ GHz
Coefficient λ (coef of energy)	{0, 1, 2, 3, 4, 5, 10, 15, 20}

We compare the following mMIMO-based architectures: traditional mMIMO, CF-mMIMO, pairing (Greedy) NOMA, and clustering (K-means) NOMA.

B. Convergence Analysis

Fig. 1 below demonstrates the convergence characteristics of our proposed SCA optimization algorithm under different MEC server frequencies and energy-latency trade-off coefficients (λ). Our analysis focuses on two key server frequencies (5 GHz and 10 GHz) and three λ values (0, 20, and 100) to comprehensively evaluate system performance.

At 5 GHz server frequency with high energy efficiency priority ($\lambda = 100$), the system starts with the highest initial latency of approximately 0.19 ms but converges to 0.143 ms within 5 iterations. When λ decreases to 20, the initial latency improves to 0.17 ms and converges to 0.142 ms. At $\lambda = 0$, focusing purely on latency, the system achieves faster convergence, starting at 0.15 ms and stabilizing at 0.141 ms. Operating at 10 GHz delivers significantly better performance across all λ values. With $\lambda = 100$, the system converges from 0.13 ms to 0.085 ms. At $\lambda = 20$, similar convergence behavior is observed but with slightly improved initial conditions. The best performance is achieved at $\lambda = 0$, where the system quickly stabilizes at 0.082 ms within 4 iterations.

Most notably, all configurations demonstrate consistent convergence characteristics, requiring 4-6 iterations to reach stability regardless of server frequency or λ value. The 10 GHz operation achieves approximately 42% lower steady-state latency compared to 5 GHz across all energy efficiency settings, validating our algorithm's effectiveness in balancing latency minimization and energy efficiency requirements.

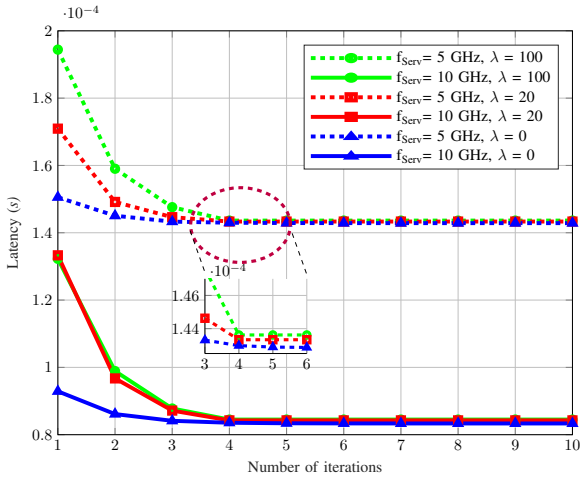


Fig. 1. System latency convergence under different MEC configurations.

C. Impact of Energy Efficiency Factor

Fig. 2 presents the relationship between system latency and energy efficiency factor (λ) for PairNOMA and ClusterNOMA architectures at two MEC server frequencies (f_{serv} = 5 GHz and 10 GHz).

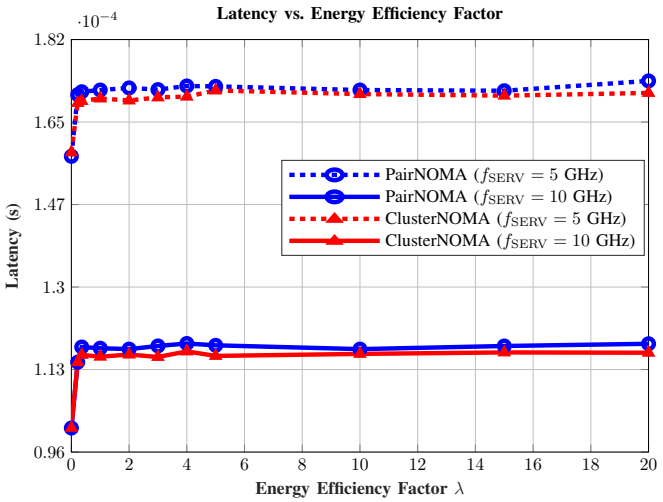


Fig. 2. Impact of energy efficiency factor on system latency under different NOMA architectures and MEC frequencies.

For f_{serv} = 5 GHz operation, both architectures exhibit similar initial latency of 1.55×10^{-4} s at λ = 0. The latency increases sharply to 1.69×10^{-4} s for ClusterNOMA and 1.71×10^{-4} s for PairNOMA at λ = 2, after which it maintains relative stability through λ = 20. ClusterNOMA consistently demonstrates marginally superior performance, achieving approximately 1.2% lower latency compared to PairNOMA. With f_{serv} = 10 GHz, both architectures show substantial performance improvements. The initial latency at λ = 0 is 0.96×10^{-4} s, rising to and stabilizing around 1.13×10^{-4} s for ClusterNOMA and 1.15×10^{-4} s for PairNOMA when $\lambda \geq 2$. This represents a significant 34% reduction in latency compared to 5 GHz operation. The results demonstrate that ClusterNOMA maintains consistently lower latency across all λ values, with enhanced performance at higher server frequencies. The stability observed for $\lambda \geq 2$ indicates that our proposed system effectively preserves low-latency performance while meeting various energy efficiency requirements.

D. Latency CDF Analysis

Fig. 3 depicts the cumulative distribution function (CDF) of system latency across different massive MIMO-based architectures, where markers denote experimental data points while curves represent fitted results. The curves characterize the probability that system latency falls below a certain threshold value.

Our proposed Kmean.NOMA-based CF-mMIMO architecture (green solid line) demonstrates superior performance, achieving a median latency of 153 μs and reaching 80th percentile at 158 μs. The steeper slope of its CDF curve indicates more consistent latency performance across users. The Pair.NOMA CF-mMIMO implementation (red dash-dotted line) follows as the second-best performer, with its curve positioned slightly to the right, reaching median latency at 155 μs. The CF-mMIMO architecture (blue dashed curve) shows moderate performance with 157 μs median latency,

while traditional mMIMO (black dotted curve) exhibits the highest latency variation with 177 μs median latency.

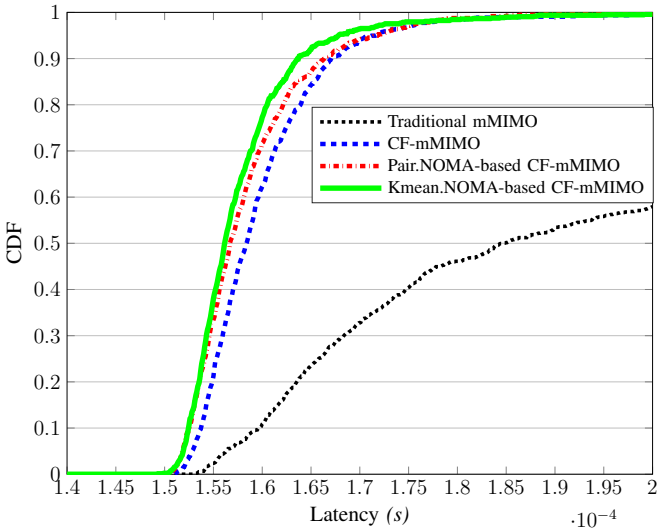


Fig. 3. CDF of latency for different massive MIMO-based architectures.

From analyzing the CDF curves in Fig. 3, we summarize the key latency metrics in Table III:

TABLE III
LATENCY PERFORMANCE COMPARISON ACROSS ARCHITECTURES

Architecture	**Latency Range (μs)**	**50th Percentile (μs)**	**80th Percentile (μs)**
Traditional mMIMO	154-200	177	200
CF-mMIMO	152-168	157	165
Pair.NOMA-based CF-mMIMO	150-163	155	163
Kmean.NOMA-based CF-mMIMO	148-158	153	158

The performance advantages of our proposed architecture are evident: it achieves 12.6%, 5.6%, and 2.5% lower median latency compared to traditional mMIMO, CF-mMIMO, and Pair.NOMA CF-mMIMO, respectively. Furthermore, its compressed latency range (148-158 μs) indicates more stable performance across network conditions. These results validate the effectiveness of our clustering-based approach in optimizing user grouping and resource allocation for latency-sensitive applications.

V. CONCLUSION

This paper has introduced an innovative clustering NOMA approach for latency optimization in CF-mMIMO systems with mobile edge computing. Our two-step optimization framework combines K-means clustering for user grouping with successive convex approximations for power and computational resource allocation.

Performance evaluation demonstrates that our proposed architecture achieves significant improvements over conventional approaches. The system converges rapidly within 4-6 iterations, delivering up to 38% latency reduction at higher MEC frequencies. The latency performance remains stable across varying energy efficiency requirements, with 80% users experiencing delays below $158\mu s$ - a substantial improvement over traditional architectures.

These results validate our framework's effectiveness in balancing computational offloading with power allocation, making it particularly suitable for delay-sensitive applications in 5G and beyond networks. Future work could explore dynamic clustering strategies and integration with distributed MEC servers to further enhance system performance.

REFERENCES

[1] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, "The road towards 6g: A comprehensive survey," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, 2021.

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6g wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.

[3] G. Interdonato and S. Buzzi, "Joint optimization of uplink power and computational resources in mobile edge computing-enabled cell-free massive MIMO," *IEEE Trans. Commun.*, vol. 72, no. 3, pp. 1804–1820, 2024.

[4] H. V. Nguyen, M. T. P. Le, T. D. Ho, P. V. Tuan, and H. Nguyen-Le, "Joint latency minimization and power allocation for mec-enabled mu-miso networks," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, 2024, pp. 753–757.

[5] M. T. P. Le, H. V. Nguyen, V. Nguyen-Duy-Nhat, and L. Sanguinetti, "Qoe-aware power allocation for aerial-relay massive MIMO networks," *IEEE Trans. Netw. Serv. Manag.*, vol. 21, no. 1, pp. 477–489, 2024.

[6] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surv. Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.

[7] S. S. Yılmaz, B. Özbek, and R. Mumtaz, "Delay minimization for massive MIMO based cooperative mobile edge computing system with secure offloading," *IEEE Open J. Veh. Technol.*, vol. 4, pp. 149–161, 2022.

[8] T. V. Thai, M. T. P. Le, H. V. Nguyen, and O.-S. Shin, "Noma-aided cell-free massive mimo with mec: A trade-off between latency and energy consumption," in *Proc. IEEE Int. Conf. Consum. Electron.-Asia (ICCE-Asia)*, 2024, pp. 1–5.

[9] G. Femenias and F. Riera-Palou, "Mobile edge computing aided cell-free massive MIMO networks," *IEEE Trans. Mobile Comput.*, vol. 23, no. 2, pp. 1246–1261, 2024.

[10] H. Zhang, F. Fang, J. Cheng, K. Long, W. Wang, and V. C. M. Leung, "Energy-efficient resource allocation in noma heterogeneous networks," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 48–53, 2018.

[11] K.-H. Nguyen, H. V. Nguyen, M. T. P. Le, L. Sanguinetti, and O.-S. Shin, "On the energy efficiency maximization of NOMA-aided downlink networks with dynamic user pairing," *IEEE Access*, vol. 10, pp. 35 131–35 145, 2022.

[12] X.-T. Dang, M. T. P. Le, H. V. Nguyen, and O.-S. Shin, "Optimal user pairing for NOMA-assisted cell-free massive MIMO system," in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, 2022, pp. 7–12.

[13] X.-T. Dang, M. T. P. Le, H. V. Nguyen, S. Chatzinotas, and O.-S. Shin, "Optimal user pairing approach for NOMA-based cell-free massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 4751–4765, 2023.

[14] M. Katwe, K. Singh, P. K. Sharma, C.-P. Li, and Z. Ding, "Dynamic user clustering and optimal power allocation in uav-assisted full-duplex hybrid noma system," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2573–2590, 2022.

[15] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci.*, vol. 622, no. C, p. 178–210, Apr. 2023.

[16] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive mimo versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, 2017.

DRL-Based Approach for RIS-Aided NOMA System in Short Packet Communications

Waqas Khalid
Institute of Industrial technology
Korea University
Sejong, South Korea
waqas283@korea.ac.kr; waqas283@gmail.com

***Abstract*—This paper investigates the power allocation problem in a reconfigurable intelligent surface (RIS)-aided non-orthogonal multiple access (NOMA) system tailored for short packet communications (SPC) in beyond 5G (B5G) and 6G networks. We propose a deep reinforcement learning (DRL)-based approach utilizing the proximal policy optimization (PPO) algorithm to maximize the sum achievable data rate by optimizing transmission power and power allocation coefficients, while adhering to power constraints and ensuring low-latency requirements of SPC. The RIS enhances signal quality by dynamically adjusting the wireless propagation environment, while NOMA enables efficient spectrum sharing for multiple users. However, the non-convex nature of the optimization problem, coupled with the dynamic channel conditions of SPC, poses significant challenges for traditional methods. The proposed DRL-based solution leverages PPO's sample efficiency and stability to adaptively handle these dynamics, offering a computationally efficient framework that reduces training data and time requirements compared to conventional optimization techniques.**

***Index Terms*—Reconfigurable intelligent surface (RIS), NOMA, Short packet communications (SPC), DRL, PPO.**

I. Introduction

The rapid evolution of wireless communication networks towards beyond 5G (B5G) and 6G is driven by the need to support emerging applications such as the Internet of Things (IoT), massive machine-type communications (mMTC), and ultra-reliable low-latency communications (URLLC). These applications demand unprecedented performance metrics, including ultra-high data rates (up to 1 Tbps), enhanced reliability (e.g., 99.9999% packet success probability), ultra-low latency (sub-millisecond), and efficient spectrum utilization to accommodate massive device connectivity [1]. Short Packet Communications (SPC) have emerged as a critical enabler for these requirements, particularly in IoT and mMTC scenarios, by facilitating the transmission of small data packets with minimal latency and high reliability. However, SPC introduces unique challenges, such as finite blocklength effects, which degrade the achievable rate and necessitate careful resource management to meet stringent latency constraints [2].

Non-orthogonal multiple access (NOMA) is a pivotal technology for 6G, enabling multiple users to share the same time-frequency resources through power-domain multiplexing, thereby significantly improving spectral efficiency compared to traditional orthogonal multiple access (OMA) schemes. The integration of reconfigurable intelligent surfaces (RIS) with NOMA further enhances system performance by dynamically manipulating the wireless propagation environment [3]. RIS, composed of numerous low-cost passive reflecting elements, can adjust the phase and amplitude of incident signals to create favorable channel conditions, mitigate interference, and extend coverage, making it a cost-effective solution for 6G networks. In RIS-aided NOMA systems, the RIS can steer signals towards desired users, improving the signal-to-interference-plus-noise ratio (SINR) and enabling efficient power allocation, which is crucial to maximize system throughput [4].

Despite these advantages, optimizing RIS-aided NOMA systems for SPC poses significant challenges. The joint optimization of transmission power and power allocation coefficients results in a non-convex optimization problem due to the non-linear relationship between the rate, SINR, and system parameters. The dynamic nature of wireless channels, coupled with the finite blocklength constraints of SPC, also introduces time-varying conditions that traditional optimization methods, such as convex optimization or iterative algorithms, struggle to address efficiently. These methods incur high computational complexity and fail to adapt to real-time channel variations, limiting their applicability in dynamic 6G environments [5].

To overcome these challenges, deep reinforcement learning (DRL) offers a powerful framework for solving complex optimization problems by learning optimal policies through interactions with the environment [5]. DRL can effectively handle the non-convexity and dynamic nature of RIS-aided NOMA systems, providing adaptive and scalable solutions. In this paper, we propose a DRL-based approach using the proximal policy optimization (PPO) algorithm to optimize power allocation in an RIS-aided NOMA system for SPC. Our approach aims to maximize the sum achievable data rate while adhering to power constraints, leveraging PPO's sample efficiency and training stability to achieve a computationally efficient solution. By addressing the unique challenges of SPC in 6G, our method paves the way for efficient and adaptive resource management in future wireless networks.

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) under Grant NRF-2022R1I1A1A01071807.

979-8-3315-9431-2/25/$31.00 ©2025 IEEE

II. System Model

We consider an RIS-aided NOMA downlink network designed for short packet communications (SPC). The system comprises a base station (BS) equipped with a single antenna, serving two single-antenna users: a near user (U_N) and a far user (U_F). The RIS, consisting of L passive reflecting elements, is deployed to enhance signal quality by dynamically adjusting the wireless propagation environment [6]. The channel between the BS and the RIS is denoted as $\mathbf{h}_{BR} \in \mathbb{C}^{L\times 1}$, while the channels from the RIS to users U_N and U_F are $\mathbf{g}_N \in \mathbb{C}^{1\times L}$ and $\mathbf{g}_F \in \mathbb{C}^{1\times L}$, respectively. Additionally, direct channels from the BS to U_N and U_F are denoted as $h_{BN} \in \mathbb{C}$ and $h_{BF} \in \mathbb{C}$, respectively, accounting for potential line-of-sight (LoS) paths. All channels follow a Rayleigh fading model, with elements modeled as independent complex Gaussian random variables, i.e., $\mathbf{h}_{BR} \sim \mathcal{CN}(0, \sigma_{BR}^2\mathbf{I}_L)$, $\mathbf{g}_N \sim \mathcal{CN}(0, \sigma_N^2\mathbf{I}_L)$, $\mathbf{g}_F \sim \mathcal{CN}(0, \sigma_F^2\mathbf{I}_L)$, $h_{BN} \sim \mathcal{CN}(0, \sigma_{BN}^2)$, and $h_{BF} \sim \mathcal{CN}(0, \sigma_{BF}^2)$, where σ_{BR}^2, σ_N^2, σ_F^2, σ_{BN}^2, and σ_{BF}^2 represent the large-scale path loss and shadowing effects. We assume perfect CSI is available at the BS for analytical traceability [7, 8], though this is idealized; practical imperfections, such as channel estimation errors, will be explored in future work. The RIS phase-shift matrix can be defined as $\mathbf{\Theta} = \text{diag}\left(e^{j\phi_1}, \ldots, e^{j\phi_L}\right)$, where $\phi_l \in [0, 2\pi)$ is the phase shift for the l-th reflecting element ($l = 1, \ldots, L$). For simplicity, we assume the RIS elements are passive with unit amplitude (i.e., no energy loss during reflection), focusing on phase adjustments to optimize signal propagation [9, 10].

The BS transmits a superimposed NOMA signal $x = \sqrt{\alpha_N P}x_N + \sqrt{\alpha_F P}x_F$, where P is the total transmit power, α_N and α_F are the power allocation coefficients with $\alpha_N + \alpha_F = 1$, and x_N and x_F are the unit-power symbols for U_N and U_F, respectively.

The received signals at U_N and U_F are:

$$y_N = (h_{BN} + \mathbf{g}_N\mathbf{\Theta}\mathbf{h}_{BR})\left(\sqrt{\alpha_N P}x_N + \sqrt{\alpha_F P}x_F\right) + n_N, \tag{1}$$

$$y_F = (h_{BF} + \mathbf{g}_F\mathbf{\Theta}\mathbf{h}_{BR})\left(\sqrt{\alpha_N P}x_N + \sqrt{\alpha_F P}x_F\right) + n_F, \tag{2}$$

where $n_N, n_F \sim \mathcal{CN}(0, \sigma^2)$ are the additive white Gaussian noise (AWGN) terms at U_N and U_F, respectively.

In NOMA, user decoding order and power allocation are determined based on channel conditions. The effective channel gains are defined as $G_N = |h_{BN} + \mathbf{g}_N\mathbf{\Theta}\mathbf{h}_{BR}|$ for U_N and $G_F = |h_{BF} + \mathbf{g}_F\mathbf{\Theta}\mathbf{h}_{BR}|$ for U_F. Given that U_N is closer to the BS and typically experiences a stronger channel (i.e., $G_N > G_F$), we assign less power to U_N and more to U_F, i.e., $\alpha_N < \alpha_F$, to ensure fairness and maximize the sum rate. Successive interference cancellation (SIC) is applied at the stronger user (U_N) [3]. Specifically, U_N first decodes x_F (the far user's signal) by treating x_N as interference, requiring the SINR for decoding x_F at U_N to satisfy the target rate. If successful, U_N subtracts x_F from the received signal and decodes its own signal x_N with no interference. Conversely, U_F, as the weaker user, decodes its signal x_F directly, treating x_N as interference. The signal-to-interference-plus-noise ratios (SINRs) are:

$$\gamma_{N,F} = \frac{|h_{BN} + \mathbf{g}_N\mathbf{\Theta}\mathbf{h}_{BR}|^2 P\alpha_F}{|h_{BN} + \mathbf{g}_N\mathbf{\Theta}\mathbf{h}_{BR}|^2 P\alpha_N + \sigma^2}, \tag{3}$$

$$\gamma_N = \frac{|h_{BN} + \mathbf{g}_N\mathbf{\Theta}\mathbf{h}_{BR}|^2 P\alpha_N}{\sigma^2}, \tag{4}$$

$$\gamma_F = \frac{|h_{BF} + \mathbf{g}_F\mathbf{\Theta}\mathbf{h}_{BR}|^2 P\alpha_F}{|h_{BF} + \mathbf{g}_F\mathbf{\Theta}\mathbf{h}_{BR}|^2 P\alpha_N + \sigma^2}, \tag{5}$$

where $\gamma_{N,F}$ is the SINR at U_N for decoding x_F, γ_N is the SINR at U_N for decoding x_N after SIC, and γ_F is the SINR at U_F for decoding x_F.

III. Problem Formulation

To evaluate the system performance under SPC conditions, we focus on the achievable data rate, which is critical for meeting the low-latency and high-reliability requirements of 6G applications such as IoT and mMTC. In SPC, the block length K is small (e.g., $K \leq 100$ channel uses) due to stringent latency constraints, leading to finite blocklength effects [11]. Given a target block error rate (BLER) ϵ_u and block length K, the achievable rate for user $u \in \{U_N, U_F\}$ in bits per channel use (bpcu) is approximated as:

$$R_u = \log_2(1 + \gamma_u) - \sqrt{\left(1 - \frac{1}{(1+\gamma_u)^2}\right)\frac{1}{K}}\frac{Q^{-1}(\epsilon_u)}{\ln 2}, \tag{6}$$

where γ_u is the SINR for user u (i.e., γ_N for U_N and γ_F for U_F), $Q^{-1}(\cdot)$ is the inverse Q-function, and ϵ_u is the target BLER for user u. The second term accounts for the rate penalty due to finite block length.

The optimization problem is formulated to maximize the sum achievable rate of both users, ensuring efficient resource utilization while meeting SPC constraints:

$$\max_{P,\,\alpha_N} \quad R_N(P, \alpha_N) + R_F(P, \alpha_N), \tag{7a}$$

$$\text{subject to} \quad 0 \leq P \leq P_{\max}, \tag{7b}$$

$$0 \leq \alpha_N \leq 1, \quad \alpha_F = 1 - \alpha_N, \tag{7c}$$

$$R_N \geq R_{\text{th}}, \quad R_F \geq R_{\text{th}}. \tag{7d}$$

where $P_{\max}$ is the maximum transmission power, α_N and α_F are the power allocation coefficients for users U_N and U_F, respectively, and R_{th} is the minimum rate threshold to ensure quality of service (QoS) for each user. The SINR γ_u depends non-linearly on P and α_N, as shown in the system model, making the optimization problem non-convex. Additionally, the dynamic wireless environment, characterized by time-varying channels and the finite blocklength constraints of SPC, further complicates the problem, rendering traditional optimization methods, such as gradient descent or convex relaxation, inefficient due to their high computational complexity and inability to adapt in real-time. Therefore, we adopt a DRL-based approach to achieve efficient and adaptive optimization.

IV. Proposed Solution: DRL with PPO Algorithm

To address the power allocation challenge in the RIS-aided NOMA system, we employ a DRL framework based on the Proximal Policy Optimization (PPO) algorithm, which is well-suited for learning optimal policies in complex, dynamic environments due to its balance of sample efficiency, training stability, and robustness to hyperparameter variations.

A. DRL Framework

We model the power allocation task as a Markov Decision Process (MDP) with the following components:

- **Agent**: BS selects the transmission power P and power allocation coefficient α_N at each time step t.
- **State Space** ($\mathcal{S}$): The state s_t includes the CSI of all links, i.e., $s_t = \{\mathbf{h}_{BR}, \mathbf{g}_N, \mathbf{g}_F, h_{BN}, h_{BF}\}$, capturing the dynamic channel conditions.
- **Action Space** ($\mathcal{A}$): The action $a_t = [P, \alpha_N]$ is continuous, subject to the constraints $0 \leq P \leq P_{\max}$ and $0 \leq \alpha_N \leq 1$.
- **Reward Function** (r): The reward is designed to maximize the sum achievable rate while satisfying the constraints:

$$r_t = \begin{cases} R_N + R_F, & \text{if } 0 \leq P \leq P_{\max}, \\ & 0 \leq \alpha_N \leq 1, \\ & R_N, R_F \geq R_{\text{th}}, \\ -1, & \text{otherwise.} \end{cases} \tag{8}$$

- **Transition Probability** ($\mathcal{P}$): The environment transitions from state s_t to s_{t+1} based on the action a_t and stochastic channel variations, following the Rayleigh fading model.

B. PPO Algorithm

PPO, a policy gradient method, updates the policy by maximizing a clipped surrogate objective function, ensuring stable learning by preventing excessively large policy updates. The objective function is defined as:

$$\mathcal{L}(\theta) = \mathbb{E}\left[\min\left(r(\theta)A_t, \text{clip}(r(\theta), 1-\epsilon, 1+\epsilon)A_t\right)\right], \tag{9}$$

where θ represents the policy parameters (neural network weights), $r(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ is the probability ratio between the current policy π_θ and the previous policy $\pi_{\theta_{\text{old}}}$, A_t is the advantage function, and ϵ is the clipping hyperparameter (typically set to 0.2) to constrain policy updates.

The training process includes the following steps:

- **Initialization**: Set initial policy parameters θ_0 and value function parameters ϕ_0, typically using a neural network with random weights.
- **Data Collection**: For each iteration k, collect a set of trajectories $D_k = \{(s_t, a_t, r_t, s_{t+1})\}_{t=0}^{T-1}$ by interacting with the environment for T time steps using the current policy π_{θ_k}.
- **Advantage Estimation**: Compute the advantage A_t using Generalized Advantage Estimation (GAE):

$$A_t = \sum_{l=0}^{\infty} (\gamma\lambda)^l \delta_{t+l},$$

where $\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t)$ is the temporal difference error, γ is the discount factor (e.g., 0.99), λ (e.g., 0.95) balances bias and variance, and $V_\phi(s_t)$ is the state-value function approximated by a neural network.

- **Policy Update**: Update the policy parameters θ by maximizing $\mathcal{L}(\theta)$ using mini-batch gradient ascent over D_k, typically for a few epochs (e.g., 10).
- **Value Function Update**: Update the value function parameters ϕ by minimizing the mean squared error:

$$\mathcal{L}_V(\phi) = \mathbb{E}\left[\left(V_\phi(s_t) - (r_t + \gamma V_\phi(s_{t+1}))\right)^2\right].$$

To enhance training efficiency, experience tuples (s_t, a_t, r_t, s_{t+1}) are stored in a replay buffer, and mini-batches are sampled to improve data efficiency. Before each episode, the environment is reset, and key hyperparameters (e.g., learning rate, discount factor, clipping range, entropy coefficient) are adjusted as needed. The advantage estimates help the agent prioritize actions that yield higher long-term rewards, ensuring stable and robust policy updates for the dynamic RIS-aided NOMA system.

V. Numerical Results

In this section, we present the numerical results to evaluate the performance of the proposed DRL-based PPO approach for power allocation in the RIS-aided NOMA system for SPC. We compare the sum achievable data rate (in bits per channel use, bpcu) of the proposed method against a traditional optimization method without RIS (No-RIS), where the RIS phase shifts are not utilized, and only direct channels are considered. The simulations are conducted using MATLAB, with the system parameters listed in Table I.

TABLE I
Simulation Parameters

Parameter	Value
Number of RIS elements (L)	20
Block length (K)	100
Target BLER (ϵ_u)	10^{-3}
Maximum transmit power ($P_{\max}$)	0 to 30 dBm
Noise power (σ^2)	-90 dBm
Path loss exponents (σ^2_{BR}, σ^2_N, σ^2_F, σ^2_{BN}, σ^2_{BF})	-2.5, -2.0, -3.0, -2.8, -3.2
Minimum rate threshold (R_{th})	0.5 bpcu
DRL episodes	1000
Learning rate	0.001
Discount factor (γ)	0.99
Clipping parameter (ϵ)	0.2

Fig. 1 shows the sum achievable rate ($R_N + R_F$) versus the maximum transmit power $P_{\max}$ for the proposed DRL-PPO method and No-RIS scheme. The proposed DRL-PPO approach consistently outperforms No RIS scheme across all power levels, achieving up to a 35% higher sum rate compared to No-RIS case at $P_{\max} = 30$ dBm. This improvement is attributed to the DRL's ability to adaptively optimize P and α_N, leveraging the RIS to enhance the effective channel gains and mitigate interference. The No-RIS scheme suffers from weaker channel gains due to the absence of RIS reflections, leading to significantly lower rates. Fig. 2 illustrates the sum rate versus

the number of RIS elements L at $P_{\max} = 20$ dBm. As L increases from 10 to 50, the sum rate of the proposed DRL-PPO method improves by approximately 20%, demonstrating the benefit of additional reflecting elements in enhancing signal quality. The No-RIS scheme remains unaffected by L. Fig. 3 depicts the convergence behavior of the DRL-PPO algorithm over 1000 episodes. The sum rate converges to a stable value of approximately 41.65 bpcu after around 100 episodes, indicating the algorithm's efficiency in learning the optimal policy. The rapid convergence, facilitated by PPO's sample efficiency and stability, underscores the practicality of the proposed approach for real-time 6G applications.

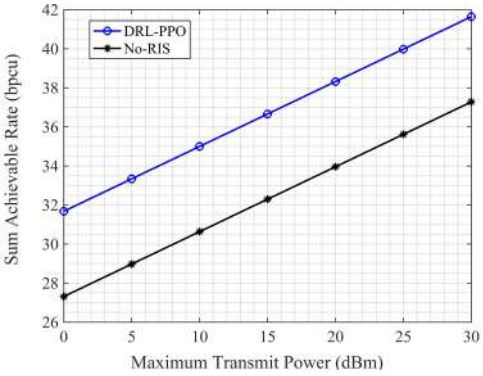


Fig. 1. Sum achievable rate ($R_N + R_F$) versus the maximum transmit power $P_{\max}$.

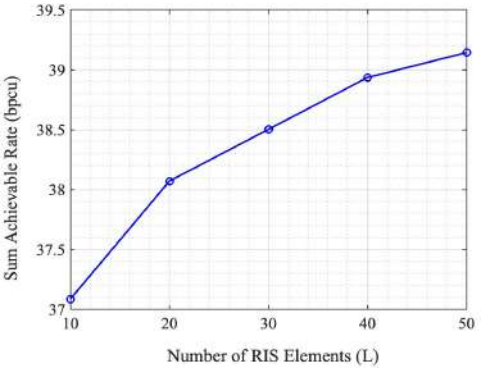


Fig. 2. Sum achievable rate ($R_N + R_F$) vs. maximum transmit power $P_{\max}$.

These results highlight the effectiveness of the DRL-PPO method in optimizing power allocation for RIS-aided NOMA systems under SPC constraints, offering significant performance gains over traditional methods and paving the way for adaptive resource management in future wireless networks.

VI. Conclusion

This paper proposed a DRL-based approach using the PPO algorithm to optimize power allocation in an RIS-aided NOMA system for SPC in 6G networks. By addressing the non-convex optimization challenges and dynamic channel conditions, the proposed method achieved up to a 35% higher sum rate compared to the No-RIS baseline at $P_{\max} = 30$ dBm, with rapid convergence within 100 episodes. The results demonstrate the effectiveness of DRL-PPO in enhancing spectral efficiency and signal quality, paving the way for adaptive resource management in future wireless systems.

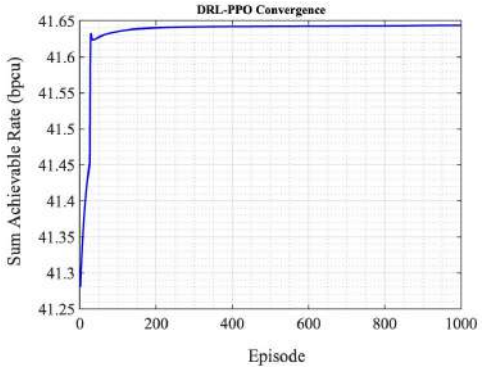


Fig. 3. Sum achievable rate ($R_N + R_F$) versus maximum transmit power $P_{\max}$.

References

[1] W. Khalid, A. -A. A. Boulogeorgos, T. Van Chien, J. Lee, H. Lee and H. Yu, "Optimal Operation of Active RIS-Aided Wireless Powered Communications in IoT Networks," *IEEE Internet of Things Journal*, vol. 12, no. 1, pp. 390-401, 1 Jan. 2025.

[2] N. T. Y. Linh, P. N. Son, and V. N. Q. Bao, "Intelligent reflecting surface-assisted beamforming-noma networks for short-packet communications: Performance analysis and deep learning approach," *IET Communications*, vol. 17, no. 16, pp. 1940–1954, Aug. 2023.

[3] W. Khalid and H. Yu, "Security Improvement With QoS Provisioning Using Service Priority and Power Allocation for NOMA-IoT Networks," *IEEE Access*, vol. 9, pp. 9937-9948, Jan. 2021.

[4] M. R. A. Ruku, M. Ibrahim, A. S. M. Badrudduza, I. S. Ansari, W. Khalid, and H. Yu, "Effects of co-channel interference on RIS empowered wireless networks amid multiple eavesdropping attempts," *ICT Exp.*, vol. 10, no. 3, pp. 491–497, Jun. 2024.

[5] R. Zeng, Z. Feng, Z. Zhang, N. Ying, H. Wang and Y. Yao, "Intelligent Reflective Surface Resource Allocation Algorithm Based on Deep Reinforcement Learning in Internet of Vehicles System," *IEEE Communications Letters*, vol. 28, no. 8, pp. 1885-1888, Aug. 2024.

[6] W. Khalid, M. A. U. Rehman, T. Van Chien, Z. Kaleem, H. Lee and H. Yu, "Reconfigurable Intelligent Surface for Physical Layer Security in 6G-IoT: Designs, Issues, and Advances," *IEEE Internet of Things Journal*, vol. 11, no. 2, pp. 3599-3613, Jan 2024.

[7] W. Khalid, Z. Kaleem, R. Ullah, T. Van Chien, S. Noh and H. Yu, "Simultaneous Transmitting and Reflecting-Reconfigurable Intelligent Surface in 6G: Design Guidelines and Future Perspectives," *IEEE Network*, vol. 37, no. 5, pp. 173-181, Sept. 2023.

[8] W. Khalid, H. Yu, J. Cho, Z. Kaleem and S. Ahmad, "Rate-Energy Tradeoff Analysis in RIS-SWIPT Systems With Hardware Impairments and Phase-Based Amplitude Response," *IEEE Access*, vol. 10, pp. 31821-31835, Mar. 2022.

[9] W. Khalid, H. Yu, D. -T. Do, Z. Kaleem and S. Noh, "RIS-Aided Physical Layer Security With Full-Duplex Jamming in Underlay D2D Networks," *IEEE Access*, vol. 9, pp. 99667-99679, Jul. 2021.

[10] W. Khalid and H. Yu, "Sum Utilization of spectrum with spectrum handoff and imperfect sensing in interweave multi-channel cognitive radio networks," *Sustainability*, vol. 10, no. 6, pp. 1764–1782, May 2018.

[11] J. Zheng, Q. Zhang, and J. Qin, "Average block error rate of downlink noma short-packet communication systems in nakagami-m fading channels," *IEEE Communications Letters*, vol. 23, no. 10, pp. 1712–1716, Jul. 2019.

Enhancing Data Preprocessing Layer for Power Transformer Fault Diagnosis Using DGA Combining Fuzzy Logic and Duval Triangle 1

Kim Anh Nguyen*
The University of Danang - University of Science and Technology
Da Nang, Vietnam
https://orcid.org/0000-0003-3408-847X
*Corresponding author

Huy Hoang Le
Faculty of Electrical Engineering The University of Danang - University of Science and Technology
Da Nang, Vietnam
105200406@sv1.dut.udn.vn

Ba Tu Phung
Faculty of Electrical Engineering The University of Danang - University of Science and Technology
Da Nang, Vietnam
105200478@sv1.dut.udn.vn

Huy Vu Tran
Faculty of Electrical Engineering The University of Danang - University of Science and Technology
Da Nang, Vietnam
thvu@ncs.dut.udn.vn

***Abstract*— The Duval Triangle 1 method, when combined with fuzzy logic, has been widely utilized in the diagnostics of power transformer faults due to its straightforwardness and interpretability. Nevertheless, the conventional approach is fundamentally constrained by its dependence on only three gas components—methane (CH_4), acetylene (C_2H_2), and ethylene (C_2H_4)—which limits its diagnostic capability in situations where other gases such as hydrogen (H_2) and ethane (C_2H_6) are crucial. To overcome these limitations, this study introduces two significant enhancements. Firstly, a refined fuzzy logic framework is developed to address zone boundary ambiguities within the Duval Triangle, thereby enhancing classification precision and achieving a diagnostic accuracy of 90%. Secondly, a data preprocessing layer is incorporated, facilitating the analysis of a more comprehensive set of input gases (%CH_4, %C_2H_2, %C_2H_4, %H_2, and %C_2H_6). This layer performs normalization, outlier detection, and dimensionality adjustment, further elevating diagnostic performance to 97.1%. The proposed improvements are validated using real-world datasets, demonstrating substantial gains in both accuracy and robustness. The results underscore the synergistic advantage of integrating fuzzy logic inference with intelligent data preprocessing, offering a scalable and reliable solution for contemporary transformer condition monitoring.**

Keywords—dissolved gas analysis, power transformer fault diagnostic, Duval triangle method, fuzzy logic, data processing

I. INTRODUCTION

Transformer fault diagnostics play a crucial role in maintaining the reliability and longevity of electrical power systems. One of the most widely used techniques for fault diagnosis is the dissolved gas analysis (DGA), which involves analyzing the gases generated in transformer oil to identify potential faults. The Duval triangle 1 method (DTM) has been traditionally employed for interpreting these gases, providing a simple and effective tool for identifying faults based on the percentage of specific gases such as CH_4, C_2H_2, and C_2H_4 [1,6]. However, the method has limitations, especially when other gases, such as H_2 and C_2H_6, become more prevalent in the transformer oil [2,5].

The traditional DTM is also constrained by its inability to handle complex datasets effectively, often struggling with boundary ambiguities and less accurate diagnoses when dealing with data from transformers with varied gas compositions [3,4]. To address these limitations, fuzzy logic has been integrated into transformer fault diagnostics, offering a more flexible and accurate approach to classifying faults [7,8]. Furthermore, data preprocessing techniques such as normalization, outlier detection, and dimensionality reduction can significantly improve the performance of these diagnostic systems by ensuring the quality of the input data [9,10].

This study aims to enhance the performance of the fuzzy logic and the Duval triangle 1 method (denoted FL—DTM) by incorporating a data preprocessing layer. The proposed methodology addresses the boundary ambiguities of traditional DTM and improves fault classification accuracy, especially when dealing with datasets containing a broader range of dissolved gases. Through these improvements, the accuracy of fault diagnosis can be increased from 90% using fuzzy logic alone to 97.1% with the addition of a preprocessing layer, offering a more robust and reliable solution for transformer condition monitoring.

II. PROPOSED METHODOLOGY

The proposed methodology aims to enhance transformer fault diagnostics by integrating the conventional Duval triangle 1 method with fuzzy logic and a robust data preprocessing layer called data preprocessed fuzzy logic—Duval triangle 1 method (DPFD). This approach addresses the limitations of traditional methods, particularly in addressing complex gas compositions and boundary ambiguities.

A. Dissolved Gas Analysis and Traditional Duval Triangle 1 Method

DGA is a widely utilized technique for assessing the condition of oil—immersed transformers through the analysis of concentrations of key gases dissolved in the transformer oil [2,5]. These gases are generated as byproducts of thermal and electrical faults within the transformer [4]. The conventional

Duval triangle 1 method simplifies fault diagnostics by focusing on three specific gas percentages: methane (CH_4), acetylene (C_2H_2), and ethylene (C_2H_4) [6]. By plotting these gas concentrations on a triangular coordinate system, as presented in Fig. 1, the method identifies fault types based on predefined regions corresponding to various failure modes, including partial discharge (PD), thermal faults (T1—T3), and electrical faults (D1—D2) [6,10]. While the Duval triangle 1 method demonstrates efficacy in numerous cases, its reliance on only three gas percentages constrains its diagnostic scope. Scenarios in which other gases, such as hydrogen (H_2) and ethane (C_2H_6), play a significant role may result in misclassifications [3,9]. Furthermore, the distinct boundaries of the fault regions can lead to ambiguity for data points in proximity to the edges, thereby reducing diagnostic reliability [7].

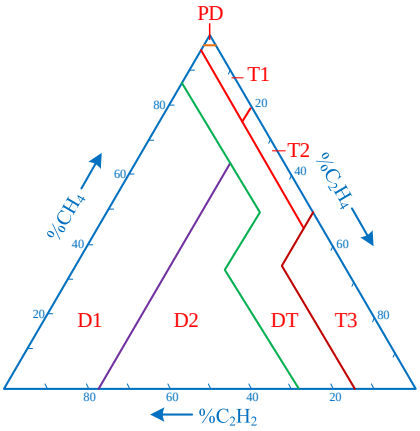


Fig. 1. Duval's triangular coordinate system.

B. Data Preprocessing Layer

To address the limitations of the traditional Duval Triangle 1 method, a data preprocessing layer is implemented to enhance the quality and consistency of input data. This layer is particularly crucial when analyzing datasets involving five key gas percentages: CH_4, C_2H_2, C_2H_4, H_2, and C_2H_6. The preprocessing layer comprises the following steps, as presented in Fig. 2.

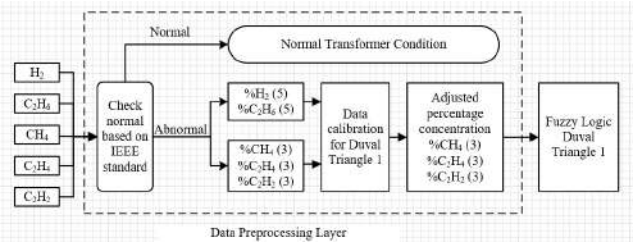


Fig.2. Data preprocessing layer's architecture.

Initial Condition Assessment: The system initially evaluates the transformer's condition based on the IEEE — 2019 standard [2]. If the result indicates a normal condition, the transformer is classified as healthy, and no further processing is necessary. If a fault is detected, the process proceeds to the subsequent step.

1) Gas percentage calculation: Let %X be the percentage of gas X. The preprocessing layer calculates the gas percentages in two stages: For five gases (5): $\%H_2$ and $\%C_2H_6$. For three gases (3): $\%CH_4$, $\%C_2H_4$, and $\%C_2H_2$. This ensures accurate categorization and prioritization of gases based on their diagnostic relevance.

2) Data processing: The calculated gas percentages undergo processing through a data calibration block. This block's mechanism involves the transformation of input percentages from five gases into optimized percentages for the three gases corresponding to the inputs of the DTM. Fig. 3 illustrates the algorithmic architecture of this data calibration block. As depicted, additional conditions derived from extensive studies on operational real — world data [11] have been incorporated to adjust the percentages of the three input gases. These adjustments address the limitations of the Duval triangle method. For instance, the absence of $\%H_2$, which is crucial for diagnosing PD faults, and $\%C_2H_6$, which primarily indicates thermal faults, frequently results in misdiagnosis when $\%H_2$ or $\%C_2H_6$ concentrations are elevated in DGA samples. Through the integration of these additional conditions, the preprocessing layer enhances the accuracy of the fault diagnosis process.

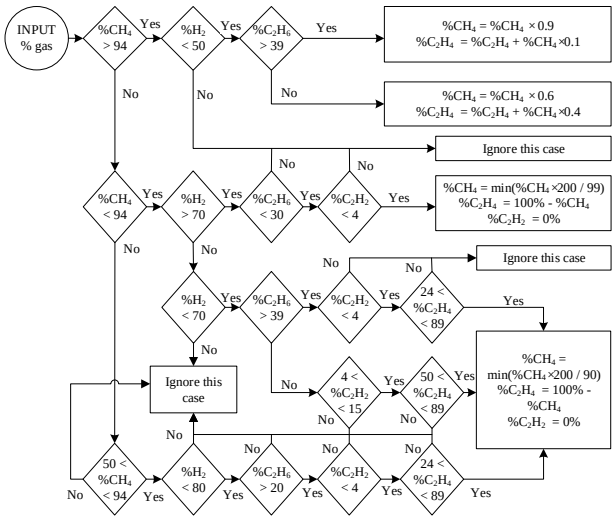


Fig. 3. Data calibration block's flowchart.

3) Optimized output: The output of the preprocessing block yields optimized gas percentages for $\%CH_4$, $\%C_2H_4$, and $\%C_2H_2$. These adjusted values serve as inputs for the Duval triangle 1 method, enhancing the accuracy and reliability of fault classification. By implementing this structured preprocessing workflow, the proposed methodology ensures that the input data is both accurate and optimized for subsequent diagnostic analysis.

C. Integrating Fuzzy Logic and Duval Triangle 1 Method

The integration of fuzzy logic with the Duval triangle 1 method (FL—DTM) provides a robust approach to diagnosing transformer faults by addressing the inherent uncertainties in dissolved gas analysis (DGA). The system's output membership function D, defined within a range of 0 to 10, is categorized into

fault types from F1 (partial discharge) to F5 (arcing), as presented in Table I [6,7].

TABLE I. FAULT CODE FOR OUTPUT MEMBERSHIP FUNCTION OF FL—DTM

Fault DTM method	FL DTM	Output
Partial Discharges (PD)	PD (F1)	0 < D ≤ 2
Thermal faults, T < 300 ℃ (T1)	Low thermal (F2)	2 < D ≤ 4
Thermal faults, 300 ℃ < T < 700 ℃ (T2)	High thermal (F3)	4 < D ≤ 6
Thermal faults, T > 700 ℃ (T3)		
Mixture of thermal and electrical faults (DT)	DT (F4)	6 < D ≤ 8
Discharges of low energy (D1)	Arcing (F5)	8 < D < 10
Discharges of high energy (D2)		

The architecture of the FL—DTM system is illustrated in Fig. 4. It employs a multiple—input single—output (MISO) configuration, where the inputs comprise three optimized gas ratios—%CH_4, %C_2H_4, and %C_2H_2—derived from a preprocessing layer designed to refine raw DGA data [2,8]. The output D represents the diagnostic result based on the Duval Triangle 1 zones, enhanced with fuzzy logic reasoning. The fuzzy inference system incorporates 21 rules (as shown in Fig. 5), constructed using expert knowledge and historical fault datasets [3,9,11]. The Mamdani inference method is utilized to model the relationships between input gas ratios and fault categories, while the centroid method is employed for defuzzification, ensuring accurate and interpretable results. This integration effectively enhances diagnostic accuracy by combining the precision of the Duval triangle method with the flexibility of fuzzy logic to address overlapping and uncertain data points.

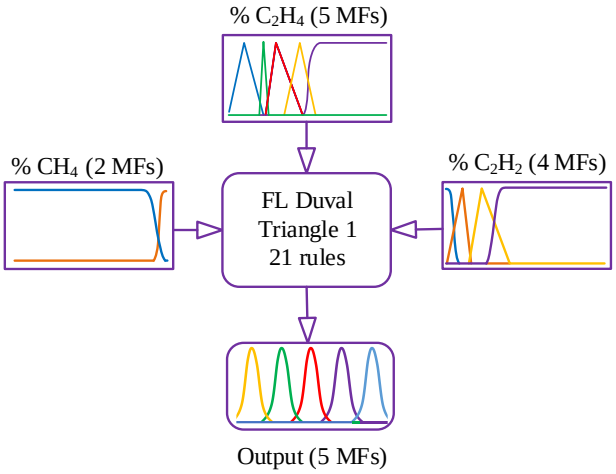


Fig. 4. FL—DTM system's architecture.

The centroid defuzzification method is employed to compute a crisp output by determining the center of gravity of the aggregated fuzzy set. Let D_{output} be the defuzzied diagnostic output. Then D_{output} is defined as follows:

$$D_{output} = \frac{\int \mu_{output}(x).xdx}{\int \mu_{output}(x)dx} \quad (1)$$

Where $\mu_{output}(x)$ represents the aggregated membership function of the output variable, and x represents the continuous range of possible output values.

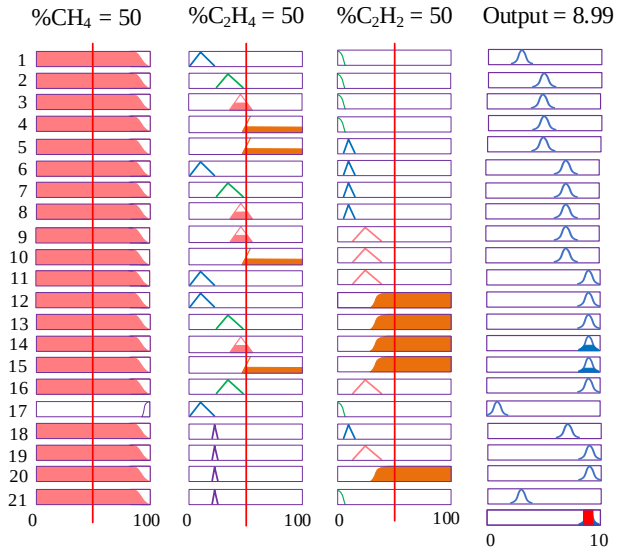


Fig. 5. FL—DTM system's rules.

III. EXPERIMENTAL SETUP AND RESULTS

The DPFD method was tested using historical DGA datasets [11]. The system, integrating a preprocessing layer, fuzzy inference with 21 rules, and Duval triangle 1, showed improved diagnostic accuracy and reliability.

A. *Simulation Model using Matlab/Simulink Software*

The model is simulated in Matlab/Simulink, as shown in Fig. 6, by incorporating all the steps illustrated in Fig. 2, including data preprocessing, fuzzification, and defuzzification, to display the results. The input data processing algorithms were implemented within Matlab function blocks, adhering to the algorithmic flowchart depicted in Fig. 3.

B. *Results and Discussion*

To validate the proposed model, 12 random DGA samples ranging from no fault to severe fault (D) were tested using four diagnostic methods: DTM, DP—DTM, FL—DTM, and DPFD. The results, presented in Table II, demonstrate the accuracy of each method as follows: 4/12 for DTM, 7/12 for FL—DTM, 9/12 for DP—DTM, and 12/12 for DPFD. These results elucidate the efficacy and limitations of each approach. The conventional DTM and FL—DTM methods exhibited reduced performance with data outside the scope of the three primary input gases, particularly affecting diagnoses of PD, T1, and normal conditions (N). While DP—DTM addressed these limitations by rectifying specific deficiencies, it demonstrated inadequate performance in handling data near boundary regions effectively. The DPFD method exhibited superior performance,

overcoming the challenges encountered by the other methods. Its capacity to handle diverse datasets and mitigate boundary—region issues underscores its robustness and reliability in fault diagnosis.

TABLE II. IMPROVED DIAGNOSTIC RESULTS FROM THE PROPOSED METHOD

No.	Input Data					DTM	DP DTM	FL DTM	DPFD	ACT
	H_2	CH_4	C_2H_6	C_2H_4	C_2H_2					
1	6.5	7.2	0.8	1.6	0	T1	N	F2	N	Normal
2	18.9	2.9	0.4	2.3	0	T2	N	F3	N	Normal
3	166	21	38	6	0	T2	PD	F2	F1	PD
4	183	6	0	5	0	T2	PD	F3	F1	PD
5	11	46	155	18	0	T2	T1	F3	F2	T1
6	29	71	158	20	0	T2	T2	F2	F2	T1
7	400	940	210	820	24	T2	T2	F3	F3	T2
8	16	73.3	113.5	194.8	0	T3	T3	F3	F3	T3
9	7.4	19.8	66.1	2.7	4	D1	D1	F4	F4	DT
10	150	130	9	55	30	D2	D2	F4	F4	DT
11	85.2	19.7	2.6	34	32.1	D2	D2	F5	F5	Arcing
12	163	26	7	19	133	D1	D1	F5	F5	Arcing

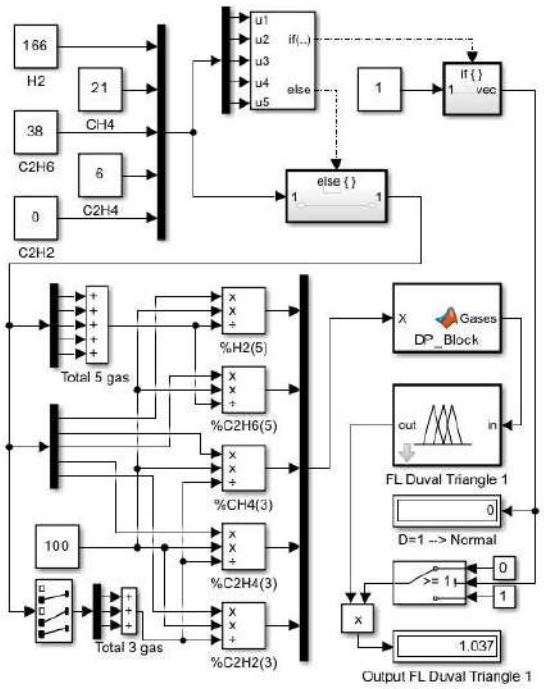


Fig. 6. Proposed system's simulation on Matlab/Simulink.

In comparison to the work by [9], this study presents significant advancements in transformer fault diagnosis. Wani et al. proposed the fault interpretation matrix (FIM) to diagnose incipient transformer faults using DGA. Their approach augmented traditional methods such as Rogers' ratio, IEC ratio, and DTM with FL, achieving diagnostic accuracies of 69.57%, 70.91%, and 78.87%, respectively, as reported in Fig. 7. Despite these enhancements, the standalone fuzzy logic-based methods exhibited inconsistent sensitivity and reliability across fault types, necessitating the integration of results via the FIM to improve diagnostic consistency. The FIM achieved an overall accuracy of 86% when tested on 515 DGA samples. While this integration improved consistency, FIM demonstrated several limitations, including variability in sensitivity among methods, reliance on static rule-based prioritization, and challenges in diagnosing mixed or boundary fault cases.

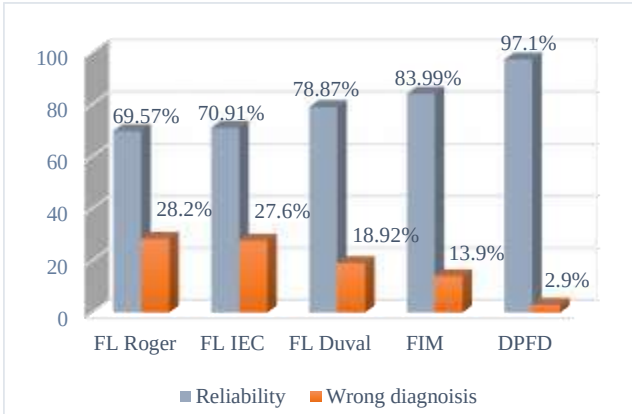


Fig. 7. Accuracy comparison between the proposed DPFD model and the FIM model, as presented in the study [9].

In contrast, the proposed DPFD model addresses these limitations and demonstrates superior diagnostic performance. Evaluated on a larger dataset of 758 real-world DGA samples [11], the DPFD model achieves an accuracy of 97.1%, representing a substantial improvement over the FIM's performance. This significant enhancement highlights the robustness and reliability of the DPFD model in managing diverse and complex datasets while delivering higher diagnostic precision.

By building upon the strengths of previous methodologies and incorporating novel preprocessing and diagnostic strategies, the DPFD model establishes itself as a state-of-the-art tool for practical transformer fault diagnosis and condition monitoring. This research underscores its potential to establish a new benchmark in the field of incipient fault detection.

IV. CONCLUSION

The integration of fuzzy logic with the Duval triangle 1 method, enhanced by a robust data preprocessing layer, significantly advances the accuracy and reliability of transformer fault diagnostics. The proposed DPFD model effectively addresses the limitations of traditional approaches by incorporating optimized input data handling, resolving boundary ambiguities, and improving fault classification in diverse and complex datasets. Simulation results demonstrate the model's exceptional performance, achieving high diagnostic accuracy across various fault scenarios, including those involving overlapping gas compositions and edge—case conditions. This enhancement underscores the importance of combining systematic data preprocessing with advanced fuzzy reasoning to create a comprehensive diagnostic framework. By bridging the gaps in existing methods and delivering consistent, precise outcomes, the DPFD model establishes a strong foundation for future innovations in transformer condition monitoring.

Future endeavors will focus on further refining preprocessing techniques and expanding the applicability of this approach to encompass a wider range of diagnostic challenges

in power system reliability. Moreover, forthcoming research will explore the integration of multi-method DGA interpretation with advanced artificial intelligence algorithms, in conjunction with innovative machine learning techniques such as Quantum Machine Learning, to enhance fault diagnosis capabilities. These complementary strategies aim to significantly improve the accuracy and reliability of power transformer diagnostics.

ACKNOWLEDGMENT

This work was supported by The University of Danang–University of Science and Technology, code number of Project: T2024-02-39.

REFERENCES

[1] M. J. Heathcote, Electric Power Transformer Engineering, 3rd ed., IEEE Power Energy Mag., vol. 11, no. 5, 2013, pp. 94–95.

[2] IEEE C57. 104 - 2019, "IEEE Guide for the Interpretation of Gases Generated in Mineral Oil-Immersed Transformers," November 2019.

[3] S. A. Wani, A. S. Rana, S. Sohail, O. Rahman, S. Parveen, and S. A. Khan, "Advances in DGA based condition monitoring of transformers: A review," Renewable Sustainable Energy Rev., vol. 149, pp. 111347, June 2021.

[4] A. Nanfak, I. Fofana, E. Samuel, C. Hubert Kom, F. Meghnefi, and M. G. Ngaleu, "Traditional fault diagnosis methods for mineral oil–immersed power transformer based on dissolved gas analysis: Past, present and future," IET Nanodielectr., vol. 7, no. 3, pp. 97–130, April 2024.

[5] BSI British Standards, "Mineral oil-impregnated electrical equipment in service. Guide to the interpretation of dissolved and free gases analysis," BS EN 60599, January 2016.

[6] M. Duval, "The duval triangle for load tap changers, non–mineral oils and low temperature faults in transformers," IEEE Electr. Insul. Mag., vol. 24, no. 6, pp. 22–29, November 2008.

[7] S. Mofizul Islam, T. Wu, and G. Ledwich, "A novel fuzzy logic approach to transformer fault diagnosis," IEEE Trans. Dielectr. Electr. Insul., vol. 7, no. 2, pp. 177–186, April 2000.

[8] S. Hmood, A. Abu-Siada, M. A. S. Masoum and S. M. Islam, "Standardization of DGA interpretation techniques using fuzzy logic approach," 2012 IEEE International Conference on Condition Monitoring and Diagnosis, Bali, Indonesia, 2012, pp. 929–932.

[9] S. A. Wani, D. Gupta, G. Prashal, and S. A. Khan, "Smart Diagnosis of Incipient Faults Using Dissolved Gas Analysis-Based Fault Interpretation Matrix (FIM)," Arabian J. Sci. Eng., vol. 44, no. 8, pp. 6977–6985, March 2019.

[10] M. Duval and A. Depabla, "Interpretation of gas–in–oil analysis using new IEC publication 60599 and IEC TC 10 databases," IEEE Electr. Insul. Mag., vol. 17, no. 2, pp. 31–41, March 2001.

[11] N. V. Nga, N. H. Chien, D. Truc, T. D. Tho, N. V. Luc, T. H. Vu, "Research on application of artificial intelligence in diagnosis of potential failures in transformers by dissolved gas analysis method," Univ. Danang J. Sci. Technol., vol. 22, pp. 30–35, October 2024.

Simulation-Based Analysis of Grid Impact from Large-Scale HDEV Charging Infrastructure

1st Thomas Oberliessen
ie3
TU Dortmund University
Dortmund, Germany
0000-0001-5805-5408

2nd Daniel Feismann
ie3
TU Dortmund University
Dortmund, Germany
0000-0002-3531-9025

3rd Florian Klausmann
Institute for Industrial Engineering
Fraunhofer
Stuttgart, Germany
0000-0002-7953-4296

4th Felix Otteny
Institute for Industrial Engineering
Fraunhofer
Stuttgart, Germany
0009-0002-5249-9803

5th Christian Rehtanz
ie3
TU Dortmund University
Dortmund, Germany
0000-0002-8134-6841

Abstract—The integration of large-scale heavy-duty electric vehicle (HDEV) charging infrastructure presents significant challenges for existing power distribution grids. This paper evaluates the grid impact of megawatt-level HDEV charging using a coupled simulation approach. A detailed mobility model from previous work simulates HDEV arrivals, parking durations, and energy requirements at highway rest areas over a year, generating realistic charging profiles. These profiles, serve as input for quasi-static time series power system simulations. The analysis employs SimBench benchmark distribution grids (Rural, Semi-Urban, Commercial, Urban). Three distinct scenarios for connecting the HDEV charging infrastructure are investigated across 36 combined simulation cases. The paper focuses on the impact on transformer utilization, line utilization, and voltage magnitudes, comparing scenarios with and without HDEV charging. Results indicate that while transformer capacity appears generally sufficient to accommodate the additional load, line utilization emerges as a critical bottleneck. The findings highlight that significant grid reinforcement, primarily upgrading distribution line capacities, will be essential for the widespread deployment of HDEV charging.

Index Terms—Heavy-Duty Electric Vehicle, Grid Impact Analysis, Power System Simulation, Distribution Grid, Grid Reinforcement, Power Flow Analysis

I. Introduction and Related Work

The electrification of heavy-duty transport is crucial for reducing greenhouse gas emissions in the transportation sector. While HDEV offer significant environmental benefits, their integration poses unique challenges for power distribution grids due to their high power demand and concentrated charging patterns. Highway rest areas, where heavy duty vehicles typically stop for mandatory breaks, are particularly challenging locations for charging infrastructure deployment, often requiring megawatt-level charging capacity that can significantly impact local distribution grids.

This research was done as part of the HoLa project (FKZ 03EMF0404C), funded by the Federal Ministry for Economic Affairs and Climate Action (BMWK)

Studies on the grid impact of heavy-duty electric vehicles are limited, with most literature focusing on light-duty electric vehicles. Some conclusions might be drawn from high power charging for light-duty vehicles as in [1]–[3]. As for HDEVs specifically, in [4] the authors analyzed the impacts of HDEV depot charging on electricity distribution systems. The study found that depot charging power requirements can be met with current light-duty EV charging levels ($\leq$100 kW per vehicle) and that most substations could accommodate 100 HDEVs without upgrades. Managed charging is highlighted as important for reducing peak loads. The authors in [5] proposed a methodology to model the load profile of high-power charging stations for HDEVs and assessed grid impacts in Norway, finding that a regional substation's capacity was exceeded with a 25% HDEV share, and the thermal limit at 50%. Extending driver breaks and reducing charging power showed potential for peak reduction. In [6] the authors examined the grid capacity impact of public fast charging for a fully electrified long-haul truck fleet in southern Sweden using agent-based simulations and probabilistic load flow analysis, finding that aggregated HDEV charging led to overloads in 6 of 18 primary substation transformers. The authors in [7] evaluated the grid impacts of different charging methods for electric buses in a typical urban European distribution grid, finding that every station charging caused higher simultaneity and voltage drops, with LV charging deemed unsuitable for high power. In [8] the authors presented a systematic methodology for grid impact analysis of HDEV charging stations, using voltage load sensitivity matrix for location identification, showing that poor charging station placement could cause significant voltage sags. Reactive power support from smart chargers was proposed as a mitigation strategy.

Little work has been done on the specific grid impact

979-8-3315-9431-2/25/$31.00 ©2025 IEEE

of HDEV on different grid topologies and varying realistic grid operating conditions. This paper combines a previously developed detailed mobility model for HDEV charging demand with grid utilization simulations across different network topologies, grid utilization scenarios and HDEV charging connection scenarios. The simulation results allow detailed time series analysis of the grid impact of HDEV charging on the grid infrastructure. Such analyses are crucial for integrating grid extension needs of HDEVs into the grid planning process.

II. Methodology for Charging Profile Generation

HDEV charging profiles are derived using a two-part simulation tool, combining a mobility model for HDEV arrivals and energy demand with a charging infrastructure model for simulating charging events. The methodology is developed and described in detail in [11] but will be briefly summarized here.

A. Mobility Model

This model simulates truck arrivals and parking at a highway rest area over one year in minute resolution. Parking occupancy is predicted using regression models linking highway traffic to dynamic rest area occupancy data, considering time lags and different time clusters (e.g., weekday, weekend). Parking durations are assigned based on legal requirements (45min break, 9-11h rest, 45h weekly rest) using probability distributions derived from synthetic trip chains. Driving distances are assigned using synthetic European origin-destination freight flow data and shortest path calculations. This yields time-resolved truck arrivals with associated parking durations and travel distances.

B. Charging Infrastructure Model

This component simulates charging processes based on the mobility model output and defined infrastructure parameters. A fraction of arriving trucks are designated as Battery HDEVs based on an electrification rate. HDEV parameters include energy consumption (e.g., 1.2 kWh/km) and battery capacity E_{batt}. The infrastructure comprises Megawatt Charging System (MCS) for breaks and Night Charging System (NCS) for rest periods, each with a defined number of charging points (CPs) and maximum power P_{CP}. MCS power follows a state of charge (SOC)-dependent curve $\epsilon_{truck}(SOC)$, while NCS power is constant; overall efficiency ρ is applied. A queuing system manages CP contention, with waiting times limited by a SOC-dependent willingness-to-wait.

Charging demand is determined based on arrival SOC and charging strategy. Arrival SOC (SOC_{start_1}) is calculated assuming a drive from full (e.g., 4.5h), consuming energy E_{trip}.

$$SOC_{start\_1} = (E_{batt} - E_{trip})/E_{batt} \times 100 \qquad (1)$$

MCS strategies include full recharge during the break (Break Charging) or charging only needed energy for the next leg plus reserve (On-Demand Charging). The scenario used within this paper utilizes break charging.

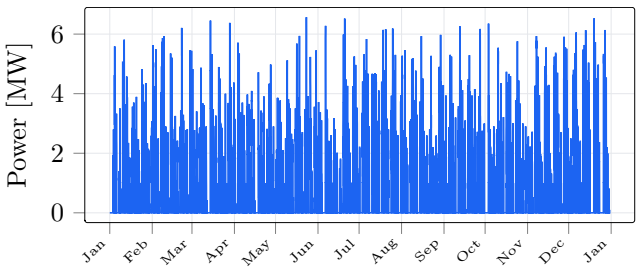


Fig. 1. Yearly HDEV charging profile for a 20% share of HDEV's at a highly frequented highway rest area.

The charging process is simulated in discrete time steps (e.g., 1 minute) using MATLAB/SIMULINK. Charging power $P_{charge}(SOC)$ is calculated considering CP limits, vehicle request, and efficiency.

$$P_{charge}(SOC) = P_{CP} \times \epsilon_{truck}(SOC) \times \rho \qquad (2)$$

Vehicle SOC is updated at each step based on energy transferred $E_{charge}(t)$.

$$SOC(t+1) = SOC(t) + (E_{charge}(t)/E_{batt}) \times 100 \qquad (3)$$

The simulation manages CP occupancy, charging completion (target SOC or duration), minimum charge times, and queuing logic.

The specific charging profile used in this paper generated by the simulation tool is shown in Figure 1 and corresponds to break charging scenario S1a in [11]. It is for a specific highly frequented german highway rest area for a 20% share of HDEV's, representing a significant share of the total heavy-duty truck traffic. The rest area has 93 parking spots, 7 MCS and 36 NCS charging points. The peak power of the charging profile is 6.57 MW and the yearly energy demand is 5022 MWh.

III. Methodology for Grid Utilization

A. Simulation Scenarios

This work's primary goal is to evaluate the impact of HDEV charging integration on various distribution grid types under different operational conditions. Grid-specific data, including topologies and utilization scenarios, is sourced from the SimBench benchmark dataset [9]. Grid topologies studied include Rural, Commercial, Semi-Urban, and Urban network types. Grid utilization scenarios encompass "Today (0)," "Tomorrow (1)," and "The Day after Tomorrow (2)," characterized by increasing penetrations of renewable energy sources and modern loads such as electric vehicles and heat pumps. A scenario combination is denoted as e.g. "Rural 1", which denotes a rural grid with the utilization scenario "Tomorrow". Each scenario includes distinct load and generation time series profiles assigned to individual grid buses. Different

approaches for connecting HDEV charging infrastructure to the grid are considered due to their varying potential impacts, particularly concerning the electrical distance to the High Voltage / Medium Voltage (HV/MV) transformer substation. Three distinct connection scenarios were defined to represent a range of possibilities. The "Best Case" scenario assumes connection directly at the HV/MV substation busbar. The "Median Case" involves connection to a feeder at a mean electrical distance, representing an average connection point along a line, calculated as the mean aggregated line length at the mean bus position of that line. The "Worst Case" simulates connection at the electrical endpoint of the longest feeder in the grid. Combining the four grid types, three utilization scenarios, and three HDEV charging connection scenarios results in a total of 36 distinct simulation scenarios.

B. Grid Utilization Simulation

Grid utilization simulations are performed using the energy system simulator SIMONA [10]. The SimBench grid models and associated utilization scenario data, such as load and generation profiles, are converted into the PowerSystemDataModel format compatible with SIMONA. To isolate the specific impact of HDEV charging, baseline simulations are first established for the 12 combinations of grid type and utilization scenario without considering HDEV charging loads. For each of these 12 baseline combinations, a full one-year time series simulation is conducted. Subsequently, a one-year time series simulation is performed for each of the 36 defined HDEV charging scenarios, incorporating the HDEV charging load at the specified connection point within the respective grid and utilization context. Within each simulation, both baseline and HDEV charging scenarios, the state of the power system is determined for every time step over the simulated year. This involves executing a power flow calculation based on the aggregated net power derived from the specific load and generation profiles assigned to each individual bus. The net complex power injection at each bus i, denoted as $S_i = P_i + jQ_i$, is determined by the sum of complex power consumption $S_{load,l,i}$ from all loads l at bus i plus the sum of complex power generation $S_{gen,k,i}$ from all generators k at bus i, where generation is defined as negative:

$$S_i = \sum_l S_{load,l,i} + \sum_k S_{gen,k,i} \tag{4}$$

where P_i and Q_i are the net active and reactive power injections at bus i, respectively. The core of the simulation is solving the non-linear power flow equations for the network. For each bus i in a network of N buses, these equations relate the net active power injection P_i and reactive power injection Q_i to the bus voltage magnitudes $|V|$ and angles θ, and the network admittances $Y_{ij} = G_{ij} + jB_{ij}$:

$$P_i = \sum_{j=1}^{N} |V_i||V_j|(G_{ij}\cos(\theta_i - \theta_j) + B_{ij}\sin(\theta_i - \theta_j)) \tag{5}$$

$$Q_i = \sum_{j=1}^{N} |V_i||V_j|(G_{ij}\sin(\theta_i - \theta_j) - B_{ij}\cos(\theta_i - \theta_j)) \tag{6}$$

These equations are typically solved iteratively using numerical methods like the Newton-Raphson algorithm to find the unknown voltage magnitudes and angles for each time step. [12]

IV. Results

This section presents the analysis of grid impacts from integrating HDEV charging infrastructure across different grid types and scenarios. The results focus on three key aspects: transformer utilization, line utilization, and voltage magnitudes. Each aspect is evaluated by comparing scenarios with and without HDEV charging to isolate the specific impact of the charging infrastructure. The analysis considers four grid types (Rural, Semi-urban, Commercial, and Urban) and three temporal scenarios (today/0, tomorrow/1, and the day after tomorrow/2) to provide a comprehensive assessment of grid impacts.

A. Transformer Utilization

A fundamental precondition for integrating HDEV charging locations is the availability of sufficient transformer capacity. The transformer capacity is defined as t and expressed as a percentage of the rated power. To assess the impact of HDEV charging, we compare the transformer utilisation between the base scenario and the HDEV charging scenario. The difference between these scenarios, denoted as Δ_t, is calculated as the difference between the transformer capacity in the HDEV charging scenario and the base scenario:

$$\Delta_t = t_{\text{HDEV}} - t_{\text{base}} \tag{7}$$

To ensure N-1 security in the distribution grid, the maximal permissible transformer utilization is limited to 50% of the rated power. Figure 2 presents an analysis of transformer utilization across the four grid types (Rural, Semi-urban, Commercial, and Urban) and three temporal scenarios (today, tomorrow, and the day after tomorrow). The figure consists of two panels: the upper panel showing absolute transformer utilization (t) and the lower panel showing the relative change in utilization (Δ_t). The box plots reveal distinct patterns across grid types, with rural grids exhibiting the highest variability in both absolute utilization and relative changes. Interestingly, it reduces extreme outliers in the rural and semiurban scenarios 1 and 2. When investigating the utilisation delta between the base and the HDEV charging scenarios in the lower plot it becomes visible that the integration of HDEV charging can even significantly decrease transformer utilizations. Values range from 10 % to -10 % for the rural

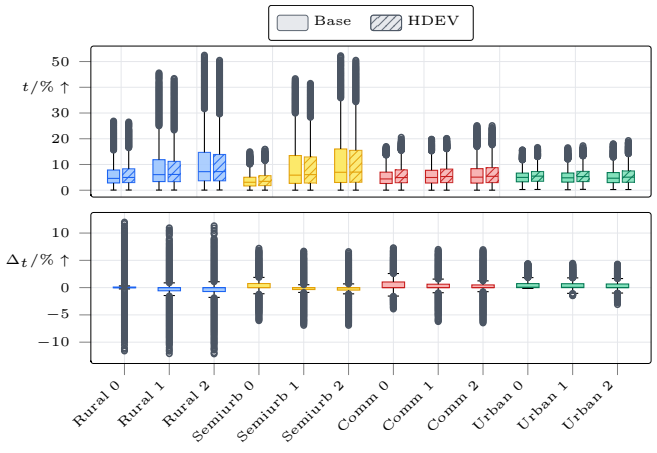


Fig. 2. Absolute trafo utilization and trafo utilization delta after inclusion of HDEV charging for different simulation scenarios.

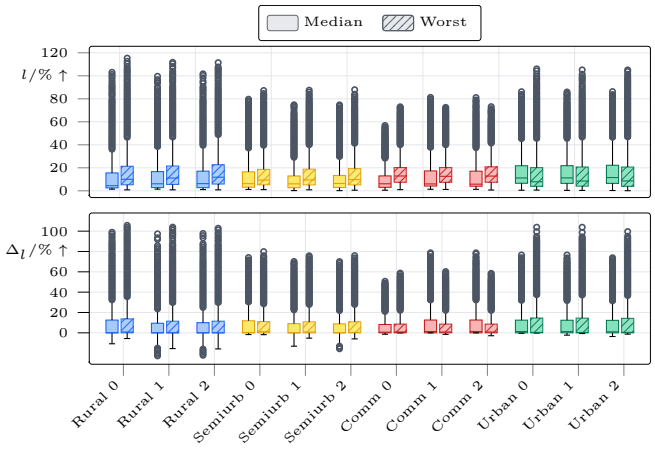


Fig. 3. Line utilization and line utilization delta after inclusion of HDEV charging for different simulation scenarios.

grids and 5 % to -3 % for the urban grids. Negative utilisation changes are explained by the available photovoltaic generation. During net generation the transformer current is reversed and flows into the high voltage grid. Increasing the load within the grid reduces the reversed current and therefore reduces transformer utilization. This indicates that there can be some synergystic effects between HDEV charging and increasing levels of photovoltaic generation. Relative differences between the grid topologies is explained by the different installed hardware. Rural grids tend to have lower rated transformers. In summary transformer with 5 % to 10 % remaining capacity depending on the transformer rating can presumably handle the addition of highly frequented 20 % HDEV charging stations.

B. Line Utilization

Since HDEV charging infrastructure represents a substantial concentrated load, the current ratings of the lines have to be able to handle the additional stress. The line utilization is defined as l and expressed as a percentage of the rated current. To assess the impact of HDEV charging, we compare the line utilization between the base scenario and the HDEV charging scenario. The difference between these scenarios, denoted as Δ_t, is calculated as the difference between the line utilization in the HDEV charging scenario and the base scenario:

$$\Delta_l = l_{\text{HDEV}} - l_{\text{base}} \tag{8}$$

As for transformer utilization, the generally permissable line utilization is 50 % to ensure N-1 security. Figure 3 presents the absolute line utilization and the delta in line utilization across the different simulation scenarios. The utilizations are shown for each line in the path between the HDEV charging location and the HV/MV transformer. We only show the "Median Case" and "Worst Case" scenarios, as in the "Best Case" scenario the charging infrastructure is directly connected to the substation so it is not connected via existing distribution grid lines. For all topology/scenario combinations we can observe absolute line utilizations exceeding 50 %. Utilizations even exceed 100 % for the rural grids due to smaller thermal ratings of the cables. The delta of the line utilizations in the lower subplot show that indeed the HDEV charging induces most of the load on the distribution grid lines. In summary the results show that connection of HDEV charging to existing medium voltage grid is challenging due to limited capacity of distribution grid cables and would require sufficient grid extension.

C. Voltage Magnitudes

Grid operators need to ensure voltage levels stay within +/- 10% of rated voltage. We compare voltage drops between base and HDEV charging scenarios, where:

$$\Delta_v = v_{\text{HDEV}} - v_{\text{base}} \tag{9}$$

Figure 4 presents the bus voltages and the delta bus voltages across the different simulation scenarios. The voltages are shown for each bus in the path between the HDEV charging location and the HV/MV transformer. We only show the "Median Case" and "Worst Case" scenarios, as in the "Best Case" scenario the charging infrastructure is directly connected to the substation so it is not connected via existing distribution grid lines and therefore no considerable voltage drop occurs. The rural scenarios show very high voltage drops in absolute terms but also compared to the different grid topologies. A major reason is the higher line length typical for rural areas, which lead to significantly higher voltage drops. With voltage drops up to 10 % to 15 % depending on the charging connection points a connection without grid reinforcement would not be possible due to exceeding the voltage limits of +/- 10 %. For other grid topologies, the voltage delta reaches up to 2 % for the median connection

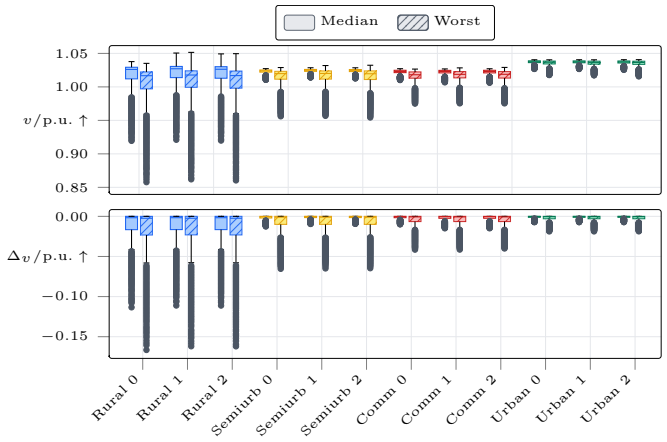


Fig. 4. Bus voltage magnitude and bus voltage magnitude delta after inclusion of HDEV charging for different simulation scenarios.

scenario, which should not be problematic for secure grid operation. Across the worst case scenario, the voltage drop for semiurban is significantly higher than for the commercial scenario which is significantly higher than the one for the urban scenario. While a voltage drop of up to 7 % of the semiurban grids would be critical this problem might be relieved by an increase in transmission capacity. Since the line utilization indicated a need for grid extension, the additional transmission capacity would lower the voltage drops significantly. In summary the voltage drops should be in an acceptable range after the transmission capacity increase for most scenarios.

V. Conclusion and Outlook

This paper presents a methodology for analyzing the grid integration of HDEV charging infrastructure, combining a detailed model for HDEV charging demand with grid utilization simulations across different network topologies and grid utilization scenarios. The analysis evaluates three critical grid parameters: transformer utilization, line utilization, and voltage magnitudes, across four grid types (Rural, Semi-urban, Commercial, Urban), three grid utilization scenarios with differing amount of renewables and modern loads and two HDEV charging connection scenarios. The results reveal that transformer capacity is generally sufficient to handle the HDEV charging loads, with utilization changes ranging from -10% to +10%. However, line utilization emerges as the primary bottleneck, with many scenarios exceeding the 50% N-1 security limit and rural grids even surpassing 100% utilization. Voltage drops are most severe in rural areas (up to 15%), while urban and commercial grids show more manageable impacts (up to 2-7%). These findings suggest that while transformer capacity is often adequate, significant grid reinforcement, particularly of distribution lines, will be necessary to accommodate large-scale HDEV charging infrastructure. Future work should focus on developing detailed grid reinforcement strategies and exploring the potential of smart charging approaches to mitigate grid impacts. Smart charging can reduce the peak load of HDEV charging significantly and therefore reduce the need for grid reinforcement. As this work focused on 20 % electrification rate, future work should also consider the impact of higher electrification rates and analyze if an MV grid connection is still feasible. Additionally, the interaction between HDEV charging and increasing renewable energy penetration warrants further investigation, as initial results indicate potential synergies in transformer utilization.

References

[1] K. Yunus, H. Zelaya De La Parra, and M. Reza, “Distribution Grid Impact of Plug-In Electric Vehicles Charging at Fast Charging Stations Using Stochastic Charging Model,” in Proc. 2011 IEEE PES Innov. Smart Grid Technol. Middle East (ISGT Middle East), Jan. 2011

[2] X. Kong, S. T. Lee, A. A. Abd Musa, and C. S. Tan, “A Study on the Impacts of DC Fast Charging Stations on Power Distribution System,” in Proc. 2014 IEEE Conf. Energy Convers. (CENCON), Johor Bahru, Malaysia, Oct. 2014

[3] B. Pea-da and S. Dechanupaprittha, “Impact Analysis of Fast Charging to Voltage Profile in PEA Distribution System by Monte Carlo Simulation,” in Proc. 2015 7th Int. Conf. Inf. Technol. Electr. Eng. (ICITEE), Chiang Mai, Thailand, Oct. 2015

[4] B. Borlaug, M. Muratori, M. Gilleran, D. Woody, W. Muston, T. Canada, A. Ingram, H. Gresham, and C. McQueen, “Heavy-Duty Truck Electrification and the Impacts of Depot Charging on Electricity Distribution Systems,” Nature Energy, vol. 6, no. 6, pp. 673–682, Jun. 2021.

[5] K. K. Fjaer, V. Lakshmanan, B. N. Torsaeter, and M. Korpas, “Heavy-Duty Electric Vehicle Charging Profile Generation Method for Grid Impact Analysis,” in Proc. 2021 Int. Conf. Smart Energy Syst. Technol. (SEST), Vaasa, Finland, Sep. 2021, pp. 1–6.

[6] A. Jansson, M. Ingelström, O. Samuelsson, and F. J. Márquez-Fernández, “Grid Capacity Impact from the Charging of Electrified Long-Haul Trucks,” in Proc. 2025 IEEE Texas Power Energy Conf. (TPEC), College Station, TX, USA, Feb. 2025, pp. 1–6.

[7] D. Stahleder, D. Reihs, S. Ledinger, and F. Lehfuss, “Impact Assessment of High Power Electric Bus Charging on Urban Distribution Grids,” in Proc. 45th Annu. Conf. IEEE Ind. Electron. Soc. (IECON), Lisbon, Portugal, Oct. 2019, pp. 4304–4309.

[8] X. Zhu, B. Mather, and P. Mishra, “Grid Impact Analysis of Heavy-Duty Electric Vehicle Charging Stations,” in Proc. 2020 IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. (ISGT), Washington, DC, USA, Feb. 2020, pp. 1–5.

[9] S. Meinecke, D. Sarajlić, S. R. Drauz, A. Klettke, L.-P. Lauven, C. Rehtanz, A. Moser, and M. Braun, “SimBench—A Benchmark Dataset of Electric Power Systems to Compare Innovative Solutions Based on Power Flow Analysis,” Energies, vol. 13, no. 12, p. 3290, Jan. 2020.

[10] J. Hiry, C. Kittl, D. Sen Sarma, T. Oberließen, S. Peter, D. Feismann, J. Bao, J. Hohmann, and M. Staudt, “SIMONA - A Discrete-Event Distribution Grid Simulation Environment,” Software, Version 3.0.0, Aug. 2023. Available: https://github.com/ie3-institute/simona

[11] F. Klausmann and F. Otteny, “Simulation-Based Tool for Strategic and Technical Planning of Truck Charging Parks at Highway Sites,” World Electric Vehicle Journal, vol. 15, no. 11, p. 521, Nov. 2024.

[12] J. J. Grainger and W. D. Stevenson, Power System Analysis. New York: McGraw-Hill, 1994.

A New Design of a Quadruped Robot for Teaching

Ba-Phuc Huynh*
Department of Robotics Engineering, Faculty of Electronics and Telecommunications, VNU University of Engineering and Technology, Vietnam National University,
Hanoi, Vietnam
phuchb@vnu.edu.vn

Xiem Hoang Van
Department of Robotics Engineering, Faculty of Electronics and Telecommunications, VNU University of Engineering and Technology, Vietnam National University,
Hanoi, Vietnam
xiemhoang@vnu.edu.vn

Minh Dinh Bao
Department of Robotics Engineering, Faculty of Electronics and Telecommunications, VNU University of Engineering and Technology, Vietnam National University,
Hanoi, Vietnam
minhdinh@vnu.edu.vn

***Abstract*—Quadruped robots are increasingly favored for their ability to handle uneven terrain where wheeled robots fall short. This study presents a compact, low-cost quadruped robot designed for educational and research purposes, featuring high-torque servo motors and joints optimized for flexibility and space efficiency. A dual-layer operating system integrating multithreading and multiprocessing enables smooth coordination. An experimental prototype confirms the design's practicality in teaching and integrated technology research.**

Keywords—quadruped robot, dual-layer robot operating system

I. INTRODUCTION

Quadruped robots have drawn growing attention for their superior mobility over rough terrains, where wheeled robots often struggle [1]-[2]. Their ability to traverse uneven ground makes them ideal for critical tasks like search and rescue, inspections, and exploration in areas inaccessible to humans [3]. Beyond practical uses, they are also valuable tools in education and research, offering hands-on experience in robotics, control systems, and AI-related technologies [4]-[6].

The evolution of quadruped robots at Boston Dynamics, exemplified by models like BigDog and Spot, has demonstrated remarkable progress in areas such as dynamic stability, energy-efficient locomotion, and autonomous navigation. These robots combine sophisticated control systems, terrain-optimized mechanical designs, and state-of-the-art sensors to achieve high-performance mobility. BigDog [7], originally developed for military use, was a trailblazer in transporting heavy payloads across uneven and challenging landscapes. In contrast, Spot [8] features a more compact and adaptable form, tailored for commercial and industrial tasks like inspections and mapping, showcasing its versatility across a range of real-world applications.

Beyond Boston Dynamics, other research initiatives have contributed significantly to the field. The MIT Mini Cheetah [9], for instance, represents a shift toward lightweight, agile, and cost-effective quadrupeds capable of fast and nimble motion without compromising functionality. Meanwhile, Unitree Robotics has played a key role in advancing quadruped platforms for both academic and practical use. Their Unitree A1 [10] stands out for delivering strong performance at a competitive price point. Similarly, Xiaomi has entered the quadruped robot space with its CyberDog 2 [11], an enhanced version of its predecessor. Designed with affordability and advanced AI integration in mind, the CyberDog 2 offers valuable potential for educational and research applications in robotics.

Quadruped robots have become an invaluable resource in the fields of education and research, offering a hands-on platform for learning robotics fundamentals, developing locomotion algorithms, and testing sophisticated control strategies. These robots allow students and researchers to engage directly with advanced topics such as gait generation, balance regulation, and dynamic mobility. By interacting with physical robotic systems, learners gain a deeper understanding of how artificial intelligence integrates with hardware, how motion control can be refined through real-world experimentation, and how sensor-based navigation systems operate. This experiential approach not only demystifies complex theories but also encourages creativity and fosters problem-solving skills-essential traits for addressing real-world robotics challenges.

Among the many educational platforms available, LoCoQuad [12] stands out for its affordability and accessibility. Designed with low-cost MG90S servo motors and 3D-printed parts, LoCoQuad emphasizes modularity and torque efficiency. Its user-friendly design enables easy assembly and customization, making it an ideal entry point for educators and researchers aiming to explore robotics without substantial investment. This balance of functionality and cost positions LoCoQuad as a powerful tool for hands-on experimentation and skill development.

Another notable platform is RealAnt [13], an open-source quadruped robot tailored for reinforcement learning applications. Featuring a lightweight design and high-performance servos, RealAnt is built for durability and consistency, particularly in demanding learning environments. Its open architecture allows seamless transition between simulation environments and physical hardware, providing learners and researchers with a practical means to apply and refine machine learning algorithms in real-world scenarios.

This work has been supported by VNU University of Engineering and Technology under project number CN24.05.

The Q-Robot [14] is a more recent addition to the landscape of educational quadrupeds. Built entirely from budget-friendly electronic components and 3D-printed PLA parts, it supports ROS and Wi-Fi connectivity. Q-Robot is designed with simplicity and cost-effectiveness in mind, making it a strong candidate for classroom environments and self-guided robotics education.

Lastly, PADWQ [15] represents an advanced yet accessible option in the open-source quadruped space. Built with standard components and 3D-printed materials, it incorporates semi-direct drive joints and a GPU-supported onboard computer. This allows PADWQ to perform complex tasks such as adaptive terrain navigation and motion planning, demonstrating how low-cost materials and thoughtful design can support high-level robotic capabilities.

Despite considerable progress in the field, achieving an optimal balance among cost, system complexity, and functional capabilities continues to pose a significant challenge in the development of quadruped robots intended for educational and research applications. High-performance systems typically entail prohibitive costs, whereas more affordable alternatives often suffer from limited robustness and functionality. In response to this issue, the present study proposes the design and development of a novel quadruped robot that seeks to bridge this gap by prioritizing simplicity, cost-effectiveness, and the integration of essential features, including a compact, servo-driven mechanical architecture and an efficient dual-layer operating system (DLOS). Furthermore, a sensorless landing force estimation module, which originally developed by the authors in a prior study [16], is incorporated into the operating system framework. This integration obviates the need for dedicated force sensors, thereby simplifying the mechanical design, reducing overall costs, and enhancing the robot's operational resilience, particularly in wet environments.

The rest of this paper is organized as follows: Section II outlines the overall design of the quadruped robot. Section III introduces DLOS. Section IV covers experimental results, and Section V concludes with key findings and future research directions.

II. Model Design

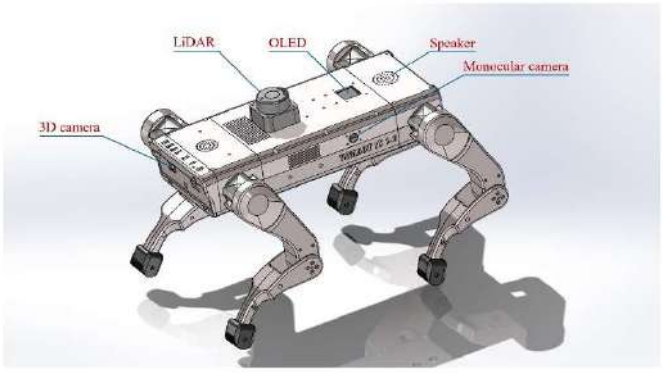


Fig. 1. The 3D CAD of the quadruped robot.

Fig. 1 shows the 3D model of a 4.2 kg quadruped robot designed with 3D-printed parts to reduce cost. Each of its four legs has three degrees of freedom powered by serial bus servo motors, with a compact and simple mechanical structure that maintains full mobility. The robot measures 390 mm in length, 290 mm in width, and its body height ranges from 115 mm to 278 mm during operation. Its vision system includes a front 3D camera and three monocular cameras on the sides and rear, capable of real-time object and terrain recognition, and optionally integrated with a rear-mounted LiDAR for SLAM.

As shown in Fig. 2, the power supply (12.6 VDC, 10 Ah) runs along the underside, with control and function boards stacked above, and two ultrasonic sensors mounted underneath at the front and back for terrain detection. Additional components inside the upper body cover an IMU for balance, GPS for navigation, an OLED display, speakers, and monocular cameras for visual input.

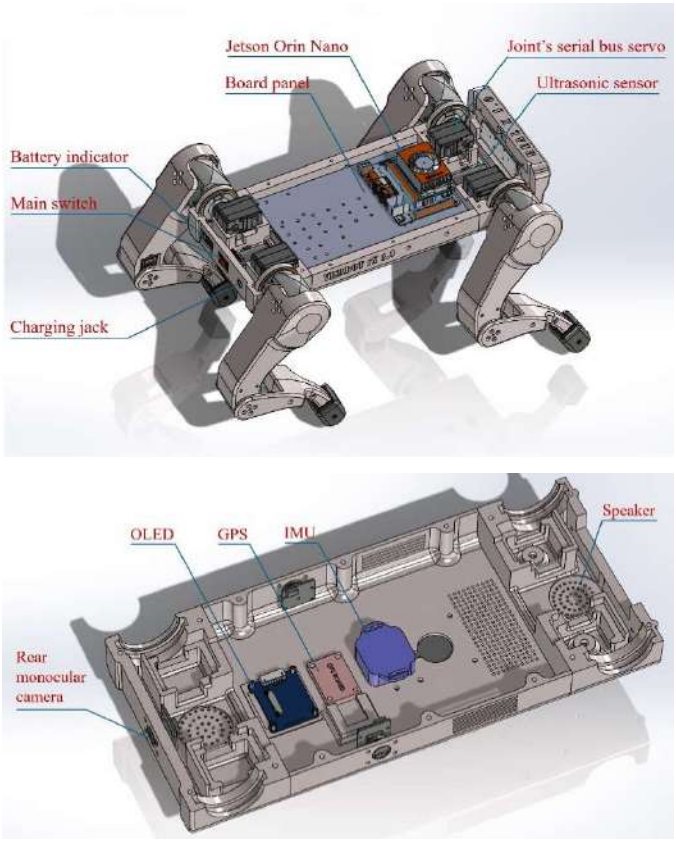


Fig. 2. Component Layout.

Fig. 3 illustrates the robot body design, split into two halves measuring 160 × 365 × 36 mm each, with the thinnest part only 3 mm thick. Made from Hyper PLA for high-speed, accurate, and durable 3D printing, the upper and lower body weigh 575 g and 615 g respectively. The upper body features ventilation holes and side fans placed above the Jetson Orin Nano board to maintain temperatures below 40 ℃, while the lower body is reinforced for better structural strength. Components are tightly fitted, and the upper and lower sections are joined using 14 M3 bolts, with threaded inserts embedded in the plastic for durability.

Each leg has three degrees of freedom: hip pitch, hip roll, and knee pitch. Servos are compactly arranged, supported by body-embedded bearings to handle dynamic loads. The servos

measure 45.22 × 24.72 × 35 mm, weigh 74.5 g, operate at 9–12.6 V, and deliver 50 kg·cm stall torque with 12-bit encoders for accurate positioning and load sensing. The 75 g hip joint made from ABS, precisely housing the hip roll servo. The upper leg, 139 mm long and 155 g, with space for the knee servo, printed from Hyper PLA. The lower leg, 139 mm long and 70 g, also made from Hyper PLA, and connected to the upper leg via the knee servo. The foot is printed with TPU plastic to ensure smoother steps. and prevent slipping.

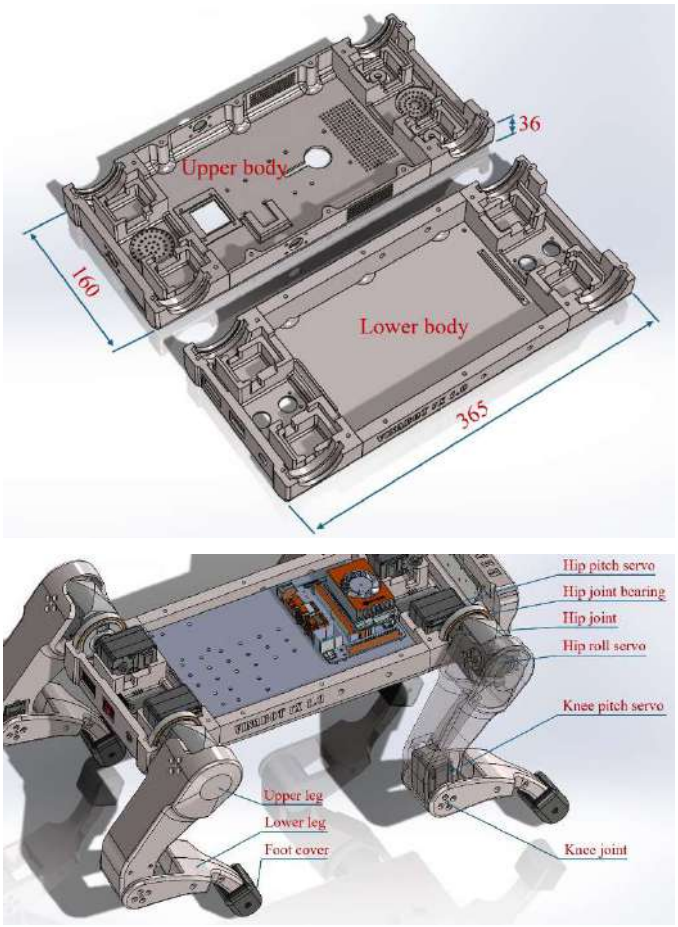


Fig. 3. Body design.

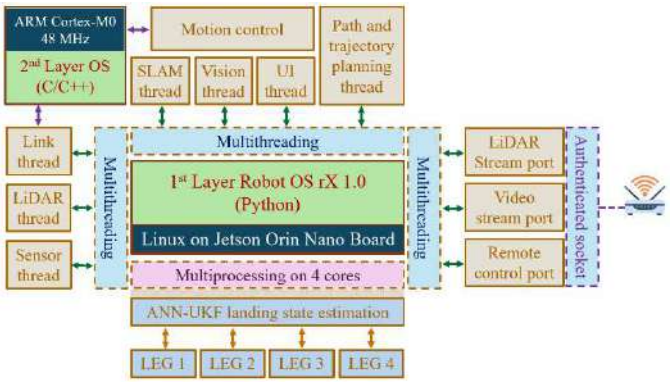


Fig. 4. The dual-layer robot operating system.

III. 4 DLOS Architecture

Fig. 4 shows the DLOS of the robot. The first layer, built in Python on Linux and running on the Jetson Orin Nano, handles core control tasks, while the second layer, coded in C/C++ on a 32-bit Arduino, processes IMU data for balance and motion. This division allows quicker reactions to terrain changes, with both layers communicating via UART at 1 Mbps. The data packet format is described in Fig. 5.

The first-layer OS communicates with the second-layer OS by sending control instructions or data acquisition commands. Each request packet starts with a 1-byte start code, which is always set to 0x1C. Following that is a single byte representing the command code. Based on the nature of the command, additional data might be included. If so, the third byte in the sequence indicates how many data bytes follow. The packet ends with a 1-byte checksum, computed from all preceding bytes in the packet. Similarly, when the second-layer OS sends a response back to the first-layer OS, it uses the same packet structure. The key distinction is that the command code byte is replaced with a response code byte. The accompanying data bytes contain information related to the robot's balance system.

The request data packet's format (From 1st layer OS to 2nd layer OS):

Start code	Command code	Data size	Data	Checksum
1 byte (0x1C)	1 byte	1 byte	0 – 255 bytes	1 byte

Transmission order → Optional bytes

The response data packet's format (From 2nd layer OS to 1st layer OS):

Start code	Response code	Data size	Data	Checksum
1 byte (0x52)	1 byte	1 byte	0 – 255 bytes	1 byte

Transmission order → Optional bytes

Fig. 5. The data format transmitted between two robot operating system layers.

The second-layer operating system is primarily tasked with processing IMU data to support the robot's balance, whereas the first-layer OS handles more computationally intensive functions. To maintain high performance and seamless operation, the system leverages a combination of multi-threading and multi-processing techniques. Multi-threading is utilized to manage communication with the robot's peripheral devices, while multi-processing is reserved for executing sensorless landing force estimation across all four legs—a task that demands substantial computing resources.

This sensorless landing force estimation technique, detailed in the author's previous work [16], employs a hybrid approach combining an artificial neural network (ANN) with an Unscented Kalman Filter (UKF). It allows the estimation of landing force magnitudes without relying on physical force sensors in the robot's feet. Each estimator operates in a separate process, taking servo angles and torque load readings as inputs. The resulting output provides crucial data on the impact forces encountered between the robot's feet and the terrain or obstacles.

By eliminating force sensors, the system design benefits from reduced hardware complexity and lower maintenance or replacement costs, all while retaining the capability to gather essential force information. This ensures accurate gait and trajectory adjustments during locomotion.

Fig. 6. Prototype of the robot made with 3D printing

Fig. 7. Testing process of body balance feature.

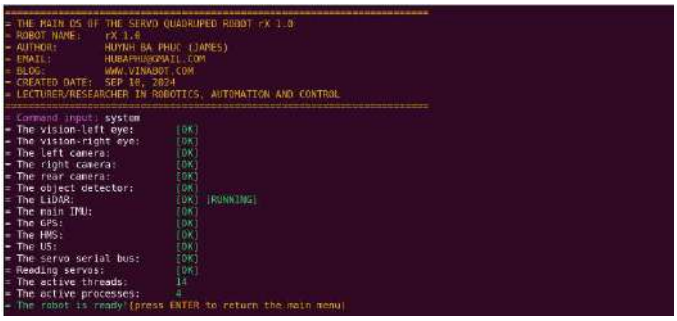


Fig. 8. The command line interface of the DLOS. In the picture is the command to check system information.

IV. Experiment

Fig. 6 shows the prototype of the quadruped robot, weighing 4.2 kg with dimensions of 390 mm × 290 mm and a height range of 115–278 mm. It can walk at 0.06 m/s and perform basic motions like walking, standing, sitting, and predefined actions. The robot's components and their key technical specifications are summarized in Table 1. Despite a production cost under $2,000, it integrates advanced technologies suitable for education and research. Fig. 7 demonstrates its balance test on a tilting board, where calibrated IMU data enables it to maintain roll and pitch angles with 0.1 ° precision. Fig. 8 displays the robot's command-line interface, which allows users to monitor system and sensor states, check vision and servo status, control basic movements, and access APIs for integrating custom control algorithms and research applications.

TABLE I. Technical specifications of components

Device	Parameters	Scope of application
Battery	Li-Po 12.6V 10Ah.	Power supply.
Jetson Orin Nano	CPU: 6-core Arm® Cortex®-A78AE v8.2 64-bit 1.5 GHz, 1.5 MB L2 + 4 MB L3; GPU: 1024-core 625 MHz N-VIDIA Ampere architecture with 32 Tensor Cores; AI performance: 40 TOPS; Memory: 8 GB 128-bit LPDDR5, 68 GB/s; Storage: 128 GB NVMe SSD; Power: 15 W.	1st layer OS.
Arduino Nano 33 IoT	MCU: SAMD21 48 MHz Cortex®-M0+ 32-bit low power ARM, 256 KB SRAM, 1 MB flash; Size: 18 ×45 mm; Weight: 5 g.	2nd layer OS.
IMU WT901C	Accelerometer: X, Y, Z, 3-axis ±16 g, Accuracy: 0.01 g, Resolution: 16 bit, Stability: 0.005 g; Gyroscope X, Y, Z, 3-axis ±2000 °/s, Resolution: 16 bit, Stability: 0.05 °/s; Magnetometer X, Y, Z, 3-axis ±4900 μT, 0.15 μT/LSB typ. (16-bit); Angle/ Inclinometer X, Y, Z, 3-axis X, Z-axis: ±180 °Y ±90 °(Y-axis 90 °is singular point), Accuracy: X, Y-axis: 0.05 °, Z-axis: 1 °(after magnetic calibration)	Balancing.
GPS WTGPS+BD	Signal reception: BDS/GPS/GLONASS /GALILEO/QZSS/SBAS	Positioning.
RPLIDAR C1	Ranging distance: 0.05~12.00 m (white target, 70% reflectivity); 0.05~6.00 m (black target, 10% reflectivity); Ranging frequency: 5000 Hz; Scanning range: 360 °; Scanning frequency: 8~12 Hz; Angular resolution: 0.72 °; Scanning accuracy: ±30 mm; UART 460800 bps; Size: 55.6 ×55.6 ×41.3 mm.	SLAM
Ultrasonic sensor	Ultrasonic frequency: 40 kHz; Measuring range: 2~350 cm; Resolution: 1 cm; Measurement angle: 15 degree; Trigger signal: 10 uS TTL; Echo signal: TTL; Size: 20 ×40 mm.	Measure body height and terrain surface
IMX219-83 Stereo Camera	Sensor: Sony IMX219, 8 MP; Resolution: 3280 ×2464 (per camera); Size: 24 ×85 mm;	Main vision for object detection, SLAM.
IMX335 5 MP USB Camera	Sensor: IMX335, 5 MP; Resolution: 2592 ×1944; Size: 25 ×24 mm.	Side and rear vision.
1.5-inch RGB OLED	Controller: SSD 1351; Resolution: 128(H) RGB ×128(V); Pixel size: 0.045(H) ×0.194(V) mm; Display color: 65 K color; Dimension: 26.855 × 26.855 mm.	Display control information.
Serial bus servo	Running degree: 360 °; Gear ratio: 1/345; Operating voltage: 12 V; No load speed: 75 RPM; Stall torque: 50 Kgf.cm; No load current: 330 mA; Angle sensor: 12 bits magnetic coding; UART 1 Mbps; Output: position, velocity, load torque; Size: 45.22 × 24.72 ×35 mm; Weight: 74.5 g.	Joint actuator.

V. Conclusion and Future Work

This paper presents a compact, low-cost quadruped robot designed for easy 3D printing in lab settings, while still supporting high-performance hardware for educational and research use. It features a DLOS to aid in developing control solutions. The focus is on overall structure and concept rather than detailed component optimization, which is suitable for the robot's small size. Future work will enhance both the DLOS and design, aiming to offer an open-source platform for the research community

Acknowledgment

This work has been supported by VNU University of Engineering and Technology under project number CN24.05.

References

[1] P. Biswal and P.K. Mohanty, "Development of Quadruped Walking Robots: A Review," Ain Shams Engineering Journal, vol. 12, pp. 2017–2031, 2021.

[2] M.S. Lopes, A.P. Moreira, M.F. Silva, and F. Santos, "A Review of Quadruped Manipulators," In: Moniz, N., Vale, Z., Cascalho, J., Silva, C., Sebasti ão, R. (eds) Progress in Artificial Intelligence. EPIA 2023, Lecture Notes in Computer Science, pp. 14115, 2023.

[3] A. Majithia, D. Shah, J. Dave, et al, "Design, Motions, Capabilities, and Applications of Quadruped Robots: A Comprehensive Review," Frontiers in Mechanical Engineering, vol. 10, 2024.

[4] H. Taheri and N. Mozayani, "A Study on Quadruped Mobile Robots," Mechanism and Machine Theory, vol 190, pp. 105448, 2023.

[5] Y. Fan, Z. Pei, C. Wang, M. Li, Z. Tang, and Q. Liu, "A Review of Quadruped Robots: Structure, Control, and Autonomous Motion," Advanced Intelligent Systems, vol 6, pp. 2300783, 2024.

[6] A. Hamrani, M. M. Rayhan, T. Mackenson, D. McDaniel, and L. Lagos, "Smart Quadruped Robotics: A Systematic Review of Design, Control, Sensing and Perception," Advanced Robotics, pp. 1–27, 2024.

[7] M. Raibert, K. Blankespoor, G. Nelson, R. Playter and the BigDog Team, "BigDog, the Rough-Terrain Quadruped Robot," Boston Dynamics, Waltham, MA 02451, USA, 2008.

[8] Spot by Boston Dynamics. Available online: https://www.bostondynamics.com/products/spot (accessed on 06 December 2024).

[9] J. Chu, "Mini cheetah is the first four-legged robot to do a backflip," MIT News Office, 2019. Available online: https://news.mit.edu/2019/mit-mini-cheetah-first-four-legged-robot-to-backflip-0304 (accessed on 06 December 2024).

[10] "Unitree A1," Available online: https://www.unitree.com/a1 (accessed on 06 December 2024).

[11] "CyberDog 2," Available online: https://www.mi.com/cyberdog2 (accessed on 06 December 2024).

[12] M. Bernal and J. Civera: LoCoQuad, "A Low-Cost Arachnoid Quadruped Robot for Research and Education," arXiv:2003.09025, 2020.

[13] R. Boney, J. Sainio, M. Kaivola, A. Solin, and J. Kannala: RealAnt, "An Open-Source Low-Cost Quadruped for Education and Research in Real-World Reinforcement Learning," arXiv:2011.03085, 2022.

[14] C.F. Joventino, J.H.M. Pereira, J.A. Fabro, J. Lima, and A.S. de Oliveira, "A Versatile Opensource Quadruped Robot Designed for Educational purposes," In: ROBOT 2022: Fifth Iberian Robotics Conference, Lecture Notes in Networks and Systems, vol. 589, 2023.

[15] J. Kim, T. Kang, D. Song, and S-J. Yi, "Design and Control of an Open-Source, Low Cost, 3D Printed Dynamic Quadruped Robot," Applied Sciences, vol. 14, pp. 11330, 2024.

[16] B-P Huynh and J. Bae, "ANN-UKF-based estimator for landing forces in quadruped robots," International Journal of Intelligent Robotics and Applications, 2024.

Improving the Reliability of Oil—Immersed Power Transformer Fault Diagnosis Based on the Evaluation of Dissolved Gas Component Input Vectors

Huy Vu Tran
Faculty of Electrical Engineering
The University of Danang - University of Science and Technology
Da Nang, Vietnam
thvu@ncs.dut.udn.vn

Kim Anh Nguyen*
The University of Danang - University of Science and Technology
Da Nang, Vietnam
https://orcid.org/0000-0003-3408-847X, nkanh@dut.udn.vn
*Corresponding author

Dinh Duong Le
Faculty of Electrical Engineering
The University of Danang - University of Science and Technology
Da Nang, Vietnam
ldduong@dut.udn.vn

Duc Hanh Dinh
Faculty of Mechanical Engineering
The University of Danang - University of Science and Technology
Da Nang, Vietnam
ddhanh@dut.udn.vn

***Abstract*—Fault diagnosis plays a pivotal role in ensuring the reliability of oil-immersed power transformers, with Dissolved Gas Analysis (DGA) widely recognized as an essential diagnostic tool. However, the effectiveness of DGA methods is often compromised by inconsistent input vector formulations, leading to challenges in accurately classifying faults. This study focuses on addressing these limitations by systematically optimizing DGA input vectors to improve diagnostic reliability. Eighteen input vector types, including concentration—based, ratio—based, and graphical methods such as the Duval triangle and Pentagon, were evaluated using SMOTE—GBDT models. The findings revealed that integrating ratio—based and graphical features enhance classification accuracy by 5—10%. Notably, input vector *X*12, combining gas concentrations and Duval features, achieved a high accuracy of 98.70%, while simpler vectors like input vector *X*7 performed less effectively at 89.77%. By accurately identifying faults such as overheating, arcing, and partial discharges, this study provides timely and accurate information about potential transformer faults even in very early stages, helping power companies make timely decisions for maintenance, stable operation and optimization of the power system.**

Keywords—power transformer, dissolved gas analysis, machine learning, ensemble learning, synthetic minority over—sampling technique, gradient boosted decision tree

I. INTRODUCTION

The reliability and efficiency of power systems are fundamentally dependent on the operational health of oil—immersed power transformers, which serve as critical components for stable energy transmission and distribution [1]. Early detection of incipient faults in transformers is crucial to minimizing downtime, reducing maintenance costs, and preventing catastrophic failures [2]. Among various diagnostic methods, Dissolved Gas Analysis (DGA) has gained prominence for its capability to detect and classify faults such as overheating, arcing, and partial discharges through the analysis of dissolved gas concentrations in transformer oil.

Despite its widespread use, achieving consistent and reliable fault diagnosis with DGA remains a challenge. A major limitation lies in the selection and representation of input vectors derived from dissolved gas data [3,4]. Traditional approaches, such as concentration—based and ratio—based methods, often exhibit limited accuracy and robustness, particularly under complex fault scenarios or varying operating conditions [5]. Graphical methods, including the Duval triangle and Pentagon, offer valuable diagnostic insights but require further refinement to improve fault classification outcomes. These challenges underscore the need for a systematic approach to optimize DGA input vector formulations [6,7].

This study addresses these challenges by systematically evaluating 18 types of input vectors, encompassing concentration—based, ratio—based, and graphical representations. Advanced machine learning techniques, specifically Gradient Boosted Decision Trees (GBDT) enhanced with SMOTE, were employed to assess the diagnostic performance of these vectors [8,9]. The results highlight that combining multiple input vector types, such as ratio—based and graphical features—significantly improves diagnostic reliability, achieving up to 5–10% higher accuracy compared to traditional methods. For instance, input vector 12, which integrates individual dissolved gas concentrations, Duval Pentagon features, and Duval triangle 1, 4, and 5 features, demonstrated the highest accuracy of 98.70%, significantly outperforming simpler methods like vector 7 (Duval triangle 4), which achieved 89.77%.

The implications of these findings are profound for power transformer maintenance and reliability management. Improved fault diagnosis accuracy enables utilities to implement condition—based maintenance strategies, reducing operational risks, optimizing resource allocation, and extending the lifespan of transformers. This study contributes to the development of more reliable and effective diagnostic frameworks, advancing the goals of system reliability, operational efficiency, and sustainability in power systems.

II. PROPOSED NOVEL APPROACH

This section outlines the proposed methodology for improving the reliability of oil—immersed power transformer fault diagnosis based on an optimized evaluation of DGA dissolved gas component input vectors. The approach integrates systematic data preprocessing, advanced machine learning

979-8-3315-9431-2/25/$31.00 ©2025 IEEE

models, and a comprehensive analysis of input vector formulations derived from DGA data.

A. Data Collection and Preprocessing

The first step in the proposed approach is to collect a large dataset of dissolved gas analysis measurements from oil—immersed power transformers. This data should include the concentrations of the key dissolved gas components, such as hydrogen H_2), methane, ethylene (C_2H_4), and carbon monoxide, as well as the associated fault types [10,11].

1) Data Collection: The study uses a dataset containing dissolved gas concentrations (ppm) from various fault scenarios in oil—immersed power transformers. The dataset comprises measurements of five key gas components such as H_2, CH_4, C_2H_2, C_2H_4 and C_2H_6 which are critical indicators of faults such as overheating, arcing, and partial discharges. The dataset also includes fault labels derived from expert diagnostic reports, covering common transformer fault types. The data spans various operational and fault conditions, ensuring diverse and comprehensive coverage for model training and evaluation.

2) Data Preprocessing: To ensure the dataset is suitable for machine learning and to enhance the reliability of the proposed approach, several preprocessing steps were conducted:

a) Data Cleaning: Ensuring the quality and reliability of the training data is crucial for developing an accurate fault diagnosis model. Inconsistent, missing, or outlier values in the gas measurements were addressed. Missing values were imputed using mean values of corresponding fault categories, while extreme outliers were removed to maintain data integrity and prevent bias in model training.

b) Data Normalization: Gas concentration values (ppm) were normalized to a consistent range (i.e., from 0 to 1) to ensure that all features contribute equally during model training [12].

c) Data Balancing: The dataset was imbalanced, with certain fault categories (e.g., partial discharges) being underrepresented [1,13]. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic samples for minority classes, achieving a balanced dataset that improves model performance across all fault categories [8].

d) Data Splitting: The preprocessed dataset is then split into training (75%), validation (25%), and test (independent of training and validation dataset, not through model training process) sets to evaluate the performance of the proposed approach [14,15].

By following these preprocessing steps, the dataset was prepared for robust machine learning analysis, ensuring that the proposed method's performance reflects real—world fault scenarios and challenges.

B. Gradient Boosted Decision Trees

Gradient Boosted Decision Trees (GBDT) was selected as the core machine learning algorithm for fault diagnosis in this study due to its proven effectiveness in handling complex classification tasks and its ability to model non—linear relationships. GBDT operates under the boosting framework, where weak learners (decision trees) are combined iteratively to create a strong learner, with each tree focusing on reducing the residual errors of its predecessor, resulting in a robust and accurate model [16]. The algorithm is particularly effective for fault diagnosis as it can capture complex patterns and non—linear relationships between dissolved gas input vectors and fault types, ensuring precise classification. Additionally, GBDT incorporates a gradient—based optimization approach, which enhances its robustness to noisy data and outliers which common issues in DGA datasets [17].

GBDT model provides inherent feature importance metrics, enabling the identification of the most significant dissolved gas components and input vector features, which aids in understanding their contributions to specific fault types [18]. The model's performance was evaluated on an independent test set using accuracy, precision, recall, F1—score, and confusion matrices, demonstrating its ability to distinguish fault types such as overheating, arcing, and partial discharges with high reliability. In summary, GBDT's ability to handle diverse input vector formulations and its gradient—based optimization make it an ideal choice for improving the accuracy and robustness of DGA—based fault diagnosis in oil—immersed power transformers.

C. Performance Evaluation

The performance of the proposed fault diagnosis approach was rigorously evaluated using a set of widely recognized metrics, including accuracy, precision, recall, and F1—score, to provide a comprehensive assessment of its diagnostic reliability [13]. Accuracy measures the overall correctness of the model, while precision and recall evaluate its ability to correctly classify specific fault categories without misclassification. The F1—score, as a harmonic mean of precision and recall, offers a balanced metric that accounts for both false positives and false negatives, ensuring a more nuanced evaluation of the model's effectiveness.

D. Input Vectors Based on DGA Dissolved Gas Components

This study evaluates 18 input vector formulations derived from DGA data to identify the most effective representations for fault diagnosis. Each input vector is described in the following way.

1) Input vector X1: X1 contains the raw concentrations of five dissolved gas components:

$$X1 = \left[H_2, CH_4, C_2H_6, C_2H_4, C_2H_2\right].$$

2) Input vector X2 (Pentagon): X2 normalizes gas concentrations as percentages:

$$X2 = \left[\%H_2, \%CH_4, \%C_2H_6, \%C_2H_4, \%C_2H_2\right].$$

3) Input vector X3 (IEC): This input vector is based on the diagnostic criteria proposed in the IEC 60599 standard, which uses three gas ratios as input features:

$$X3 = \left[\frac{C_2H_2}{C_2H_4}, \frac{CH_4}{H_2}, \frac{C_2H_4}{C_2H_6}\right].$$

4) Input vector X4 (Roger): X4 uses the classic Roger's gas ratios as the input features and is defined as:

$$X4 = \left[\frac{C_2H_2}{C_2H_4}, \frac{CH_4}{H_2}, \frac{C_2H_4}{C_2H_6}, \frac{C_2H_6}{CH_4} \right].$$

*5) Input vector X*5 *(Dornenburg): X*5 uses the Doernenburg gas ratios as the input features, which are another widely used approach for DGA—based fault diagnosis:

$$X5 = \left[\frac{C_2H_2}{C_2H_4}, \frac{CH_4}{H_2}, \frac{C_2H_4}{C_2H_6}, \frac{C_2H_2}{CH_4} \right].$$

*6) Input vectors X*6, *X*7, and *X*8 *(Duval triangle Variations):* These vectors are based on the graphical methods, including Duval triangle 1, 4, and 5 methods (denoted *X*6, *X*7, and *X*8 respectively) which use the concentrations of three key dissolved gases to diagnose fault types. *X*6, *X*7, and *X*8 are as follows:

- $X6 = [\%CH_4, \%C_2H_4, \%C_2H_2]$
- $X7 = [\%H_2, \%CH_4, \%C_2H_6]$
- $X8 = [\%CH_4, \%C_2H_4, \%C_2H_6]$

*7) Input vectors X*9—*X*18 *(combined input vectors):* Input vectors combining raw gas concentrations, ratios, and graphical methods arre evaluated. Where,

- *X*9 = *X*8 + *X*7 + *X*6. This input vector combines the features from the Duval triangle 1, 4, and 5 methods to leverage the information from multiple dissolved gas components;
- *X*10 = *X*1 + *X*2, which combines the individual dissolved gas component concentrations with the additional features from the Duval Pentagon method;
- *X*11 = *X*2 + *X*9, which combines the features from the Duval Pentagon and the Duval triangle 1, 4, and 5 methods;
- *X*12 = *X*1 + *X*11 = *X*1 + *X*2 + *X*9, which combines all the proposed input features, including the individual dissolved gas component concentrations, the Duval Pentagon features, and the Duval triangle 1, 4, and 5 features;
- *X*13 = *X*1 + *X*3, which combines the individual dissolved gas component concentrations with the IEC—based features;
- *X*14 = *X*1 + *X*4, which combines the individual dissolved gas component concentrations with the Roger's ratio—based features;
- *X*15 = *X*1 + *X*5, which combines the individual dissolved gas component concentrations with the Doernenburg ratio—based features;
- *X*16 = *X*3 + *X*4 + *X*5. This vector combines the IEC—based, Roger's ratio—based, and Doernenburg ratio—based features;
- *X*17 = *X*1 + *X*16 = *X*1 + *X*3 + *X*4 + *X*5. This input vector combines the individual dissolved gas component concentrations with the IEC—based, Roger's ratio—based, and Doernenburg ratio—based features;

*X*18 = *X*11 + *X*17 = *X*2 + *X*9 + *X*1 + *X*3 + *X*4 + *X*5. This input vector combines all the proposed input features, including the individual dissolved gas component concentrations, the IEC—based, Roger's ratio—based, and Doernenburg ratio—based features, as well as the Duval Pentagon and Duval triangle 1, 4, and 5 features.

III. Experimental Results

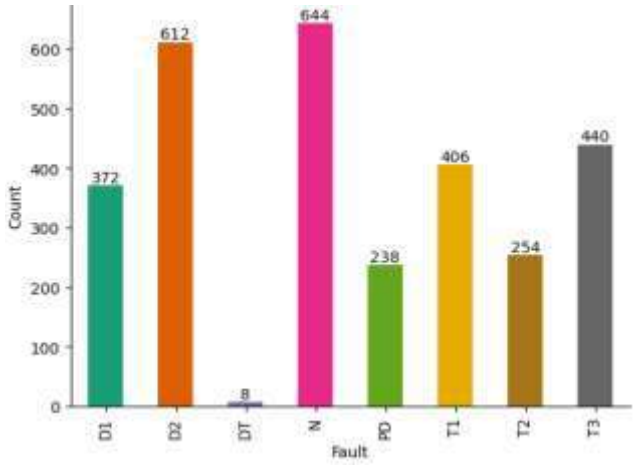
In this section, we demonstrate the effectiveness of the proposed approach in improving the reliability of oil—immersed power transformer fault diagnosis through the evaluation and optimization of DGA dissolved gas component input vectors. The diagnostic performance varies significantly among different input vectors, emphasizing the important role of input vector selection in accurate fault classification.

Comprehensive experiments were conducted utilizing a dataset of 2,974 DGA samples, as illustrated in Fig. 1(a). These samples were employed for training and validation purposes, alongside an independent testing dataset comprising 595 DGA samples. Both datasets were derived from reputable literature and operational data of power transformers within Vietnam's power grid over the preceding decade, especially world data [19]. The SMOTE technique was subsequently applied to the original dataset, resulting in an expanded dataset of 5,152 samples, as depicted in Fig. 1(b). This augmented dataset was partitioned into two subsets: 75% (3,864 samples) for training and 25% (1,288 samples) for validation, adhering to the widely accepted standard 75%/25% ratio in machine learning research.

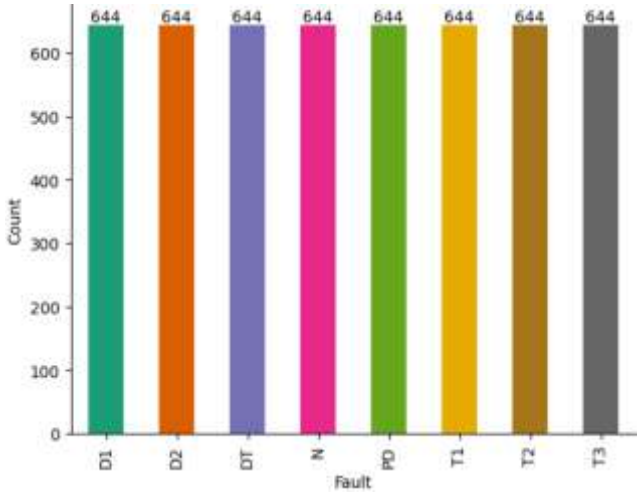
The model achieves a training accuracy of 98.70% and a testing accuracy of 98.32%, as shown in Table I. These results highlight the model's strong generalization ability across both subsets. For the testing subset, macro average metrics include precision (98.26%), recall (98.29%), and F1—Score (98.27%). The lower precision compared to recall indicates the model's strong detection of faulty patterns, though some false positives may occur. Additionally, the weighted average metrics, precision (98.33%), recall (98.32%), and F1—Score (98.32%), the accuracy of the SMOTE—GBDT model when applying the input vector as vector12.

Fig. 2 illustrates the confusion matrices for the SMOTE—GBDT model using *X*12, evaluated on both the training dataset and the testing dataset. The matrix on the training dataset shows a high degree of accuracy, with most true labels aligning closely with predicted labels across all categories. Categories such as DT, T3, and N demonstrate excellent performance, with minimal misclassifications, achieving near—perfect predictions. However, slight misclassifications are observed in categories D1 and D2, likely due to the synthetic data introduced by the SMOTE algorithm.

In the testing dataset, while the overall accuracy remains strong, misclassifications are slightly more pronounced compared to the training set. For example, T3, T2, and T1 maintain high predictive performance, but PD, D1, and D2 show some degree of misclassification, indicating potential challenges in these categories. This discrepancy highlights the gap between

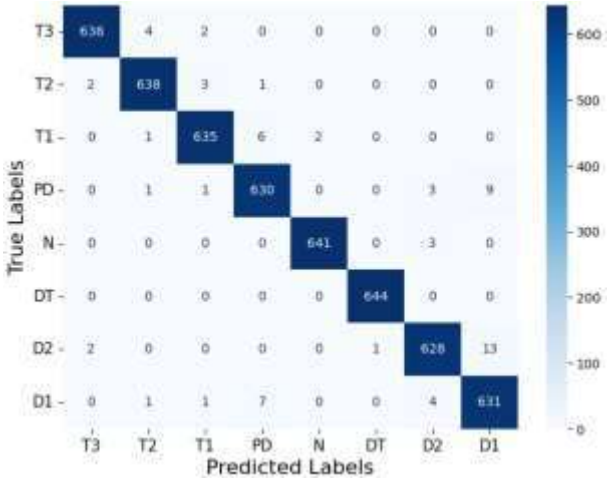


(a)

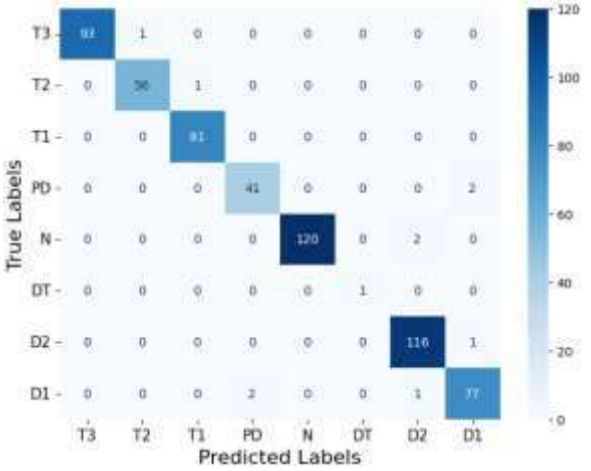


(b)

Fig. 1. Original dataset distribution before (a) and after (b) applying the SMOTE algorithm.



(a)



(b)

Fig. 2. Confusion matrices of SMOTE—GBDT and input vector *X*12 model (a) on training dataset and (b) on testing dataset.

training and real—world testing scenarios, emphasizing the importance of further optimization to improve generalization on unseen data.

As shown in Fig. 3, the chart illustrates the accuracy rates of the SMOTE—GBDT model evaluated across 18 input vectors. The model demonstrates high performance, with 16 out of 18 vectors achieving accuracy rates exceeding 98%, indicating effective training. *X*12 achieves the highest accuracy at 98.70%, suggesting strong alignment with the model. Other vectors, such as *X*10, *X*11, and *X*18, also perform consistently well, maintaining accuracy around 98.45%. However, some vectors, particularly *X*6 (92.10%) and *X*7 (89.77%), exhibit relatively lower accuracy, indicating that the model encounters challenges with the data associated with these vectors. While the implementation of the SMOTE algorithm effectively addressed data imbalance and contributed to high overall accuracy, these results suggest the need for additional preprocessing or parameter optimization to enhance performance on these specific vectors.

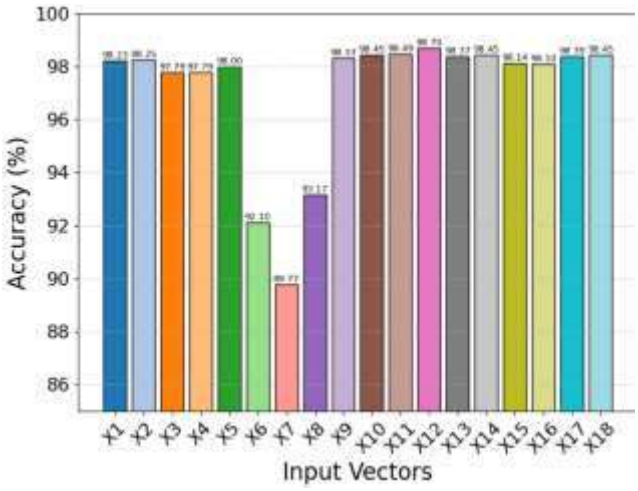


Fig. 3. Accuracy of each input vector.

TABLE I. PERFORMANCE METRICS OF THE PROPOSED SMOTE—GBDT AND INPUT VECTOR12 MODEL

Model	Training			Testing		
Metrics	**Precision**	**Recall**	**$F_{1—Score}$**	**Precision**	**Recall**	**$F_{1—Score}$**
D1	96.63%	97.98%	97.30%	96.25%	96.25%	96.25%
D2	98.43%	97.52%	97.97%	97.48%	99.15%	98.31%
DT	99.84%	100.00%	99.92%	100.00%	100.00%	100.00%
N	99.69%	99.53%	99.61%	100.00%	98.36%	99.17%
PD	97.83%	97.83%	97.83%	95.35%	95.35%	95.35%
T1	98.91%	98.60%	98.76%	98.78%	100.00%	99.39%
T2	98.91%	99.07%	98.99%	98.25%	98.25%	98.25%
T3	99.38%	99.07%	99.22%	100.00%	98.94%	99.47%
Accuracy	—	—	**98.70%**	—	—	**98.32%**
Macro avg.	**98.70%**	**98.70%**	**98.70%**	**98.26%**	**98.29%**	**98.27%**
Weighted avg.	**98.70%**	**98.70%**	**98.70%**	**98.33%**	**98.32%**	**98.32%**

TABLE II. F1—SCORE BY FAULT TYPE OF EACH INPUT VECTOR

	***X*1**	***X*2**	***X*3**	***X*4**	***X*5**	***X*6**	***X*7**	***X*8**	***X*9**	***X*10**	***X*11**	***X*12**	***X*13**	***X*14**	***X*15**	***X*16**	***X*17**	***X*18**
D1	96.56	96.70	96.16	96.07	96.15	92.53	90.12	92.36	96.78	96.71	96.71	97.30	96.54	96.85	96.48	96.31	96.70	96.70
D2	97.42	97.74	96.82	96.98	97.36	93.65	88.67	93.11	98.05	97.89	98.28	97.97	97.74	98.05	97.50	97.51	97.97	97.82
DT	99.77	99.69	99.31	99.61	99.61	93.19	93.57	94.98	99.77	99.84	99.77	99.92	99.77	99.77	99.77	99.69	99.84	99.84
N	99.15	98.67	97.17	97.65	98.36	92.09	81.19	88.06	98.67	99.38	99.14	99.61	99.30	99.30	98.91	98.36	99.15	99.22
PD	97.59	97.12	96.11	96.73	97.04	88.58	92.47	93.57	97.20	97.50	97.42	97.83	97.51	97.51	97.27	96.96	97.43	97.43
T1	98.52	98.84	98.05	98.83	98.68	91.55	92.16	95.97	98.91	98.91	99.22	98.76	98.76	98.91	98.52	98.91	99.07	98.91
T2	97.90	98.29	97.19	97.98	95.32	90.71	90.85	92.62	98.52	98.37	98.76	98.99	98.37	98.45	98.14	98.37	98.37	98.68
T3	98.99	98.99	98.06	98.45	95.78	94.35	89.48	94.44	98.76	99.07	99.07	99.22	98.99	98.76	98.52	98.68	98.68	98.99

Having constructed all the input vectors previously, an analysis was performed to calculate the F1—score for each type of error corresponding to each input vector. Table II illustrates a detailed analysis of the F1—score that leads to the choice of input vectors. The F1—score demonstrates a high level of consistency across vectors, with the majority of vectors (*X*1 to *X*18) achieving high scores across all categories, often exceeding 90%, indicating a well—trained model. Among these, *X*12 stands out, delivering the highest F1—score across all vectors. Certain categories, such as DT and T3, exhibit exceptional and consistent performance, with scores approaching or surpassing 99%.

Despite the limited number of samples in the original dataset, DT achieves a high F1—score due to the application of the SMOTE algorithm for balancing. This algorithm generated interpolated data, increasing the sample size for DT from 8 to 644. However, this process inadvertently introduced a bias toward the DT error type, which may partially explain its elevated performance in comparison to other categories. This highlights the importance of careful oversight when applying data augmentation techniques to prevent unintended biases in model evaluation.

The results of the study demonstrate the effectiveness of the proposed approach in enhancing the reliability of oil—immersed power transformer fault diagnosis through the evaluation and optimization of DGA dissolved gas component input vectors. Diagnostic performance exhibited significant variation across different input vector formulations, underscoring the critical role of input vector selection in achieving accurate fault classification. Input vectors that combined ratio—based, graphical, and percentage—based representations, such as input vectors *X*11 (98.49%) and *X*12 (98.70%), consistently outperformed others, yielding the highest accuracy and reliability. These vectors effectively captured distinct fault characteristics, enabling precise differentiation between fault types.

The Gradient Boosted Decision Trees (GBDT) model played a pivotal role in these results. Its capacity to model complex, non—linear relationships and handle noise and outliers contributed significantly to the improvements in fault classification. Furthermore, the feature importance analysis provided by GBDT identified key dissolved gas components and input vector features, offering insights into their contributions to fault diagnosis. Comparative analysis confirmed that the proposed method consistently surpassed traditional diagnostic techniques, such as ratio—based and graphical methods, particularly in complex fault scenarios.

IV. CONCLUSION AND PERSPECTIVES

This study proposed and evaluated a novel approach to improving the reliability of fault diagnosis for power transformers by systematically analyzing and optimizing DGA input vectors. The findings demonstrated that the selection and combination of input vector formulations are critical in achieving accurate and reliable fault classification. Among the 18 input vectors tested, advanced combinations such as $X11$ and $X12$, which integrate ratio-based, graphical, and percentage-based features, achieved the highest diagnostic accuracies of 98.70% and 98.49%, respectively. Performance analysis revealed that while basic input vectors, such as raw gas concentrations ($X1$) and Duval Pentagon ($X2$), provided satisfactory results with accuracies around 98%, more advanced combinations captured fault-specific characteristics more effectively, leading to incremental improvements. Simpler representations, including ratio-based and graphical methods such as Duval Triangle-based vectors (e.g., $X6$, $X7$, and $X8$), showed relatively lower accuracy, with $X7$ achieving only 89.77%. This highlights their limitations in complex fault scenarios and emphasizes the importance of integrating diverse and complementary input vector features to enhance diagnostic reliability.

The adoption of GBDT played a pivotal role in the superior performance of the proposed approach. GBDT's ability to model non-linear relationships, handle noise, and provide feature importance analysis proved essential for robust fault classification. The feature importance insights further aid in understanding the contributions of individual dissolved gas components and input vector features, enhancing both interpretability and practical diagnostic decision-making.

The implications of these findings are significant for power transformer maintenance. This study demonstrates the potential of combining advanced machine learning models with optimized input vector formulations to address challenges in transformer fault diagnosis. Future research will focus on validating the approach across larger and more diverse datasets, integrating real-time diagnostic systems, and incorporating additional fault indicators to further enhance reliability and applicability.

While the current study has demonstrated promising results in diagnosing transformer faults based on degradation information derived from liquid insulation, it remains limited by the exclusion of solid insulation degradation. Future research will focus on developing an integrated diagnostic model that combines both liquid and solid insulation indicators—such as degree of polymerization, 2-furfuraldehyde concentration, and paper moisture content—to provide a more comprehensive and accurate assessment of transformer health index. This approach is expected to enhance the reliability of fault diagnosis and provide effective support for implementing condition-based maintenance strategies.

REFERENCES

[1] L. Wang, T. Littler, and X. Liu, "Hybrid AI model for power transformer assessment using imbalanced DGA datasets," IET Renewable Power Gener., vol. 17, no. 8, pp. 1912—1922, 2023.

[2] A. Wajid et al., "Comparative Performance Study of Dissolved Gas Analysis (DGA) Methods for Identification of Faults in Power Transformer," Int. J. Energy Res., vol. 2023, no.1 pp. 9960743, 2023.

[3] M. Demirci, H. Gozde, M.C. Taplamacioglu, "Fault Diagnosis of Power Transformers with Machine Learning Methods Using Traditional Methods Data," Int. J. Tech. Phys. Probl. Eng., vol. 13, no. 4, pp. 225–230, 2021.

[4] O. Kherif, Y. Benmahamed, M. Teguar, A. Boubakeur, S. S. M. Ghoneim, "Accuracy Improvement of Power Transformer Faults Diagnostic Using KNN Classifier With Decision Tree Principle," IEEE Access, vol. 9, pp. 81693–81701, 2021.

[5] S. A. Ward, S. G. E.—D. Ibrahim, A. A. El—Faraskoury, D. A. Mansour, and M. Badawi, "Identification of Transformer Oil incipient Faults Based on the Integration between Different DGA Techniques," Delta Univ. Sci. J., vol. 6, no. 1, pp. 412—421, 2023.

[6] S. A. Wani, A. S. Rana, S. Sohail, O. Rahman, S. Parveen, and S. A. Khan, "Advances in DGA based condition monitoring of transformers: A review," Renewable Sustainable Energy Rev., vol. 149, p. 111347, 2021.

[7] F. Guerbas, Y. Benmahamed, Y. Teguar, R. A. Dahmani, M. Teguar, E. Ali, M. Bajaj, S. A. D. Mohammadi, and S. S. M. Ghoneim, "Neural networks and particle swarm for transformer oil diagnosis by dissolved gas analysis," Sci. Rep., vol. 14, no. 9271, pp. 1–14, 2024.

[8] R. M. A. Velásquez, "A comprehensive analysis for wind turbine transformer and its limits in the dissolved gas evaluation," Heliyon, vol. 10, no. 20, 2024.

[9] Y. Benmahamed, O. Kherif, M. Teguar, A. Boubakeur, and S. S. M. Ghoneim, "Accuracy Improvement of Transformer Faults Diagnostic Based on DGA Data Using SVM—BA Classifier," Energies, vol. 14 no. 10, p. 2970, 2021.

[10] S. D. Patil, A. J. Patil, M. Dharme, and R. K. Jarial, "DGA Based Ensemble Learning Approach for Power Transformer Fault Diagnosis," 2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT), IEEE, Jun. 09, 2023.

[11] H. Suwarno, R. A. Sutikno, R. A. Prasojo, and A. Abu‐Siada, "Machine learning based multi—method interpretation to enhance dissolved gas analysis for power transformer fault diagnosis," Heliyon, vol. 10, no.4, 2024.

[12] S. Kim et al., "A Semi—Supervised Autoencoder With an Auxiliary Task (SAAT) for Power Transformer Fault Diagnosis Using Dissolved Gas Analysis," IEEE Access, vol. 8, pp. 178295—178310, 2020.

[13] P. A. R. Azmi, M. Yusoff, and M. T. M. Sallehud—din, "Improving transformer failure classification on imbalanced DGA data using data—level techniques and machine learning," Energy Rep., vol 13, pp. 264—277, 2025.

[14] Y. D. Almoallem, I. B. M. Taha, M. I. Mosaad, L. Nahma, and A. Abu‐Siada, "Application of Logistic Regression Algorithm in the Interpretation of Dissolved Gas Analysis for Power Transformers," Electronics, vol. 10, no.10 p. 1206, 2021.

[15] G. Santamaría—Bonfil, G. Figueroa, M. A. Zuniga—Garcia, C. G. A. Ramos, and A. Bassam, "Power Transformer Fault Detection: A Comparison of Standard Machine Learning and autoML Approaches," Energies, vol. 17, no. 1, p. 77, 2023.

[16] C. Bentéjac, A. Csörgő, G. Martínez—Muñoz, "A Comparative Analysis of Gradient Boosting Algorithms," Artif. Intell. Rev., vol. 54, pp. 1937–1967, 2020.

[17] J. Shah, "Gradient Boosting," Technical Report, November 2020, Nirma University, Ahmedabad, India

[18] L. Wang, C. Jianfei, Y. Ding, H. Yao, Q. Guo, and H. Yang, "Transformer fault diagnosis method based on SMOTE and NGO—GBDT," Sci. Rep., vol. 14, no. 1, p. 7179, 2024.

[19] N. V. Nga, N. H. Chien, D. Truc, T. D. Tho, N. V. Luc, T. H. Vu, "Research on application of artificial intelligence in diagnosis of potential failures in transformers by dissolved gas analysis method," Univ. Danang J. Sci. Technol., vol. 22, pp. 30–35, October 2024