JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY

3/2025

vol. 101

Capon DOD/DOA Estimation Algorithm for Bistatic MIMO Radar Using Dipole Antenna Arrays with Known Mutual Coupling

Ouarda Barkat 1

Evaluating Effectiveness of Implementing G.fast Technology in Ukraine's Broadband Access Networks

Vitaliy Balashov, Vasyl Oreshkov, Iryna Barba, Dmytro Stelya, and Ihor Makarov 8

Physical Layer Security for Keyhole-based NOMA Downlink Systems with a Multi-antenna Eavesdropper

Sang-Quang Nguyen and Chi-Bao Le 16

Bio-inspired Routing Algorithms for UAV-based Networks: A Survey

Santosh Kumar, Amol Vasudeva, and Manu Sood 23

Enhancing Leaf Area Segmentation by Using Attention Gates and Knowledge Distillation in UNet Architecture

A. Shamim Banu and S. Deivalakshmi 51

A Convex Optimization-based Approach for Sidelobe Level Suppression and Null Control in Antenna Arrays by Displacing a Minimum Number of Elements

Magdy A. Abdelhay 63

Optimal Filter Selection for MIMO F-OFDM Systems in 5G Wireless Communication

Fadila A. Miloudi, Mohammed S. Bendelhoum, Fayssal Menezla, and Ridha I. Bendjillali 69

Performance Optimization of M/M/1 Queues with Working Vacations and Server Breakdowns in Wireless Communication Systems

S. Muthukumar, J. Ebenesar Anna Bagyam, and K. Basarikodi 79

(*Contents continued on back cover*)

eISSN 1899-8852

Capon DOD/DOA Estimation Algorithm for Bistatic MIMO Radar Using Dipole Antenna Arrays with Known Mutual Coupling

Ouarda Barkat

University of Frères Mentouri – Constantine 1, Constantine, Algeria

https://doi.org/10.26636/jtit.2025.3.2121

Abstract — This study focuses on the joint estimation of the direction of departure (DOD) and direction of arrival (DOA) of multiple targets in bistatic multiple input multiple output (MIMO) radar systems employing orthogonal waveforms. A linear array of half-wavelength dipole antennas (HWD) with known mutual coupling is utilized. The proposed method applies a two-dimensional Capon (2D Capon) algorithm to estimate both the DOD and DOA of multiple targets. To mitigate the adverse effects of mutual coupling, an efficient compensation mechanism is integrated into the Capon direction-finding algorithm. This mechanism relies on realistic electromagnetic modeling in which mutual coupling is represented using Toeplitz-structured coupling matrices. Through computer simulations, the influence of various system parameters on the algorithms performance is evaluated, with particular emphasis on its resolution capability and estimation accuracy. The results clearly demonstrate that incorporating mutual coupling compensation significantly enhances the accuracy of the 2D Capon algorithm.

Keywords — bistatic MIMO radar, Capon method, DOD/DOA estimation, mutual coupling

1. Introduction

The concept of multiple input, multiple output (MIMO) has been widely used in the field of wireless communications in recent years [1]. Implementing this concept in radar systems allows the design of a virtual network larger than that of traditional systems [2]–[4]. These systems greatly enhance detection performance and robustness, improving target localization depending on the type of MIMO radar. The emergence of bistatic MIMO radars has further increased interest in estimating both the direction of departure (DOD) and the direction of arrival (DOA) [5]–[7].

The operating principle of bistatic MIMO radar, consisting of an array of half-wavelength dipole (HWD) antennas, is to dynamically create a beam pattern. This beam pattern is designed to have its main lobe directed toward the desired signal, enhancing detection and localization.

Consequently, various angle estimation algorithms have been developed for MIMO radars, including estimation of signal parameters via rotational invariance techniques (ESPRIT), multiple signal classification (MUSIC), and Capon algorithms.

Generally, the MUSIC algorithm is regarded as having higher estimation accuracy than the Capon algorithm. However, in MIMO radar applications, it is often observed that the Capon algorithm outperforms the MUSIC algorithm [8]–[11]. The two-dimensional Capon (2D Capon) algorithm is a well-established and effective technique for estimating both DOD and DOA in bistatic MIMO radar systems [12]–[14].

Although the algorithm itself is not new, its integration with a mutual coupling compensation strategy in a bistatic MIMO configuration constitutes a novel and valuable contribution.

In practical radar systems, mutual coupling between the elements of the array can significantly degrade performance, especially in the estimation of DOD and DOA, by introducing signal distortions that complicate the estimation process [15]–[18].

This work advances the state of the art by applying compensation techniques within the Capon framework and conducting a quantitative analysis of mutual coupling effects using realistic antenna models, such as HWD arrays.

The proposed approach offers medium to high potential impact in the radar signal processing community, as it improves the practical deployment of MIMO radar systems and enables a more robust angle estimation performance.

This paper focuses on the estimation of DOD and DOA in bistatic MIMO radar systems using the 2D Capon algorithm. However, it is well established that the 2D Capon algorithm is highly sensitive to mutual coupling between HWD antennas in the array. To evaluate the impact of mutual coupling, we used a simplified model of the HWD array. Leveraging extensive data on dipoles, we analyze an array of equidistantly spaced dipoles.

Simulation results demonstrate that the performance of the Capon algorithm deteriorates due to mutual coupling, with the degradation becoming more pronounced as the interelement spacing between antennas decreases. This performance degradation can be significantly mitigated by employing a compensating matrix that optimally adjusts the DOD and DOA estimates.

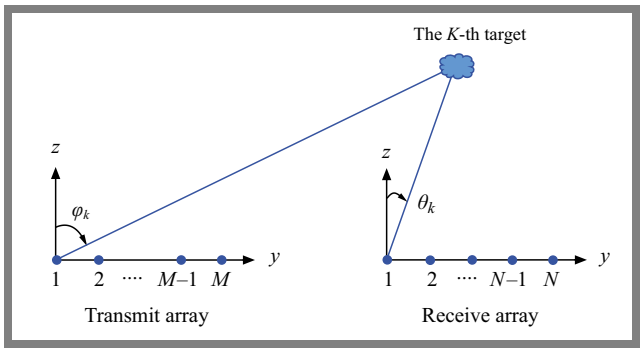

Fig. 1. Bistatic MIMO radar system.

2. The Theoretical Model

Figure 1 illustrates a bistatic MIMO radar system with K non-coherent targets. The transmitting array consists of M uniformly spaced half-wavelength antennas arranged along the y-axis with an inter-element spacing of d. Similarly, the receiving array consists of N HWD antennas, also arranged linearly, with the same spacing d between adjacent elements [19]–[21]. Each transmitting antenna emits an orthogonal signal $s_i(t)$ which is sampled every T seconds to obtain L snapshots.

In a bistatic radar system, the signals received on the receiving array after reflection from the K targets can be expressed as follows [18]:

$$X(t) = \sum_{i=1}^{K} (\beta_i \cdot \mathbf{c}_r(\varphi_i) \cdot \mathbf{a}_r(\varphi_i)) \cdot \left(\mathbf{c}_t(\theta_i) \cdot \mathbf{a}_t^T(\theta_i)\right) \cdot S(t) + Z(t) , \tag{1}$$

where: β_i is the complex amplitude of the i-th target, K is the total number of targets illuminated by the MIMO radar, φ_i and θ_i represent the DOD and DOA of the i-th target, respectively.

Typically, multiple samples are used to estimate (φ_i, θ_i), $i = 1, \ldots, K$, and the corresponding signal model with multiple snapshots L can be written as:

$$\mathbf{X}(L) = \sum_{i=1}^{K} \left(\beta_i \cdot \mathbf{c}_r(\varphi_i) \cdot \mathbf{a}_r(\varphi_i) \cdot \mathbf{c}_t(\theta_i) \cdot \mathbf{a}_t^T(\theta_i)\right) \cdot \mathbf{S}(L) + \mathbf{Z}(L) , \tag{2}$$

where:

- $\mathbf{Z}(L)$ represents the sensor noise, assumed to be non-uniform and modeled as a zero-mean Gaussian process. This assumption allows for an accurate representation of the stochastic nature of noise in this analysis.
- $\bar{\mathbf{A}}_t(\varphi_K)$ and $\bar{\mathbf{A}}_r(\theta_K)$ denote the steering matrices of the uniform linear transmit and receive arrays, respectively:

$$\bar{\mathbf{A}}_t(\varphi_K) = \left[\mathbf{a}_t(\varphi_1), \ldots, \mathbf{a}_t(\varphi_K)\right] , \tag{3}$$

$$\bar{\mathbf{A}}_r(\theta_K) = \left[\mathbf{a}_r(\theta_1), \ldots, \mathbf{a}_r(\theta_K)\right] . \tag{4}$$

The steering vectors $\bar{\mathbf{a}}_t(\varphi_i)$ and $\bar{\mathbf{a}}_r(\theta_i)$ are given by:

$$\bar{\mathbf{a}}_t(\varphi_i) = \left[1, e^{-j\frac{2\pi}{\lambda} d \sin(\varphi_i)}, e^{-j\frac{2\pi}{\lambda} 2d \sin(\varphi_i)}, \ldots, e^{-j\frac{2\pi}{\lambda}(M-1) d \sin(\varphi_i)}\right] , \tag{5}$$

$$\bar{\mathbf{a}}_r(\theta_i) = \left[1, e^{-j\frac{2\pi}{\lambda} d \sin(\theta_i)}, e^{-j\frac{2\pi}{\lambda} 2d \sin(\theta_i)}, \ldots, e^{-j\frac{2\pi}{\lambda}(N-1) d \sin(\theta_i)}\right] . \tag{6}$$

S denotes the transmitted baseband-coded waveform matrix in the following way:

$$S = [s_1, \ldots, s_M] . \tag{7}$$

In practical applications, both the transmitter and receiver are affected by mutual coupling. It is typically undesirable because energy that should be radiated outward is instead absorbed by a nearby antenna element. Similarly, energy that one antenna could have captured may be absorbed by a neighboring antenna. Consequently, mutual coupling negatively impacts the efficiency and overall performance of the antenna system. The array of HWD antennas is conceptualized as a multiport network, where the coupling matrix can be directly linked to the generalized impedance matrix of this network.

To compute the mutual coupling matrix $\bar{C}_t$ or $\bar{C}_r$, we account for the interactions among the elements of the matrix, resulting in mutual coupling effects. The array, comprising M (or N) coupled antennas is conventionally depicted as a M (or N) port network, as illustrated in Fig. 2. The mutual coupling matrix $\bar{C}_t$ or $\bar{C}_r$ can be expressed as detailed in [18], [22]–[24]:

$$\bar{C}_t = \left(\bar{Z}_{TA} + \bar{Z}_{TL}\right) \left(\bar{Z}_{Tij} + \bar{Z}_{TL}\bar{I}\right)^{-1} , \tag{8}$$

$$\bar{C}_r = \left(\bar{Z}_{RA} + \bar{Z}_{RL}\right) \left(\bar{Z}_{Rij} + \bar{Z}_{RL}\bar{I}\right)^{-1} , \tag{9}$$

where:

- $\bar{Z}_{TA}$ antenna impedance of isolated antennas in the transmitter,
- $\bar{Z}_{TL}$ terminating load in the transmitter,
- $\bar{Z}_{RA}$ antenna impedance of isolated antennas in the receiver,
- $\bar{Z}_{RL}$ terminating load in the receiver,
- $\bar{Z}_{Tij}$ mutual impedance between the i-th and j-th transmitter elements,
- $\bar{Z}_{Rij}$ mutual impedance between the i-th and j-th receiver elements.

In Eqs. (8)–(9), $\bar{C}_t$ and $\bar{C}_r$ denote the $M \times M$ and $N \times N$ mutual coupling matrices.

$\bar{C}_t$ and $\bar{C}_r$ can be written as:

$$\bar{C}_t = \begin{bmatrix} c_{11} & c_{12} & c_{13} & \ldots & c_{1M} \\ c_{21} & c_{22} & c_{23} & \ldots & c_{2M} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{M1} & c_{M2} & c_{M3} & \ldots & c_{MM} \end{bmatrix} , \tag{10}$$

$$\bar{C}_r = \begin{bmatrix} c_{11} & c_{12} & c_{13} & \dots & c_{1N} \\ c_{21} & c_{22} & c_{23} & \dots & c_{2N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{N1} & c_{N2} & c_{N3} & \dots & c_{NN} \end{bmatrix} . \tag{11}$$

In Eq. (1), it is evident that the coupling matrix influences the signal, leaving the noise unaffected. Once the coupling is characterized, compensating for the mutual coupling becomes a manageable task. Various compensation algorithms can be employed for this purpose, such as the open-circuit voltage method, the S parameter method, the full wave electromagnetic method of moments, the calibration method, and the mutual impedance methods.

Moreover, the coupling values exhibit approximate uniformity along the diagonals, allowing for modeling with a single parameter for each subdiagonal, thus resulting in a coupling matrix of the Toeplitz structure. Leveraging these insights, a compensated coupling model can be formulated as follows [23]:

$$\bar{C}_t = \begin{bmatrix} 1 & c_2 & c_3 & \dots & c_M \\ c_2 & 1 & c_2 & \dots & c_{M-1} \\ c_3 & c_2 & \ddots & \ddots & \vdots \\ \vdots & c_3 & \ddots & 1 & c_2 \\ c_M & c_{M-1} & \dots & c_2 & 1 \end{bmatrix} , \tag{12}$$

$$\bar{C}_r = \begin{bmatrix} 1 & c_2 & c_3 & \dots & c_N \\ c_2 & 1 & c_2 & \dots & c_{N-1} \\ c_3 & c_2 & \ddots & \ddots & \vdots \\ \vdots & c_3 & \ddots & 1 & c_2 \\ c_N & c_{N-1} & \dots & c_2 & 1 \end{bmatrix} . \tag{13}$$

We can put:

$$\mathbf{A}_{ct}(\varphi_i) = \mathbf{C}_t(\varphi_i) \cdot \mathbf{A}_t^T(\varphi_i) , \tag{14}$$

$$\mathbf{A}_{cr}(\theta_i) = \mathbf{C}_r(\theta_i) \cdot \mathbf{A}_r^T(\theta_i) . \tag{15}$$

Therefore, the output of $X(L)$ can be written as:

$$X(L) = \sum_{i=1}^{K} \beta_i(L) \cdot \mathbf{A}_{cr} \cdot \mathbf{A}_{ct} \cdot \mathbf{S}(t) + Z(L) . \tag{16}$$

When $\mathbf{S}^H$ is used as the matched filter matrix, the radar output of the matched filter can be formulated as follows:

$$\begin{aligned} \mathbf{Y}^{(l)} &= \frac{1}{\sqrt{T}} \mathbf{X}^{(l)} \mathbf{S}^H \\ &= \sum_{i=1}^{K} \sqrt{T}\, \beta_i(l)\, \mathbf{A}_{cr} \mathbf{A}_{ct} + \frac{1}{\sqrt{T}} \mathbf{Z}^{(l)} \mathbf{S}^H . \end{aligned} \tag{17}$$

Performing the vectorization operation on Eq. (17), we obtain [19]:

$$\mathbf{y}^{(l)} = \mathrm{vec}\left(\mathbf{Y}^{(l)}\right) . \tag{18}$$

The obtained vector $\mathbf{y}^{(l)}$ can be written as:

$$\mathbf{y}^{(l)} = \mathbf{A}(\varphi_i, \theta_i) \cdot \mathbf{B} + \mathbf{N}^{(l)} , \tag{19}$$

where:

$$\mathbf{N} = \mathrm{vec}\left(\frac{1}{\sqrt{T}} \mathbf{Z}^{(l)} \mathbf{S}^H\right) , \tag{20}$$

and $(\varphi_i^t, \theta_i^r)$ denotes the total manifold matrix with respect to both the array of the transmitter and receiver.

Then:

$$\mathbf{A}(\varphi_i, \theta_i) = \mathbf{A}_{ct}(\varphi_i) \otimes \mathbf{A}_{cr}(\theta_i) . \tag{21}$$

The covariance matrices of the received data y can be written in such a way:

$$\mathbf{C}_{MN} = \mathbf{R}_{yy} = \mathbb{E}\left[\mathbf{y}\mathbf{y}^H\right] , \tag{22}$$

$$\mathbf{C}_{MN} = \mathbf{A}\mathbb{E}\left[\mathbf{B}\mathbf{B}^H\right]\mathbf{A}^H + \mathbb{E}\left[\mathbf{N}\mathbf{N}^H\right] , \tag{23}$$

$$\mathbf{C}_{MN} = \mathbf{A}\mathbf{R}_{BB}\mathbf{A}^H + \sigma_Z^2 \mathbf{I}_{MN} . \tag{24}$$

3. Estimation by the Capon Algorithm

Capon estimation, also known as the minimum-variance distortionless response method, is an advanced signal processing technique used to estimate the parameters of a received signal in a noisy environment disrupted by interference sources.

This estimation aims to minimize the noise power in a given direction, allowing for a more accurate estimation of the parameters of the signal of interest, especially in scenarios with interference sources. This method is widely used in fields such as radar, wireless communications, and sonar processing to improve the resolution and sensitivity of communication systems.

The principle of the Capon algorithm is to find the weighting vector $\mathbf{w}(k)$ that minimizes the total output power of the beamformer while maintaining unity gain in the desired directions. This minimization can be solved using the method of Lagrange multipliers. The beamformer output is provided by [12], [14], [25]:

$$Y_f(t) = \mathbf{w}^H \cdot \mathbf{y}(t) . \tag{25}$$

Once the Y_f is obtained, it is useful to study the spatial covariance matrix of $\mathbf{Y}_f(t)$, denoted as $\mathbf{R}_{Y_f Y_f}$. Ideally, this is defined as the statistical expectation:

$$\mathbf{R}_{Y_f Y_f} = \mathbb{E}\left[\mathbf{Y}_f(t)\mathbf{Y}_f^*(t)\right] = \mathbf{w}^H \cdot \mathbf{C}_{MN} \cdot \mathbf{w} , \tag{26}$$

where, $\mathbf{C}_{MN}$ is the covariance matrix of the input signal vector $\mathbf{y}(t)$. However, since the true expectation $\mathbb{E}[\cdot]$ cannot be computed directly in practice, it is approximated using a time average on the snapshots of the L signal:

$$\widehat{\mathbf{R}}_{Y_f Y_f} = \frac{1}{L} \sum_{t=1}^{L} \mathbf{Y}_f(t)\mathbf{Y}_f^H(t) = \mathbf{w}^H\, \widehat{\mathbf{C}}_{MN}\, \mathbf{w} , \tag{27}$$

where $\widehat{\mathbf{C}}_{MN}$ is the sample covariance matrix of the received signal vector $\mathbf{y}(t)$ estimated over L snapshots.

The Capon method aims to minimize the output power while preserving the signal from the desired direction. The main objective of the Capon algorithm is to suppress interference and noise from other directions, ensuring that the desired signal remains undistorted. This optimization can be formulated as follows.

$$\min_{\mathbf{w}} \left(\mathbf{w}^H \widehat{\mathbf{C}}_{MN} \mathbf{w} \right), \tag{28}$$

subject to the constraint: $\left| w^H \cdot \mathbf{A}(\varphi_i, \theta_i) \right| = 1$.

The solution to this optimization problem yields the Capon weight vector:

$$\mathbf{w} = \frac{\widehat{\mathbf{C}}_{MN}^{-1} \mathbf{A}(\varphi_i, \theta_i)}{\mathbf{A}^H(\varphi_i, \theta_i) \widehat{\mathbf{C}}_{MN}^{-1} \mathbf{A}(\varphi_i, \theta_i)}. \tag{29}$$

Here, $\mathbf{C}_{MN}^{-1}$ represents the inverse covariance matrix of the received data see Eq. (24), which corresponds specifically to the upper triangular matrix $\mathbf{R}$ obtained from the QR decomposition of the matrix $\mathbf{C}_{MN}$.

In this paper, following the approach of [12], the covariance matrix $\mathbf{C}_{MN}^{-1}$ is used in its theoretical form for algorithm development and performance analysis. However, in practical implementations, this matrix must be estimated from measurements, typically using the sample covariance computed from received signal snapshots.

To estimate the direction parameters (φ_i, θ_i), we design a peak-searching function based on the Capon output power spectrum, defined as:

$$P(\varphi, \theta) = \frac{1}{\mathbf{A}^H(\varphi, \theta) \widehat{\mathbf{C}}_{MN}^{-1} \mathbf{A}(\varphi, \theta)}. \tag{30}$$

4. Numerical Results

In this section, the simulation results are presented to elucidate the efficacy of the proposed algorithm. We consider a bistatic MIMO radar system consisting of two uniform linear arrays, comprising M and N HWD antennas. To examine the effect of element separation d on the mutual coupling in a linear array, we simulated the real and imaginary parts of the mutual coupling impedance between two half-wavelength dipoles as a function of their separation, as shown in Fig. 2. As the distance between the elements increases, the magnitude of the mutual coupling impedance decreases and approaches zero.

In the first test, we evaluate the effectiveness and performance of the Capon method in achieving higher resolution for the joint estimation of the direction of departure (DOD) and direction of arrival (DOA) of target signals.

Figure 3 shows the root mean square error (RMSE) of the DOD/DOA estimate versus the signal-to-noise ratio (SNR) for three different uniform linear antenna array configurations: an array without coupling, an array with coupling, and an array with coupling using the compensated algorithm.

We considered four targets with departure angles of $[-30°, -10°, 20°, 60°]$ and arrival angles of $[-40°, -20°, 10°, 50°]$. The system parameters are set to $M = 16$ transmit antennas and $N = 8$ receive antennas,

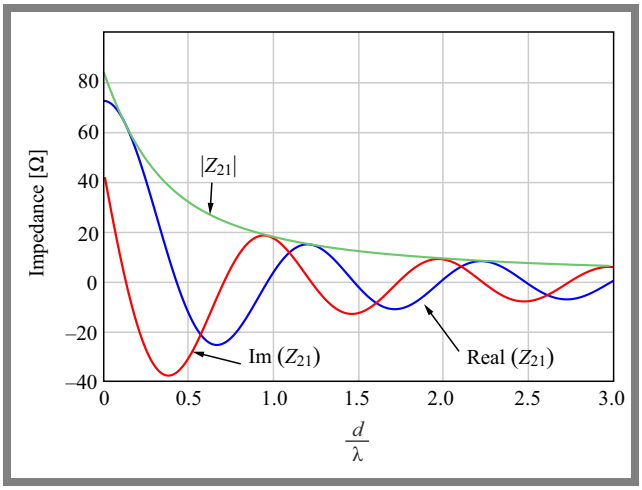

Fig. 2. Coupling impedance versus $\frac{d}{\lambda}$

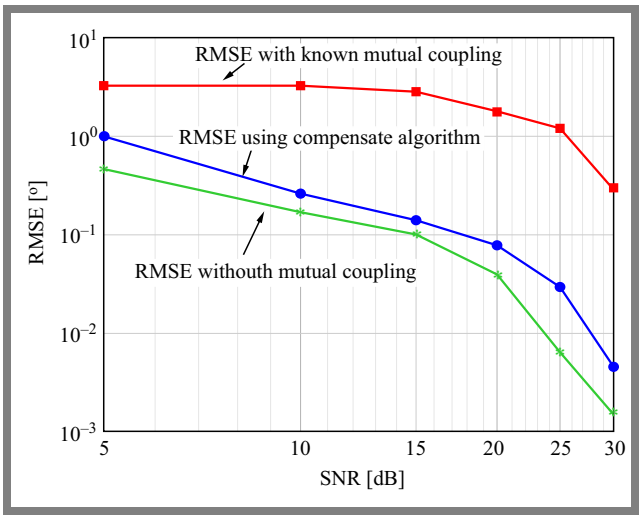

Fig. 3. RMSE versus SNR for DOD $= [-30°, -10°, 20°, 60°]$, DOA $= [-40°, -20°, 10°, 50°]$, $M = 16$, and $N = 8$.

with 250 signal snapshots. Performance is evaluated using the root mean square error (RMSE), computed over 600 Monte Carlo trials, using the following formula [26]:

$$\text{RMSE} = \sqrt{\frac{1}{VK} \sum_{v=1}^{V} \sum_{l=1}^{K} \left[(\hat{\varphi}_{l,v} - \varphi_l)^2 + \left(\hat{\theta}_{l,v} - \theta_l \right)^2 \right]}, \tag{31}$$

where, $\hat{\varphi}_{l,v}$ and $\hat{\theta}_{l,v}$ are the estimated DOD and DOA, respectively, of the K-th target in the V-th Monte Carlo trial. Here, $K = 4$ denotes the total number of targets and $V = 600$ is the number of trials used to average performance.

The results clearly demonstrate that mutual coupling effects cause significant degradation in estimation performance, particularly when the antennas are closely spaced. Moreover, the influence of known mutual coupling on the DOD/DOA estimation is SNR-dependent, with its impact generally decreasing as the SNR increases.

To better understand the behavior of the proposed algorithm, Fig. 4 illustrates the relationship between RMSE and the number of snapshots at a fixed SNR of 25 dB. The number of snapshots varies from 50 to 350. The results clearly show that

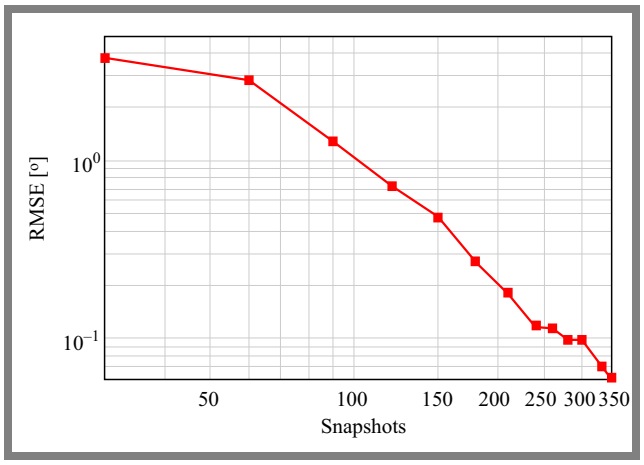

Fig. 4. RMSE versus snapshots with parameters: $M = 16$, $N = 8$, DOD $= [-30°, -10°, 20°, 60°]$, DOA $= [-40°, -20°, 10°, 50°]$, and SNR $= 25$ dB.

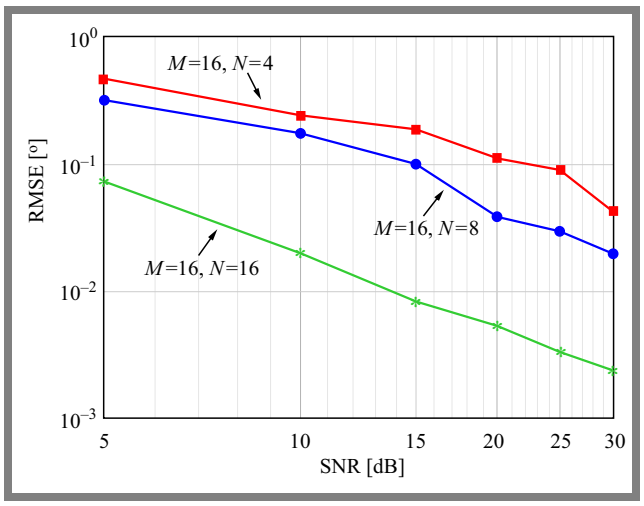

Fig. 5. RMSE versus SNR for different values of N, with $M = 16$, DOD $= [-30°, -10°, 20°, 60°]$, DOA $= [-40°, -20°, 10°, 50°]$, $L = 350$, and $d = \lambda/2$.

the accuracy of the estimation, as indicated by the RMSE, improves consistently with an increasing number of snapshots. This improvement is expected, as a higher number of snapshots enhances the estimation of the covariance matrix and effectively increases the SNR through temporal averaging, leading to more accurate DOD/DOA estimates.

In Fig. 5, the number of transmitting antennas M is fixed, while the number of receiving antennas N is varied. The results indicate that as the number of receiving antennas increases, the RMSE steadily decreases, demonstrating a consistent improvement in the accuracy of the estimation.

In Fig. 6, the number of receiving antennas N is kept constant, while the number of transmitting antennas M is varied. The results indicate that as the number of transmitting antennas increases, the RMSE steadily decreases, highlighting a significant improvement in estimation accuracy. Based on the results presented in Figs. 5 and 6, we observed that increasing the number of transmitting and receiving antennas results in a minimum RMSE, indicating more accurate estimations of the DOA and DOD.

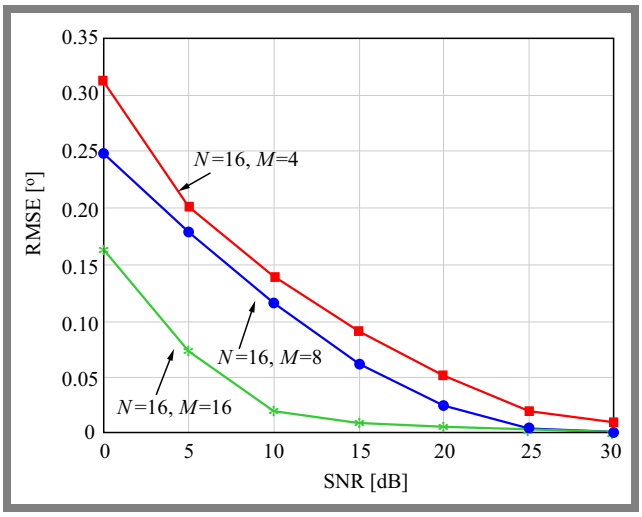

Fig. 6. RMSE versus SNR for different values of M, with $N = 16$, DOD $= [-30°, -10°, 20°, 60°]$, DOA $= [-40°, -20°, 10°, 50°]$, $L = 350$, $d = \lambda/2$, and SNR $= 25$ dB.

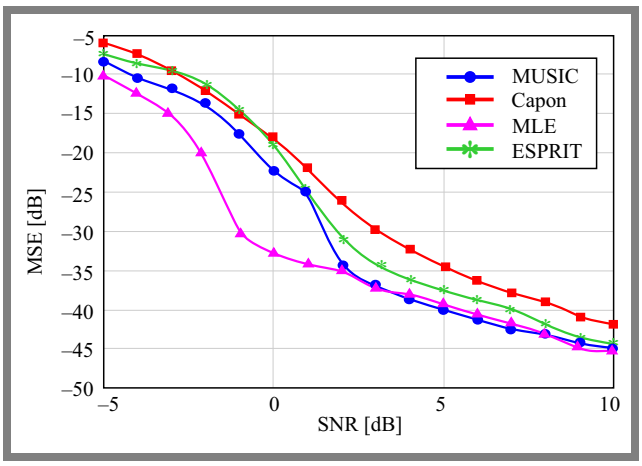

Fig. 7. Mean square error (MSE) versus SNR for DOD.

Figures 7 and 8 present the simulation results for a bistatic MIMO radar system equipped with $M = 5$ transmitting and $N = 5$ receiving antennas, both configured as uniform linear arrays (ULAs) with half-wavelength inter-element spacing. Three target scenarios are evaluated: $(30°, 45°)$, $(-8°, 30°)$, and $(0°, 5°)$, using 100 snapshots and 50 Monte Carlo trials.

Based on the results illustrated in Fig. 7 (DOD) and Fig. 8 (DOA), the mean squared error trends clearly indicate that the accuracy of angle estimation techniques improves as the signal-to-noise ratio increases. The maximum likelihood estimation (MLE) method demonstrates the best overall performance, achieving the lowest MSE.

The ESPRIT algorithm also provides high precision, with performance closely matching that of MLE, especially from 0 dB onward. Although the MUSIC method performs slightly below ESPRIT and MLE, it remains highly effective and exhibits a consistent reduction in MSE as the SNR increases.

In contrast, the Capon method shows comparatively lower performance, particularly under low-SNR conditions, indicating greater sensitivity to noise. However, beyond 5 dB, its MSE decreases significantly, suggesting improved robustness

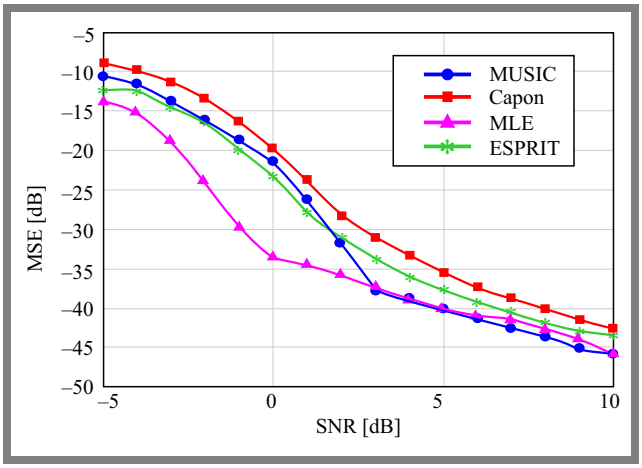

Fig. 8. Mean square error (MSE) versus SNR for DOA.

under moderate noise levels. In summary, MLE offers the highest estimation accuracy, followed by ESPRIT and MUSIC. Although Capon is less precise, it may be preferred in scenarios where reduced computational complexity is a priority [9], [27], [28].

Figure 9 shows the simulation results for signals departing from angles of $[-30°, -15°, 10°, 40°]$, also utilizing 16 antennas, an SNR of approximately 25 dB, and 350 snapshots. The element spacing in the array remains at $\lambda/2$. Here, we again observe four distinct peaks that align with the desired angles of departure.

Figure 10 presents the simulation results for signals coming from angles of $[-20°, -10°, 30°, 60°]$ using 16 antennas, an SNR of approximately 25 dB, and 350 snapshots. The spacing between the elements of the array is set to $\lambda/2$. In this case, we observe four distinct peaks corresponding to the desired angles of arrival. Furthermore, it is evident that the departure angles (DODs) and arrival angles (DOAs) can be clearly distinguished.

Figures 9 and 10 reveal that the spatial spectrum reaches its maxima at angles corresponding to DOD values of $-30°$, $-15.004°$, $9.992°$, and $40.001°$, and DOA values of $-19.995°$, $-9.999°$, $30°$, and $60°$. The precision of these estimates indicates that the angular search was performed on a finely spaced grid, probably with a step size of $0.01°$.

Figure 11 shows the estimation results for four targets, the SNR is set at 25 dB, with 350 snapshots. The results demonstrate that the DODs and DOAs are clearly observable and are automatically paired. In the following simulation, 500 Monte Carlo iterations are performed for the bistatic MIMO radar. We assume the presence of four non-coherent targets located at angles $(\varphi_1, \theta_1) = (-30°, -20°)$, $(\varphi_2, \theta_2) = (-15°, -10°)$, $(\varphi_3, \theta_3) = (10°, 30°)$, and $(\varphi_4, \theta_4) = (40°, 60°)$, respectively. It can be shown that the transmit angles (DODs) and the receive angles (DOAs) can be clearly observed.

5. Conclusions

In this paper, we examine the estimation of the direction of departure (DOD) and direction of arrival (DOA) for bistatic MIMO radar systems in the presence of known mutual coupling. Our study is grounded in fundamental electromagnetic principles and employs the Capon algorithm to achieve accurate signal estimation. Through computer simulations, we analyze the influence of various parameters on the performance of the Capon algorithm, focusing on its ability to efficiently and accurately resolve incoming signals.

Simulation results demonstrate that the DOD/DOA estimation performance improves with an increased number of array ele-

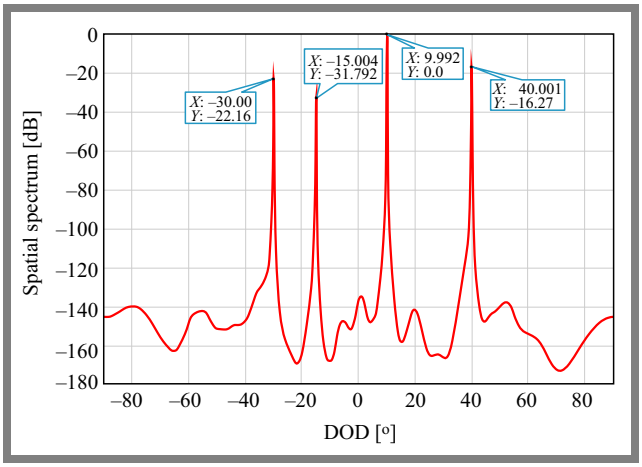

Fig. 9. Spatial spectrum versus DODs with $M_1 = N = 16$, DOD $= [-30°, -15°, 10°, 40°]$, DOA $= [-20°, -10°, 30°, 60°]$, $L = 350$, $d = \lambda/2$, and SNR $= 25$ dB.

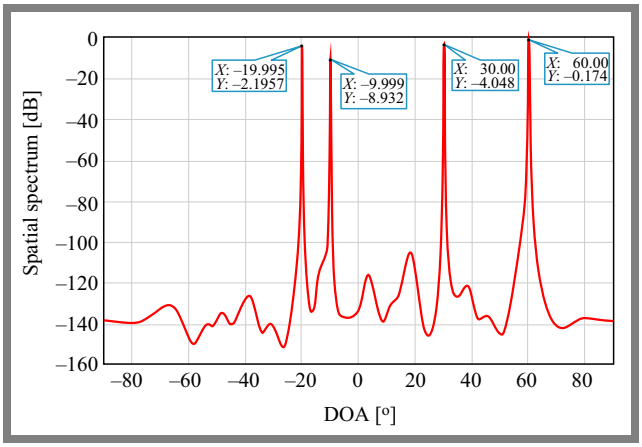

Fig. 10. Spatial spectrum versus DOAs with $M_1 = N = 16$, DOD $= [-30°, -15°, 10°, 40°]$, DOA $= [-20°, -10°, 30°, 60°]$, $L = 350$, $d = \lambda/2$, and SNR $= 25$ dB.

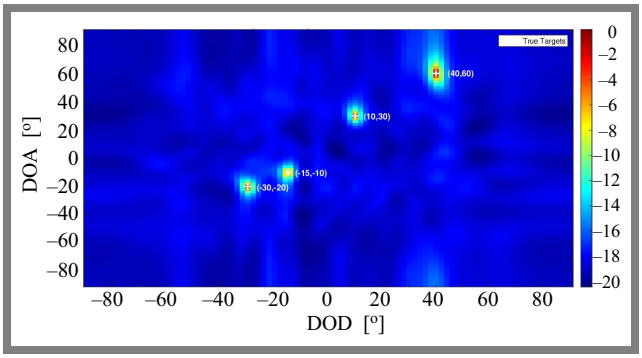

Fig. 11. Angle estimation of the proposed algorithm for four targets.

ments, a higher number of signal snapshots, and array spacing is $\lambda/2$. These enhancements result in sharper spectral peaks and reduced angular detection errors, highlighting the effectiveness of the Capon algorithm in estimating the DOD/DOA of incoming signals. However, despite these improvements, the known mutual coupling among HWD antennas introduces significant distortions to the output signal. This distortion negatively impacts the joint DOD/DOA estimation performance for multiple targets in bistatic MIMO radar systems using the Capon algorithm. To address these challenges, we recommend employing a compensation algorithm to mitigate the adverse effects of mutual coupling.

References

[1] J. Li and P. Stoica, "MIMO Radar with Colocated Antennas", *IEEE Signal Processing Magazine*, vol. 24, pp. 106–114, 2007 (https://doi.org/10.1109/MSP.2007.904812).

[2] B. Liao, "Fast Angle Estimation for MIMO Radar with Nonorthogonal Waveforms", *IEEE Transactions on Aerospace and Electronic Systems*, vol. 54, pp. 2091–2096, 2018 (https://doi.org/10.1109/TAES.2018.2847958).

[3] A.M. Haimovich, R.S. Blum, and L.J. Cimini, "MIMO Radar with Widely Separated Antennas", *IEEE Signal Processing Magazine*, vol. 25, pp. 116–129, 2008 (https://doi.org/10.1109/MSP.2008.4408448).

[4] P. Woodward, *Probability and Information Theory with Applications to Radar*, 3rd ed., Norwood, MA: Artech House, 128 p., 1980 (ISBN: 9780890061039).

[5] J. Chen, H. Gu, and W. Su, "A New Method for Joint DOD and DOA Estimation in Bistatic MIMO Radar", *Signal Processing*, vol. 90, pp. 714–718, 2010 (https://doi.org/10.1016/j.sigpro.2009.08.003).

[6] T.-Q. Xia, "Joint Diagonalization Based DOD and DOA Estimation for Bistatic MIMO Radar", *Signal Processing*, vol. 108, pp. 159–166, 2015 (https://doi.org/10.1016/j.sigpro.2014.09.010).

[7] M. Jin, G. Liao, and J. Li, "Joint DoD and DoA Estimation for Bistatic MIMO Radar", *Signal Processing*, vol. 89, pp. 244–251, 2009 (https://doi.org/10.1016/j.sigpro.2008.08.003).

[8] X. Zhang, C. Chen, and J. Li, "Angle Estimation Using Quaternion-ESPRIT in Bistatic MIMO-Radar", *Wireless Personal Communication*, vol. 69, pp. 551–560, 2013 (https://doi.org/10.1007/s11277-012-0589-3).

[9] C. Duofang, C. Baixiao, and Q. Guodong, "Angle Estimation Using ESPRIT in MIMO Radar", *Electronics Letters*, vol. 44, pp. 770–771, 2008 (https://doi.org/10.1049/el:20080276).

[10] M.L. Bencheikh and Y. Wang, "Joint DOD-DOA Estimation Using Combined ESPRIT-MUSIC Approach in MIMO Radar", *Electronics Letters*, vol. 46, pp. 1081–1083, 2010 (https://doi.org/10.1049/el.2010.1195).

[11] G. Zheng, B. Chen, and M. Yang, "Unitary ESPRIT Algorithm for Bistatic MIMO Radar", *Electronics Letters*, vol. 48, pp. 179–181, 2012 (https://doi.org/10.1049/el.2011.3657).

[12] X. Zhang and D. Xu, "Angle Estimation in Bistatic MIMO Radar Using Improved Reduced Dimension Capon Algorithm", *Journal of Systems Engineering and Electronics*, vol. 24, pp. 84–89, 2013 (https://doi.org/10.1109/JSEE.2013.00011).

[13] S.-W. Chen, C.-L. Meng, and A.-C. Chang, "DOA and DOD Estimation Based on Double 1-D Root-MVDR Estimators for Bistatic MIMO Radars", *Wireless Personal Communication*, vol. 86, pp. 1321–1332, 2016 (https://doi.org/10.1007/s11277-015-2991-0).

[14] R. Sanudin *et al.*, "Capon-like DOA Estimation Algorithm for Directional Antenna Arrays", *2011 Loughborough Antennas & Propagation Conference*, Loughborough, UK, 2011 (https://doi.org/10.1109/LAPC.2011.6114042).

[15] Z. Zhidong, Z. Jianyun, and N. Chaoyan, "Angle Estimation of Bistatic MIMO Radar in the Presence of Unknown Mutual Coupling", *2011 IEEE CIE International Conference on Radar*, Chengdu, China, 2011 (https://doi.org/10.1109/CIE-Radar.2011.6159474).

[16] C. Zhang, H. Huang, and B. Liao, "Direction Finding in MIMO Radar With Unknown Mutual Coupling", *IEEE Access*, vol. 5, pp. 4439–4447, 2017 (https://doi.org/10.1109/ACCESS.2017.2684465).

[17] B.T. Arnold and M.A. Jensen, "The Effect of Antenna Mutual Coupling in a MIMO Radar System", *IEEE Transactions on Antennas and Propagation*, vol. 67, pp. 1410–1416, 2017 (https://doi.org/10.1109/TAP.2018.2888702).

[18] P. Chen, Z. Cao, Z. Chen, and C. Yu, "Sparse DOD/DOA Estimation in a Bistatic MIMO Radar with Mutual Coupling Effect", *Electronics*, vol. 7, art. no. 341, 2018 (https://doi.org/10.3390/electronics7110341).

[19] J. Hu, E. Baidoo, and Z. Bao, "High-resolution Angle Estimation Method in Partly Calibrated Subarray-based Bistatic Multiple-input Multiple-output Radar with Unknown Non-uniform Noise", *IET Radar, Sonar and Navigation*, vol. 16, pp. 704–719, 2021 (https://doi.org/10.1049/rsn2.12214).

[20] Y. Guo, Y. Zhang, N. Tong, and J. Gong, "Angle Estimation and Self-calibration Method for Bistatic MIMO Radar with Transmit and Receive Array Errors", *Circuits, Systems, and Signal Processing*, vol. 36, pp. 1514–1534, 2017 (https://doi.org/10.1007/s00034-016-0365-9).

[21] F. Dong, C. Shen, K. Zhang, and H. Wang, "Real-valued Sparse DOA Estimation for MIMO Array System under Unknown NonUniform Noise", *IEEE Access*, vol. 6, pp. 52218–52226, 2018 (https://doi.org/10.1109/ACCESS.2018.2870257).

[22] N. Boughaba, O. Barkat, and C. Chettah, "Adaptive Beamforming Algorithm Based on MVDR for Smart Linear Dipole Array with Known Mutual Coupling", *Progress In Electromagnetics Research C*, vol. 124, pp. 125–134, 2022 (https://doi.org/10.2528/PIERC22080103).

[23] M. Bensalem and O. Barkat, "DOA Estimation of Linear Dipole Array with Known Mutual Coupling Based on ESPRIT and MUSIC", *Radio Science*, vol. 57, pp. 1–15, 2022 (https://doi.org/10.1029/2021RS007294).

[24] N. Boughaba and O. Barkat, "LMS and RLS Beamforming Algorithms Based Linear Antenna Array with Known Mutual Coupling", *Journal of Electromagnetic Waves and Applications*, vol. 37, pp. 1449–1462, 2024 (https://doi.org/10.1080/09205071.2023.2251979).

[25] X. Zhang and D. Xu, "Angle Estimation in MIMO Radar Using Reduced-dimension Capon", *Electronics Letters*, vol. 46, pp. 860–861, 2010 (https://doi.org/10.1049/el.2010.0346).

[26] H. Chen *et al.*, "Joint DOD and DOA Estimation for Bistatic MIMO Radar without Eigenvalue Decomposition", *Progress In Electromagnetics Research Letters*, vol. 80, pp. 67–74, 2018 (https://doi.org/10.2528/PIERL18100106).

[27] C. Jinli, G. Hong, and S. Weimin, "Angle Estimation Using ESPRIT without Pairing in MIMO Radar", *Electronics Letters*, vol. 44, pp. 1422–1423, 2008 (https://doi.org/10.1049/el:20089089).

[28] B. Tang, J. Tang, Y. Zhang, and Z. Zheng, "Maximum Likelihood Estimation of DOD and DOA for Bistatic MIMO Radar", *Signal Processing*, vol. 93, pp. 1349–1357, 2013 (https://doi.org/10.1016/j.sigpro.2012.11.011).

Ouarda Barkat, Professor
Department of Electronics
https://orcid.org/0000-0001-6784-8338
E-mail: barkat.ouarda@umc.edu.dz
University of Frères Mentouri – Constantine 1, Constantine, Algeria
https://www.umc.edu.dz

Evaluating Effectiveness of Implementing G.fast Technology in Ukraine's Broadband Access Networks

Vitaliy Balashov, Vasyl Oreshkov, Iryna Barba, Dmytro Stelya, and Ihor Makarov

State University of Intelligent Technologies and Telecommunications, Odessa, Ukraine

https://doi.org/10.26636/jtit.2025.3.2127

Abstract — The article examines the feasibility of implementing the G.fast technology in the process of modernizing fixed broadband access networks operated in Ukraine. An analysis of international experience in the field and of national broadband development strategies is provided. The data rates achievable by G.fast transmission systems relying on profile 106a over multi-pair TPP and UTP Cat. 5e cables are evaluated, with intrasystem interference and crosstalk taken into consideration as well. The effectiveness of applying the vectoring crosstalk compensation system to increase G.fast transmission rates is assessed. Based on the research results, recommendations are formulated for the effective deployment of G.fast in Ukraine's broadband access networks.

Keywords — broadband access, interference, multi-pair TPP cables, transmission rate, UTP Cat. 5e cable, vectoring

1. Introduction

Today, broadband Internet access is a fundamental telecommunication feature driving economic growth, facilitating social services and boosting e-Governance. One of the key challenges faced in Ukraine's digital transformation process consists in ensuring common access to high-speed Internet, regardless of the user's place of residence.

The development of broadband networks in general and fixed broadband access (FBB) networks in particular – with the latter based on copper subscriber cables (xDSL technologies) and optical solutions (FTTx or "fiber-to-the x" concepts) – is taking place under challenging conditions. These include technical disparities between regions, limited budgets for infrastructure projects, and the need for effective modernization of existing networks [1], [2].

In Ukraine, the development of FBB networks is of critical importance. Despite ongoing upgrades to the telecommunications infrastructure in urban areas, significant rural regions still have limited or no access to high-speed Internet [3], [4].

Furthermore, the large-scale war that has been ongoing since 2022 has posed additional challenges to the telecommunications infrastructure, demanding flexible solutions and the implementation of modern technologies, such as fiber to the home (FTTH), xDSL, DOCSIS, and fixed wireless access (FWA) [5].

Among the modern technologies that allow for a significant increase in the capacity of FBB networks without a complete replacement of the physical medium, G.fast takes notice [6], [7]. This technology enables transmission rates of more than 1 Gbps to be achieved on short network segments built using copper multipair telephone cables.

However, its effectiveness largely depends on the length of the subscriber line and the level of interference [8]. Thus, the application of G.fast is most appropriate when the fiber-to-the-distribution point (FTTDp) concept is relied upon, where the length of the copper segment does not exceed 250 m. In such deployments, G.fast-based FTTDp networks offer transmission rates comparable to those of fiber optic connections, while requiring significantly lower investments than full FTTH implementations [9].

Among the modern technologies that allow for a significant increase in the capacity of FBB networks without a complete replacement of the physical medium, G.fast takes notice [6], [7]. This technology enables transmission rates of more than 1 Gbps to be achieved on short network segments built using copper multipair telephone cables.

However, its effectiveness largely depends on the length of the subscriber line and the level of interference [8]. Thus, the application of G.fast is most appropriate when the fiber-to-the-distribution point (FTTDp) concept is relied upon, where the length of the copper segment does not exceed 250 m. In such deployments, G.fast-based FTTDp networks offer transmission rates comparable to those of fiber optic connections, while requiring significantly lower investments than full FTTH implementations [9].

In this context, research on the characteristics of G.fast transmission systems (TS) based on the existing infrastructure of Ukraine's fixed broadband networks, as well as potential development and modernization paths for these networks, are crucial to evaluate the real potential of this technology.

The purpose of this article is to analyze the dynamics of the development of broadband Internet in Ukraine, considering the technical aspects of introducing advanced FBB technologies. Additionally, the article aims to evaluate the effectiveness of G.fast transmission systems (G.fast TS) within Ukraine's FBB networks by modeling the characteristics of G.fast TS using traditional multi-pair telephone (TPP) and

UTP Cat. 5e cables, as an option for modernizing subscriber lines.

2. Literature Review and International Experience

Research described in the literature indicates that technologies based on the use of copper multi-pair cables, such as VDSL2 and G.fast, remain relevant in countries with well-developed infrastructure, where full replacement with optical media is economically unfeasible or would require considerable investment efforts. Article [10] notes that G.fast can achieve speeds of up to 1 Gbps over short copper loops, making it an effective solution for upgrading existing networks without the need for complete replacement of the cable.

Study [11] addresses the co-existence of G.fast and VDSL2 in FTTDp and FTTC-type networks. The authors analyze spectrum optimization methods and the degree of protection of existing services, which is a critical aspect when introducing new technologies into functioning networks.

In [12], it is emphasized that G.fast is an ideal choice for operators because it works on the existing copper telecommunication infrastructure already installed at the users' premises. However, since this infrastructure was not designed for the high frequency transmission rates used by G.fast, signal leakage may occur during the process. This radiation may directly affect the quality and reliability of radio services operating in the same frequency range. The study provides an assessment of whether radiation generated by a telecommunications network using G.fast complies with the requirements of ITU-T Recommendation K.60 and whether it may be a source of interference for radio services using the same band.

The works referred to above highlight that the key factors limiting G.fast performance include high power spectral density (PSD) of noise caused by near-end crosstalk (NEXT), far-end crosstalk (FEXT), electromagnetic interference from external sources, and signal attenuation that depends on frequency and length of line. Researchers recommend considering the cable type and design features when evaluating G.fast efficiency, as noise characteristics can vary significantly depending on the cable used.

Report [13] underscores the importance of combining government regulations with private investment to bridge the digital divide. It emphasizes the need for strategic planning and support for innovative technologies to ensure equal access to broadband Internet.

In [14], the Broadband Forum presents methodologies for calculating losses and interference characteristics for category 5e cables, including recommendations on the maximum line lengths for 106a and 212a profiles. The documents highlight the need for accurate PSD function modeling and the use of empirical data, especially in cases in which pre-installed cables are re-used.

Thus, international experience demonstrates that effective development of fixed broadband access requires:

- strategic, national level planning,
- support for innovative technologies (particularly G.fast),
- investment in digital infrastructure,
- boosting competition among providers,
- reduction of the digital divide between urban and rural areas.

There is strong evidence in global practice supporting the feasibility of using the G.fast technology for short network segments, and the negative impact of interferences is thoroughly evaluated. The results of this research may be useful for planning the modernization of fixed broadband networks, especially in urbanized areas with existing cable infrastructure.

The following sections of this article analyze the current situation in Ukraine as well as evaluate the prospects for deploying and developing fixed broadband solutions, taking into account the approaches discussed above.

3. Current State of and Development Plans for FBB

According to the Ukraine's Strategy for the Development of the Electronic Communications Sector 2030, one of the priorities is to ensure universal access to high-speed Internet regardless of location, including in rural areas [1]. FBB is a key prerequisite for the development of Ukraine's digital economy, electronic services, and innovation-oriented infrastructure. In the context of post-war recovery and Ukraine's digital transformation, FBB becomes not only a tool for accessing digital services, but also a strategic prerequisite to attract investment and develop human capital.

In 2020, with the support of the World Bank, Ukraine drew up a National Broadband Development Strategy (2020 – 2025) which addresses strategic tasks such as:

- connecting 95% of socially significant facilities (schools, hospitals, administrative service centers) to fiber-optic Internet,
- providing government subsidies to operators for connecting rural settlements,
- launching an interactive geographic information platform presenting FBB coverage [2].

In 2021, the Ukrainian Cabinet of Ministers approved an Action Plan for the Development of Broadband Internet Access for 2021 – 2022, with the objective of improving infrastructure and accessibility of the Internet [15].

According to [16], at the end of 2023, there were 8.06 million fixed Internet access lines in Ukraine, with the said result being 12% higher than the year before. The most notable growth occurred in rural areas, with the increase amounting to 25.4% and reaching 2.12 million connections. Despite the positive dynamics, only 62% of households have fixed Internet access, indicating a persistent digital divide, especially evident in rural regions.

In 2025, the government will acknowledge problems with the insufficient pace of broadband access growth. In [17],

several problems that hinder effective FBB deployment were highlighted [17]:

- lack of a unified digital platform for monitoring coverage, which complicates planning,
- fragmented responsibilities shared by different government authorities,
- limited funding, especially in the context of martial law and the need to restore damaged infrastructure,
- unequal technical resources available to specific operators – some providers are still using outdated equipment (e.g., ADSL).

According to [1], in the long-term Ukraine aims to achieve 100% nationwide broadband coverage, simultaneously aiming to complete the process of modernizing its gigabit technology infrastructure and relying on FBB to implement key national digital services.

4. Challenges and Opportunities Related to G.fast Technology

In the context of limited resources and the need for rapid network expansion, G.fast technology is seen as a promising solution to upgrade the existing copper infrastructure, especially in densely populated areas. This technology enables high data transmission rates over short distances, which is particularly relevant for apartment buildings and office centers.

However, the effectiveness of G.fast is largely dependent on the quality of the cable infrastructure and the length of the line. Ukraine's traditional public switched telephone network (PSTN) is built using multi-pair telephone cables of the TPP type. The most common telephone cable in the network is the TPP-10×2×0.4. The G.fast technology, following the FTTDp concept, utilizes the existing distribution segment of the PSTN cable infrastructure within buildings. The maximum line length within apartment buildings is limited to 250 m.

In many residential buildings, Internet connection requires modernization of the in-building network, including replacement of the cable infrastructure with twisted pair cables, typically of the UTP Cat. 5e variety. When deploying the G.fast technology, the question arises as to whether it is feasible to carry out such a modernization.

To address this issue, it is necessary to study the performance characteristics of G.fast when operating over multipair telephone cables of the TPP type and twisted pair cables, such as UTP Cat. 5e. Such a study is essential to evaluate the real potential of G.fast deployment in Ukraine, both under existing cable infrastructure conditions and in the context of its modernization, with the impact of cable infrastructure parameters (cable type, frequency and time domain characteristics, line length, and noise level) on the performance of G.fast TS taken into consideration as well.

In this study, it is assumed that the copper segments under analysis are not shared with other transmission systems such as VDSL or VDSL2. The G.fast system is deployed in a dedicated frequency band, consistent with the ITU-T recommendations, and interference from other copper-based services is excluded from the modeling.

The main criterion for determining the effectiveness of a TS is the data rate that can be achieved under specific operating conditions of the system. Therefore, to study the effectiveness of using G.fast TS in the Ukrainian FBB network and the feasibility of modernizing the cable infrastructure, it is necessary to define the initial data, determine the methods for evaluating the G.fast data rate, perform data rate calculations for G.fast based on the given initial data, analyze the results obtained, and formulate recommendations for the implementation of G.fast technology in Ukrainian fixed broadband networks. The initial data must include the following:

- identification of the characteristics of the G.fast transmission system that will influence the data rate evaluation process,
- determination of the characteristics of TPP and UTP Cat. 5e cables necessary to assess the G.fast data rate,
- specification of the conditions under which the results of the data rate evaluation process for TPP and UTP Cat. 5e cables can be compared.

5. Data Rate Evaluation

The evaluation of the data rate of the G.fast TS was carried out based on the following initial constraints:

- spectral mask of the G.fast system complies with the ITU-T G.9700 standard [6],
- frequency plan up to 106 MHz,
- used channels $i = 43 \ldots 2047$,
- number of orthogonality interval samples $N = 4096$,
- number of guard interval samples $L = 320$,
- PSD level of external additive white Gaussian noise is assumed to be uniform –140 dBm/Hz,
- cable types – TPP-10×2×0.4 and twisted pair cable UTP Cat. 5e 4×2×0.51,
- line length l from 50 to 250 m,
- number of TS operating in parallel on the multipair cable – 1 and 4.

Cable characteristics were determined by measuring cable samples at the cable manufacturer's testing laboratory. The measurement results were summarized and used to derive approximation formulas for frequency characteristics within the frequency range of up to 100 MHz. Table 1 presents the results of the process of determining the frequency characteristics for a 100-meter line.

- α intrinsic attenuation (attenuation coefficient) [dB/100 m],
- β phase coefficient [rad/100 m],
- A_N near-end crosstalk (NEXT) [dB/100 m],
- A_{ELF} equal-level far end crosstalk (ELFEXT) [dB/100 m].

To determine the data rate of the G.fast TS, we use the methodology described in [18]. The data rate is determined

Tab. 1. Frequency characteristics approximation coefficients.

Parameter	α [dB/100 m]			β [rad/100 m]	A_N [dB/100 m]		A_{ELF} [dB/100 m]	
Approximation functions	$\alpha = a + bf^c$			$\beta = d\,f$	$A_o = x - y\log f$		$A_3 = x - y\log f$	
Approximation coefficients	a	b	c	d	x	y	x	y
Cat. 5e 4×2×0.51	1.121	1.110	0.59	2.681	89.49	22.6	81.15	17.59
TPP-10×2×0.4	0.59	0.98	0.7	2.776	60.281	17.19	68.2	22.5
Note: The frequency in the formula is expressed in MHz								

by means of the signal-to-noise ratio (SNR). The total noise includes the following components:

- thermal noise, defined as AWGN with a power spectral density level of –140 dBm/Hz,
- external additive noise, which depends on the specific electromagnetic environment (in previous studies, this was modeled by increasing AWGN uniformly across the entire operating frequency range),
- crosstalk interference (XTI), originating from TS operating in parallel within a multi-pair cable,
- intrasystem interference, which, for systems using parallel transmission rate, includes two components: intersymbol interference (ISI) and interchannel interference (ICI), is collectively referred to as IS+ICI.

Therefore, to determine the data rate of the G.fast TS, it is necessary to identify all noise components that affect the system's performance.

In the simulation, we assume that we only influence the baseline thermal and AWGN at –140 dBm/Hz. The influence of specific external systems such as power line communication (PLC), which can operate in overlapping frequency bands, was not taken into account but will be considered in future studies using real-world electromagnetic compatibility data.

To determine crosstalk interference, we use the technique described in [19]. This methodology also accounts for the application of a crosstalk cancelation system based on the implementation of the vectoring method [20].

It should be noted that the impact of crosstalk interference on the G.fast transmission rate when operating over TPP-10×2×0.4 and UTP Cat. 5e cables, as well as the effectiveness of the vectoring system in compensating XTI and increasing the G.fast transmission rate, were already evaluated in [21].

However, the study [21] cannot be considered complete, as it did not take into account IS+ICI interference. Consequently, it lacks results on its potential impact on the G.fast data rate. In this work, we aim to address this research gap.

To determine IS + ICI interference, we use the methodology outlined in [22]. Since the method is based on modeling interference in the time domain, it requires an additional transformation of channel frequency characteristics into time domain characteristics, determining the impulse response (IR). The IR of the transmission channel was calculated using the inverse fast Fourier transform (IFFT) of the channel characteristics in the measured frequency domain. This IR was then used to simulate inter-symbol and inter-channel

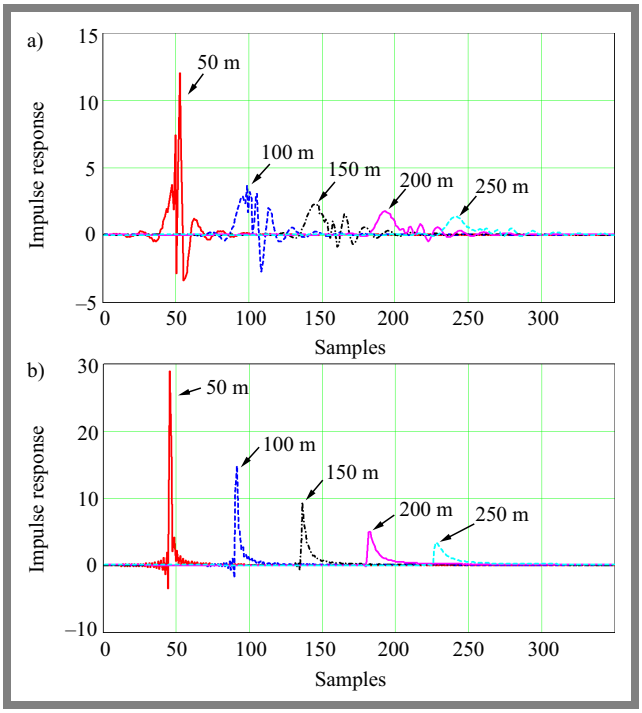

Fig. 1. Impulse response: a) to the TPP-10×2×0.4 and b) to the twisted pair cable UTP Cat. 5e 4×2×0.51.

interference components in the transmission rate calculations. Figure 1 presents the IRs for TPP and UTP cables of various lengths, illustrating signal dispersion and delay spread, i.e. parameters that are critical for IS+ICI estimation. The IFFT size corresponds to the transformation size of the modulation process in a G.fast system with a 106 MHz ($N = 4096$ samples).

6. Interference Component Analysis

Figures 2 and 3 present the calculated interference power distribution across the G.fast system channels when operating over TPP-10×2×0.4 and UTP Cat. 5e 4×2×0.51 cables with lengths of 50 and 200 m. The results demonstrate the dependence of the PSD level on the G.fast channel number $G(i)$, where the central frequency of channel i is defined as: $f_i = 51.75$ kHz $\cdot\, i$. The figures show three types of interference: –140 AWGN, IS+ICI, and XTI crosstalk interference, with their level calculated assuming parallel operation of four G.fast systems. Additionally, residual non-compensated

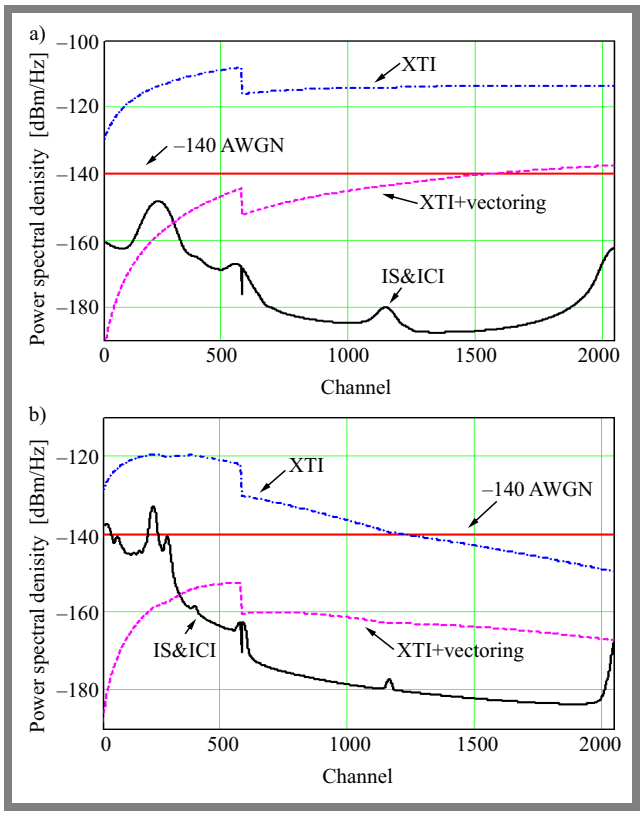

Fig. 2. PSD level of the G.fast system for: a) 50 m and b) 200 m TPP-10×2×0.4 line.

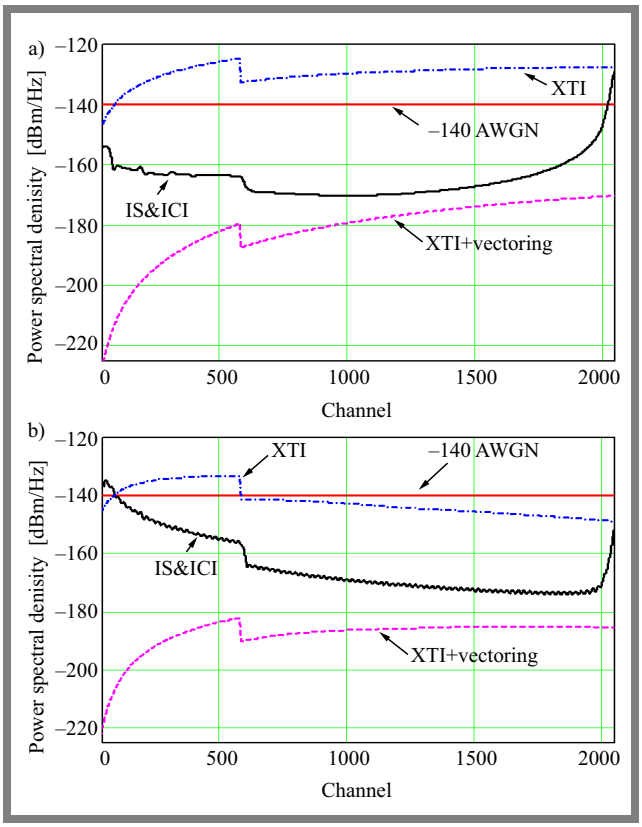

Fig. 3. PSD level of the G.fast system for: a) 50 m and b) 200 m UTP Cat. 5e line.

crosstalk is presented when the vectoring technique is applied (XTI + vectoring).

All interference power levels are compared to –140 AWGN, since under ideal TS conditions all types of interference are absent, with the exception of thermal noise, whose PSD level is –140 dBm/Hz. The presence of other interference sources increases the total interference power, and the contribution of each type of interference to the total outcome is determined by the ratio of its relative power.

The results presented in Figs. 2 and 3 allow us to conclude that cross-talk interference is the dominant type of interference, especially in the case of short lines. For a 50 m TPP cable, the PSD level of crosstalk interference exceeds thermal noise across almost all channels by approximately 25 dBm/Hz, while the IS+ICI level is significantly lower than that of –140 AWGN. The use of a crosstalk cancellation solution, such as the vectoring system, allows to reduce crosstalk levels to the –140 AWGN baseline.

For the 200-m line, a decrease in crosstalk interference is observed. However, for most G.fast channels, crosstalk still remains the dominant component, but the vectoring system lowers the crosstalk level below the –140 AWGN threshold. An increase in line length leads to greater linear distortion which, in turn, causes IS+ICI interference to rise on the lower channels over the –140 AWGN level.

For the UTP Cat. 5e cable, the qualitative conclusions remain the same. However, the quantitative assessments differ from those made for TPP, due to the distinct frequency characteristics of these cables (see Tab. 1). Lower intrinsic attenuation and higher crosstalk attenuation in the UTP Cat. 5e cable result in a lower impact of crosstalk interference (approx. 15 dB) and a higher effectiveness of the vectoring system (approx. 25 dB).

7. Analysis of G.fast Transmission Rate

In the next step, the achievable G.fast transmission rates over TPP-10×2×0.4 and UTP Cat. 5e 4×2×0.51 cables were evaluated under ideal conditions, i.e., in the absence of any interference other than thermal noise. This corresponds to the operation of a single TS (without crosstalk) over a line that is free from any linear distortions, i.e., without IS+ICI interference. The results of the transmission rate for line lengths ranging from 50 to 250 m are shown in Fig. 4.

G.fast operating on the UTP Cat. 5e cable shows a performance advantage in terms of the transmission rate. This is explained by the lower intrinsic attenuation of the UTP Cat. 5e cable. The result was expected, with the quantitative difference in transmission rates being the only unknown. In absolute values, the advantage of UTP Cat. 5e over TPP ranges from 14 to 250 Mbps, depending on the length of the line. In percentage terms, the advantage ranges from 1.2% for a 50-meter line to 73% for a 250-meter line.

Table 2 summarizes the results of the G.fast transmission rate calculations for operation over both cables. The transmission rate under ideal conditions, considering only thermal noise,

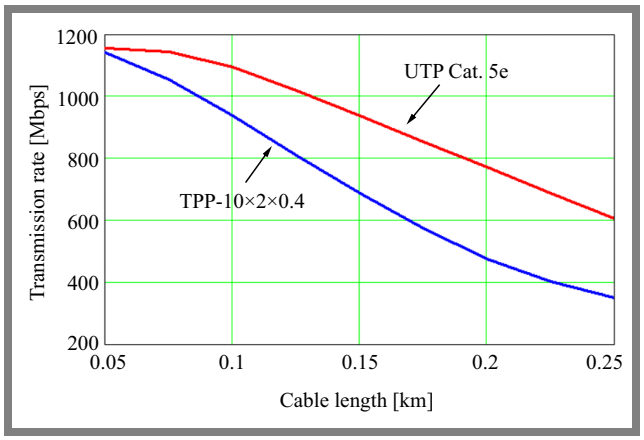

Fig. 4. G.fast system transmission rate over TPP-10×2×0.4 and twisted-pair UTP Cat. 5e cables in the absence of interference.

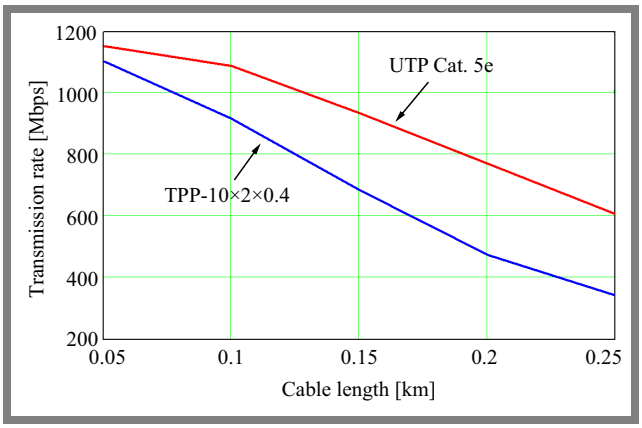

Fig. 5. G.fast system transmission rate over TPP and UTP cables, considering inter-symbol and inter-channel interference, with crosstalk interference compensation using the vectoring system (4 systems operating over a single cable).

–140 AWGN is denoted by R_0. The remaining parameters are as follows:

- R_{IS+ICI} – transmission rate considering –140 AWGN together with linear distortions leading to IS+ICI interference,
- $R_{IS+ICI+XTI}$ – transmission rate considering –140 AWGN, IS+ICI interference, and XTI from four G.fast systems operating in parallel over the cable,
- $R_{IS+ICI+vec}$ – transmission rate considering –140 AWGN, IS+ICI interference, and residual (uncompensated) crosstalk interference after applying the vectoring system.

The R_{IS+ICI} and $R_{IS+ICI+XTI}$ transmission rates are intermediate results that enable to evaluate the impact that IS+ICI and XTI interference exert on the performance of G.fast TS.

As expected, IS+ICI interference has a negligible effect on the G.fast transmission rate. For both cables, the decrease in transmission rate for R_{IS+ICI}, compared to R_0, does not exceed 2%.

The TPP-10×2×0.4 cable contains 10 pairs of wires, allowing up to ten G.fast systems to operate simultaneously. In contrast, the UTP Cat. 5e 4×2×0.51 cable contains 4 pairs, thus supporting up to four G.fast systems. To ensure equal comparison conditions in terms of XT interference influence, the transmission rate was evaluated, in both cases, assuming the operation of four systems.

Crosstalk interference has a significant impact on the G.fast transmission rate represented by $R_{IS+ICI+XTI}$. For the TPP cable, the reduction in $R_{IS+ICI+XT}$ relative to R_0 ranges from 37% to 55%, depending on the length of the line. For the UTP Cat. 5e cable, the reduction is less significant due to better mutual coupling characteristics (A_N and A_{ELF}), ranging from 5% to 13% depending on the length of the line.

In the case of $R_{IS+ICI+XT}$ transmission rates, as line length increases from 50 to 250 m, the transmission rate for the G.fast system over the TPP cable decreases from 505 to 218.5 Mbit/s, while in the case of the UTP cable, it decreases from 1010 to 578.8 Mbit/s. Therefore, the G.fast system operating over the UTP cable achieves transmission rates that are 2 to 2.6 times higher than those obtained over the TPP cable.

Such a significant limitation of the transmission rate due to crosstalk has necessitated the use of a crosstalk mitigation solution, known as the vectoring system.

The last column in Tab. 2 presents the calculated transmission rates with vectoring applied to compensate for crosstalk interference. The comparison of G.fast transmission rates for the two cable types is also shown in Fig. 5. These results account for the effects of IS+ICI and the residual XT interferences after vectoring and represent the final performance assessment of G.fast over TPP-10×2×0.4 and UTP Cat. 5e 4×2×0.51 cables.

From Tab. 2, one may conclude that the vectoring system is an effective technique for suppressing crosstalk. Compared to ideal conditions R_0, IS+ICI interference and uncompensated XT interference have a negligible impact on the performance of the G.fast system. When operating over the TPP cable, the transmission rate loss in the $R_{IS+ICI+vec}$ case ranges from 1.2 to 38 Mbit/s, which does not exceed 3.3% of the R_0 rate. For the UTP cable, the results are even more impressive. The loss of transmission rate in the $R_{IS+ICI+vec}$ case ranges from 1 to 3 Mbit/s, which does not exceed 0.4% of the R_0 rate.

Next, the transmission rates with the application of the vectoring technique between the TPP and the UTP Cat. 5e cables are compared.

As line length increases from 50 to 250 m, the G.fast system data rate over the TPP cable decreases from 1103 to 342.6 Mbit/s, while in the case of the UTP cable, it decreases from 1152 to 604.3 Mbit/s. Thus, when operating over the UTP cable, the G.fast system achieves data rates 5% to 43% higher than those obtained when operating over the TPP cable. Therefore, due to its superior characteristics, the UTP Cat. The 5e cable allows to achieve for more efficient performance of the G.fast system under all equal conditions.

Tab. 2. G.fast transmission rates for both cables [Mbps].

Length [m]	R_0	R_{IS+ICI}	$R_{IS+ICI+XTI}$	$R_{IS+ICI+vec}$
TPP-10×2×0.4				
50	1141	1141	505.056	1103
100	935.808	935.76	417.552	917.424
150	686.832	686.822	354.288	683.52
200	473.712	473.616	278.208	472.56
250	349.728	343.008	218.592	342.672
UTP Cat. 5e 4×2×0.51				
50	1155	1152	1010	1152
100	1093	1089	925.632	1089
150	937.056	935.472	842.544	935.472
200	770.016	769.632	725.472	769.632
250	605.136	604.464	578.832	604.304

8. Conclusions

Effectiveness of the G.fast technology significantly depends on the type of cable used in the network. The UTP Cat. 5e 4×2×0.51 cable provides a higher data transmission rate for G.fast systems compared to the TPP-10×2×0.4 telephone multipair cable, due to its lower attenuation and better immunity to crosstalk interference. The speed advantage is minor over short distances of 50 – 100 m, but increases with line length, reaching up to 73% at 250 m.

When more than one G.fast TS operates over a multipair cable, crosstalk interference becomes the dominant factor affecting the transmission rate. In the case of four parallel G.fast systems, the data rate loss may reach up to 55% on the TPP cable and up to 13% on the UTP cable. In such cases, the application of cross-talk interference compensation systems, such as vectoring, is mandatory. Under certain conditions, vectoring is capable of restoring nearly the full transmission rate that was achievable in the absence of crosstalk interference, especially when using UTP Cat. 5e cables.

IS+ICI interference has an insignificant impact on the transmission rate. The reduction does not exceed 2%, which means these types of interference are not a critical factor for deploying the G.fast technology in Ukraine's broadband access networks.

Based on the obtained estimates of achievable G.fast transmission rates, the following recommendations can be made:

- it is advisable to deploy G.fast on BB networks without upgrading the existing in-building cable infrastructure (preserving the TPP cable) if potential users are satisfied with a transmission rate of up to 900 Mbps at a distance of up to 100 m and up to 300 Mbps at a distance of up to 250 m,
- if potential users require higher access rates, it is necessary to upgrade the in-building cable infrastructure using UTP Cat. 5e cables. In this case, users will be able to achieve data rates of up to 1 000 Mbps at distances of up to 100 m and up to 600 Mbps at distances of up to 250 m.

References

[1] *Strategy for the Development of the Electronic Communications Sector of Ukraine*, Kyiv: Ministry of Digital Transformation of Ukraine, 2024.

[2] N. Gelvanovska-Garcia *et al.*, "Recommendations to the Ministry of Digital Transformation, Government of Ukraine on a National Broadband Strategy and Implementation Plan 2020-2025", World Bank, Washington, USA, 2020 (http://documents.worldbank.org/curated/en/896591621848142525).

[3] National Commission for the State Regulation of Communications and Informatization (NCCR), *Decision No. 298 dated 07.05.2025 on the determination of territories (settlements within their geographic boundaries) where access to universal electronic communication services (broadband Internet access services in a fixed location) must be provided*, 2025 (https://nkek.gov.ua/npas/298-07-05-2025).

[4] J. Valentín-Sívico, C. Canfield, S.A. Low, and C. Gollnick, "Evaluating the Impact of Broadband Access and Internet Use in a Small Underserved Rural Community", *Telecommunications Policy*, vol. 47, art. no. 102499, 2023 (https://doi.org/10.1016/j.telpol.2023.102499).

[5] European Commission, "Digital Decade 2024: Broadband Coverage in Europe 2023", Report, 2024 (https://digital-strategy.ec.europa.eu/library/digital-decade-2024-broadband-coverage-europe-2023).

[6] ITU-T Recommendation G.9700, "Fast Access to Subscriber Terminals (G.fast) – Power Spectral Density Specification", 2019 (https://www.itu.int/rec/T-REC-G.9700).

[7] ITU-T Recommendation G.9701, "Fast Access to Subscriber Terminals Access to Subscriber Terminals (G.fast) – Physical Layer Specification", 2019 (https://www.itu.int/rec/T-REC-G.9701).

[8] V.A. Balashov, V.I. Oreshkov, I.B. Barba, and V.V. Pedyash, "Speed Estimation of Broadband Access to Internet via xDSL Technology", *Radioelectronics and Communication Systems*, vol. 65, pp. 439–445, 2022 (https://doi.org/10.3103/S0735272722080052).

[9] D. Wei, A. Fazlollahi, G. Long, and E. Wang, "G.fast for FTTDp: Enabling Gigabit Copper Access", *2014 IEEE Globecom Workshops (GC Wkshps)*, Austin, USA, 2014 (https://doi.org/10.1109/GLOCOMW.2014.7063509).

[10] M. Timmers, M. Guenach, C. Nuzman, and J. Maes, "G.fast: Evolving the Copper Access Network", *IEEE Communications Magazine*, vol. 51, pp. 74–79, 2013 (https://doi.org/10.1109/MCOM.2013.6576342).

[11] R. Strobel and W. Utschick, "Coexistence of G.fast and VDSL in FTTDp and FTTC Deployments", *2015 23rd European Signal Processing Conference (EUSIPCO)*, Nice, France, 2015 (https://doi.org/10.1109/EUSIPCO.2015.7362554).

[12] J. Milanovic and D. Budisa, "Analysis of the Radiated Electric Field Strength from in-house G.fast2 Data Carrying Wire-line Telecommunication Network", *IET Science, Measurement and Technology*, vol. 15, pp. 478–485, 2021 (https://doi.org/10.1049/smt2.12048).

[13] Inter American Development Bank, *Broadband Policies for Latin America and the Caribbean: A Digital Economy Toolkit*, 2016 (https://doi.org/10.18235/0006525).

[14] Broadband Forum, "BBF.337 Gfast Certification Guidelines", 2025 (https://www.broadband-forum.org/testing-and-certification-programs/bbf-337-gfast-certification).

[15] Cabinet of Ministers of Ukraine, "Resolution of the Cabinet of Ministers of Ukraine on Approval of the Action Plan for Broadband Internet Development for 2021-2022", Order no. 1069-r, 2021 (https://zakon.rada.gov.ua/laws/show/1069-2021-#Text).

[16] Access to Fixed Internet: Data for 2021-2023 [Online] Available: (https://skilky-skilky.info/u-2023-rotsi-fiksovanyy-internet-maiut-62-domohospodarstva-zi-sta/)(inUkrainian).

[17] Letter No. 23 dated March 11, from the Ministry of Digital Transformation regarding remarks on the draft Cabinet of Ministers of Ukraine resolution "Issues of the National Broadband Status Platform", 2025.

[18] L.M. Liakhovetskyi, V.I. Oreshkov, and I.B. Barba, "Improvement of the Method for Evaluating the Transmission Rate of Communication Systems Using Orthogonal Harmonic Signals", *Scientific Works of O.S. Popov Odesa National Academy of Telecommunications*, no. 2, part 2, pp. 186–193, 2014 (http://nbuv.gov.ua/UJRN/Nponaz_2014_2_23).

[19] V.A. Balashov, V.I. Oreshkov, I.B. Barba, and I.V. Makarov, "Efficiency of Telecommunication Systems Transmission of Fixed Broadband Access Through Telephone Cables", *Proceedings of Odessa Polytechnic University*, no. 2, pp. 131–140, 2023 (https://doi.org/10.15276/opu.2.68.2023.14).

[20] ITU-T Recommendation G.993.5, "Self-FEXT Cancellation (vectoring) for use with VDSL2 Transceivers", 2019 (https://www.itu.int/rec/T-REC-G.993.5-201902-I).

[21] V.A. Balashov *et al.*, "Development of Fixed Wide-wide Internet Access in Ukraine", *Electronic Scientific Specialized Journal of Telecommunications Problems*, vol. 1, pp. 3–11, 2024 (https://doi.org/10.30837/pt.2024.1).

[22] V.A. Balashov, V.I. Oreshkov, I.B. Barba, and O. Iegupova, "Orthogonal Harmonic Signals of the Generalized Class", Journal of Telecommunications and Information Technology, vol. 83 pp. 64–70, 2021 (https://doi.org/10.26636/jtit.2021.146720).

Vitaliy Balashov, Professor
Department of Electronic Communications Systems
https://orcid.org/0000-0001-6122-4647
E-mail: vitaliybalashov8@gmail.com
State University of Intelligent Technologies and Telecommunications, Odessa, Ukraine
https://en.suitt.edu.ua

Vasyl Oreshkov, Ph.D.
Department of Electronic Communications Systems
https://orcid.org/0000-0001-9796-0216
E-mail: Oreshkov_VI@ukr.net
State University of Intelligent Technologies and Telecommunications, Odessa, Ukraine
https://en.suitt.edu.ua

Iryna Barba, Ph.D.
Department of Electronic Communications Systems
https://orcid.org/0000-0002-6751-0086
E-mail: irinabarba82@gmail.com
State University of Intelligent Technologies and Telecommunications, Odessa, Ukraine
https://en.suitt.edu.ua

Dmytro Stelya, M.Sc.
Department of Electronic Communications Systems
https://orcid.org/0009-0007-7330-6261
E-mail: dmutro-gepard@ukr.net
State University of Intelligent Technologies and Telecommunications, Odessa, Ukraine
https://en.suitt.edu.ua

Ihor Makarov, M.Sc.
Department of Electronic Communications Systems
https://orcid.org/0000-0002-4351-5122
E-mail: mackg9009@gmail.com
State University of Intelligent Technologies and Telecommunications, Odessa, Ukraine
https://en.suitt.edu.ua

Physical Layer Security for Keyhole-based NOMA Downlink Systems with a Multi-antenna Eavesdropper

Sang-Quang Nguyen[1] and Chi-Bao Le[2]

[1]*Posts and Telecommunications Institute of Technology, Ho Chi Minh City, Vietnam,*
[2]*Transcosmos Vietnam, Ho Chi Minh City, Vietnam*

https://doi.org/10.26636/jtit.2025.3.2116

Abstract — This paper investigates the physical layer security of downlink nonorthogonal multiple access (NOMA) systems operating over a degenerate keyhole channel in the presence of a multi-antenna eavesdropper. We propose a joint antenna selection framework with transmit antenna selection at the source and receive antenna selection at both legitimate users and eavesdroppers, thus striving to reduce hardware complexity while maximizing secrecy performance. In this framework, the efficacy of confidentiality is assessed for a specific user allocation methodology by deriving the closed-form approximate expression of secrecy outage probability (SOP). Extensive Monte Carlo simulations validate analytical results and reveal that increasing the number of antennas at the source and legitimate users dramatically lowers SOP, whereas a more capable eavesdropper raises the risk of secrecy. Our findings demonstrate that strategic antenna deployment and non-orthogonal access can effectively safeguard communications even through severely scattering environments.

Keywords — keyhole, multi-antenna, NOMA, physical layer security, secrecy outage probability

1. Introduction

Wireless systems are based on spatial diversity to boost capacity and reliability through utilization of multi-antenna techniques. In richly scattered environments, multiple transmit and receive antennas allow high spectral efficiency by creating uncorrelated transmission paths [1], [2]. However, when propagation is restricted, such as through a hallway, tunnel or narrow aperture, the so-called keyhole effect occurs and collapses the channel rank to one, degrading the benefits of using MIMO and reducing link capacity to SISO level [3]–[6].

Meanwhile, physical layer security (PLS) has emerged as a promising low-complexity approach allowing to protect wireless communications against eavesdropping, exploiting the randomness of fading channels to achieve secrecy without upper layer encryption [7], [8]. The authors of [9]–[11] studied PLS in keyhole-aided MIMO and cascaded fading scenarios. They also derived secrecy capacity and outage metrics under various relay and scheduling schemes. As a rule, these studies assume that orthogonal multiple access is relied upon and often neglect the impact of multiantenna eavesdroppers. By superimposing users signals with different power levels and employing successive interference cancelation (SIC), non-orthogonal multiple access (NOMA) can dramatically increase spectral efficiency and user connectivity in 5G and beyond networks [12], [13]. However, the security of NOMA under degenerate keyhole channels remains largely unexplored, especially when legitimate receivers and adversaries employ antenna selection to reduce hardware cost.

To address this gap, we investigate a downlink NOMA system where a multi-antenna source communicates through a single keyhole with two users and a multi-antenna eavesdropper. By combining transmit antenna selection (TAS) at the source with receive antenna selection (RAS) at the users and the eavesdropper, we derive tractable expressions for secrecy-outage probability (SOP) of both near and far users. To handle the resulting multidimensional integrals, we develop a Gauss-Chebyshev quadrature that converts them into finite sums with a negligible loss of accuracy.

The main contributions of this work are as follows:

1) We propose a method with joint TAS at the source and RAS at both legitimate users and the eavesdropper, balancing hardware simplicity against secrecy performance.
2) Using the Gauss-Chebyshev quadrature, we obtain the closed-form approximate expression of SOP expressions for both near and far NOMA users under a keyhole channel.
3) We benchmark the proposed NOMA design against a conventional orthogonal multiple access (OMA) baseline, demonstrating that NOMA superposition coding and SIC decoding yield substantially lower SOP for both users under identical antenna and power allocation settings.
4) Through extensive Monte Carlo simulations, we show how the number of antennas at the source, users, and the eavesdropper, as well as the keyhole's scattering cross section and the NOMA power split, interact to shape the SOP, offering practical guidelines for low complexity and secure deployments.

The remainder of this paper is organized as follows. Section 2 describes the system and channel models. Section 3 presents the secrecy-outage analysis and the quadrature approximation. Section 4 states the numerical results. Finally, Section 5 concludes the article.

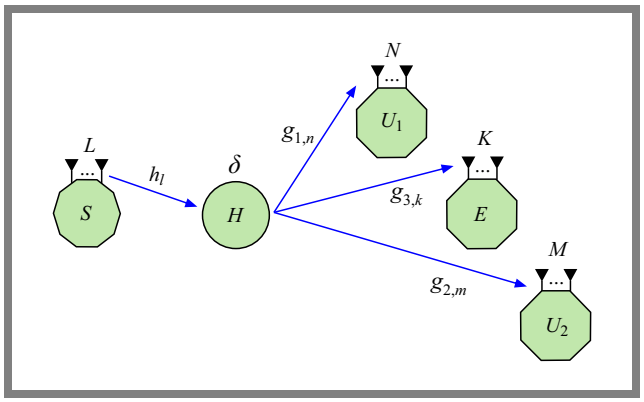

Fig. 1. Model of a keyhole-based NOMA system for downlink with a multi-antenna eavesdropper.

2. System Model

Consider a downlink NOMA communication system consisting of a source node S equipped with L antennas, a keyhole H with scattering cross-section δ, two legitimate users U_1 and U_2, and a potential eavesdropper E, as shown in Fig. 1.

Communication takes place through a keyhole with a scattering cross-section δ, separating the regions containing the users and the source. This keyhole models a propagation environment with limited scattering, such as a narrow passage or tunnel through which all signals must pass. Specifically, source S transmits information to users U_1 (near user) and U_2 (far user) via keyhole H.

Due to severe shadowing and scattering effects, direct transmission paths between the source and the users, as well as between S and E, are neglected. Therefore, all communications occur through the keyhole link.

User U_1 is equipped with N antennas, user U_2 with M antennas, and E with K antennas. Channels from H to U_1, U_2 and E are denoted by $g_{1,n}$, $g_{2,m}$, and $g_{3,k}$, respectively, while the channel from S to H is denoted by h_l. All wireless links in the network are assumed to be independent, non-selective block Rayleigh fading.

Furthermore, following common assumptions made in the literature, it is considered that the eavesdropper has complete knowledge of the relay transmission protocol and the user decoding strategies, enabling a worst-case security analysis.

The probability density function (PDF) and the cumulative distribution function (CDF) associated with random variables Z that follow an exponential distribution characterized by parameter Ω_Z can be written, respectively, as:

$$f_Z(x) = \Omega_Z^{-1} e^{-\frac{x}{\Omega_A}} , \tag{1}$$

$$F_Z(x) = 1 - e^{-\frac{x}{\Omega_A}} . \tag{2}$$

Following the NOMA principle, source S simultaneously transmits messages to both users U_1 and U_2 over the same time frequency resource by superimposing their signals with different power levels.

Specifically, the transmitted signal consists of a linear combination of two unit-power signals x_1 and x_2, U_1 and U_2, respectively. The resulting superimposed signal sent from the source can be expressed as [14]:

$$x = \sqrt{a_1}\, x_1 + \sqrt{a_2}\, x_2 , \tag{3}$$

where a_1 and a_2 are the power allocation coefficients such that $a_1 + a_2 = 1$, and typically $a_1 < a_2$ to ensure that the far user U_2 (with weaker channel conditions) receives a stronger signal.

We employ the antenna selection (AS) technique on both the transmitter and receiver sides, including the eavesdropper, to reduce system complexity while preserving its performance. At source S, which is equipped with L antennas, transmit antenna selection (TAS) is applied to choose the antenna with the strongest channel to the keyhole. Specifically, the selected transmit antenna index is given by:

$$l^* = \arg \max_{l \in \{1,\ldots,L\}} |h_l|^2 , \tag{4}$$

where h_l denotes the channel coefficient between the l-th antenna and the keyhole.

Similarly, both users and E apply the receive antenna selection technique. For user U_i, $i \in \{1,2\}$, the antenna with the strongest gain is selected as:

$$n^* = \arg \max_{n \in \{1,\ldots,N\}} |g_{1,n}|^2 , \tag{5}$$

$$m^* = \arg \max_{m \in \{1,\ldots,M\}} |g_{2,m}|^2 . \tag{6}$$

Similarly, the eavesdropper, equipped with K antennas, selects the best antenna via:

$$k^* = \arg \max_{k \in \{1,\ldots,K\}} |g_{3,k}|^2 . \tag{7}$$

The effective signal received at user U_i, $i \in \{1,2\}$ through the keyhole with antenna selection applied at both ends is given by:

$$\begin{aligned} y_1 &= \delta\, h_{l^*} g_{1,n^*} \sqrt{P_S}\, x + n_1 \\ &= \delta\, h_{l^*} g_{1,n^*} \sqrt{P_S} \left(\sqrt{a_1}\, x_1 + \sqrt{a_2}\, x_2\right) + n_1 \end{aligned} , \tag{8}$$

$$\begin{aligned} y_2 &= \delta\, h_{l^*} g_{2,m^*} \sqrt{P_S}\, x + n_2 \\ &= \delta\, h_{l^*} g_{2,m^*} \sqrt{P_S} \left(\sqrt{a_1}\, x_1 + \sqrt{a_2}\, x_2\right) + n_2 \end{aligned} , \tag{9}$$

where P_S is the total transmit power at S, x_1 and x_2 are the normalized NOMA signal with unit average power, i.e.

$$\mathbb{E}\left\{|x_1|^2\right\} = \mathbb{E}\left\{|x_2|^2\right\} = 1 ,$$

$$n_i \sim \mathcal{CN}\left(0, \sigma_i^2\right)$$

is the additive white Gaussian noise at user i in which $\mathbb{E}\{.\}$ denotes the expectation operation and $\mathcal{CN}(.\,,.)$ denotes the complex Gaussian distribution.

Similarly, the signal received at E is:

$$y_e = \delta\, h_{l^*} g_{3,k^*} \sqrt{P_S} \left(\sqrt{a_1}\, x_1 + \sqrt{a_2}\, x_2\right) + n_e . \tag{10}$$

where $n_e \sim \mathcal{CN}\left(0, \sigma_e^2\right)$ is the noise at E.

In this case, the instantaneous signal-to-interference-plus-noise ratio (SINR) at U_1 to detect x_2 for the keyhole link is

given by:

$$\gamma_{1,x_2} = \frac{\delta^2 P_S\, a_2\, |h_{l^*}|^2\, |g_{1,n^*}|^2}{\delta^2 P_S\, a_1\, |h_{l^*}|^2\, |g_{1,n^*}|^2 + \sigma_1^2} = \frac{\delta^2 \rho_S\, a_2\, |h_{l^*}|^2\, |g_{1,n^*}|^2}{\delta^2 \rho_S\, a_1\, |h_{l^*}|^2\, |g_{1,n^*}|^2 + 1}\,, \tag{11}$$

where $\rho_S = P_S/\sigma_i^2$ is the transmit signal-to-noise radio (SNR).

Suppose that U_1 can correctly cancel x_2, then, by performing SIC at U_1 to cancel signal x_2, the received SNR at U_1 to detect x_1 can be expressed by:

$$\gamma_{1,x_1} = \delta^2 \rho_S\, a_1 |h_{l^*}|^2\, |g_{1,n^*}|^2 + 1\,. \tag{12}$$

Similarly, SIC is required at U_2 to eliminate signal x_1, and SINR at U_2 can be computed to consider decoding x_2 as:

$$\gamma_{2,x_2} = \frac{\delta^2 \rho_S\, a_2\, |h_{l^*}|^2\, |g_{2,m^*}|^2}{\delta^2 \rho_S\, a_1\, |h_{l^*}|^2\, |g_{2,m^*}|^2 + 1}\,. \tag{13}$$

It is worth noting that SNR at E can be achieved by employing SIC as [15]:

$$\gamma_{E,x_2} = \delta^2\, a_2\, \rho_E\, |h_{l^*}|^2\, |g_{3,k^*}|^2\,, \tag{14}$$

$$\gamma_{E,x_1} = \delta^2\, a_1\, \rho_E\, |h_{l^*}|^2\, |g_{3,k^*}|^2\,, \tag{15}$$

where $\rho_E = P_S/\sigma_e^2$.

The normalized capacity per Hertz of bandwidth for both the user channel and the eavesdropper channel can be formulated as follows:

$$\mathcal{C}_{U_1}^{NOMA} = \log_2\left(1 + \gamma_{1,x_1}\right)\,, \tag{16}$$

$$\mathcal{C}_{U_2}^{NOMA} = \log_2\left(1 + \gamma_{2,x_2}\right)\,, \tag{17}$$

$$\mathcal{C}_{E,x_1}^{NOMA} = \log_2\left(1 + \gamma_{E,x_1}\right)\,, \tag{18}$$

$$\mathcal{C}_{E,x_2}^{NOMA} = \log_2\left(1 + \gamma_{E,x_2}\right)\,. \tag{19}$$

The secrecy capacity of keyhole-based NOMA systems for individual users U_i, $(i = 1, 2)$ can be defined as:

$$\mathcal{C}_i^{NOMA} = \left[\mathcal{C}_{U_i}^{NOMA} - \mathcal{C}_{E,x_i}^{NOMA}\right]^+\,, \tag{20}$$

where $[x]^+ = \max\{x, 0\}$.

Utilizing the order statistics, the CDF of $|h_{l^*}|^2$, $|g_{1,n^*}|^2$, $|g_{2,m^*}|^2$ and $|g_{3,k^*}|^2$ can be written as:

$$F_{|h_{l^*}|^2}(x) = 1 + \sum_{l=1}^{L} \binom{L}{l} (-1)^l\, e^{-\frac{lx}{\Omega_{h_l}}}\,, \tag{21}$$

$$F_{|g_{1,n^*}|^2}(x) = 1 + \sum_{n=1}^{N} \binom{N}{n} (-1)^n\, e^{-\frac{nx}{\Omega_{g_1}}}\,, \tag{22}$$

$$F_{|g_{2,m^*}|^2}(x) = 1 + \sum_{m=1}^{M} \binom{M}{m} (-1)^m\, e^{-\frac{mx}{\Omega_{g_2}}}\,, \tag{23}$$

$$F_{|g_{3,k^*}|^2}(x) = 1 + \sum_{k=1}^{K} \binom{K}{k} (-1)^k\, e^{-\frac{kx}{\Omega_{g_3}}}\,. \tag{24}$$

Differentiation Eqs. (21)–(24) yield the appropriate PDF as:

$$f_{|h_{l^*}|^2}(x) = \sum_{l=1}^{L} \binom{L}{l} \frac{l(-1)^{l+1}}{\Omega_{h_l}}\, e^{-\frac{lx}{\Omega_{h_l}}}\,, \tag{25}$$

$$f_{|g_{1,n^*}|^2}(x) = \sum_{n=1}^{N} \binom{N}{n} \frac{n(-1)^{n+1}}{\Omega_{g_1}}\, e^{-\frac{nx}{\Omega_{g_1}}}\,, \tag{26}$$

$$f_{|g_{2,m^*}|^2}(x) = \sum_{m=1}^{M} \binom{M}{m} \frac{m(-1)^{m+1}}{\Omega_{g_2}}\, e^{-\frac{mx}{\Omega_{g_2}}}\,, \tag{27}$$

$$f_{|g_{3,k^*}|^2}(x) = \sum_{k=1}^{K} \binom{K}{k} \frac{k(-1)^{k+1}}{\Omega_{g_3}}\, e^{-\frac{kx}{\Omega_{g_3}}}\,. \tag{28}$$

To demonstrate the dependability of such a system, we assess confidentiality performance based on SOP metrics. In the following sections, we derive the analytical formulations for the probabilities associated with SOP.

3. Secure Outage Probability

3.1. Probability of U_2

A secrecy outage occurs whenever the instantaneous secrecy rate of the far user U_2 falls below its target rate R_2. Equivalently, denoting $\theta_2 = 2^{R_2}$, we have:

$$\mathcal{S}_{U_2} = \Pr\left(\mathcal{C}_2^{NOMA} < R_2\right) = \Pr\left(\log_2 \frac{1 + \gamma_{2,x_2}}{1 + \gamma_{E,x_2}} < R_2\right)\,. \tag{29}$$

Substituting SINR expressions from Eqs. (13) and (14) yields the following:

$$\mathcal{S}_{U_2} = \Pr\left(\frac{\delta^2\, a_2\, \rho_S\, Z\, X}{\delta^2\, a_1\, \rho_S\, Z\, X + 1} < \theta_2\, \delta^2\, a_2\, \rho_E\, Z\, Y + \varsigma_2\right)\,, \tag{30}$$

where

$$X = |g_{2,m^*}|^2,\ \ Y = |g_{3,k^*}|^2,\ \ Z = |h_{l^*}|^2,$$

and

$$\varsigma_2 = \theta_2 - 1.$$

3.2. Integral Form Expression

By defining:

$$\mathcal{G}(z, y) = \frac{\theta_2\, \delta^2\, a_2\, \rho_E\, z\, y + \varsigma_2}{\delta^2\, \rho_S\, z\left[a_2 - a_1\left(\theta_2\, \delta^2\, a_2\, \rho_E\, z\, y + \varsigma_2\right)\right]}\,, \tag{31}$$

one shows by means of standard order statistics arguments that one must confront Eqs. (23) and (24) – that:

$$\mathcal{S}_{U_2} = \sum_{m=1}^{M} \sum_{k=1}^{K} \sum_{l=1}^{L} (-1)^{m+k+l} \binom{M}{m} \binom{K}{k} \binom{L}{l} \times \int_0^{\infty} \int_0^{\infty} e^{-\frac{ky}{\Omega_{g_3}}}\, e^{-\frac{lz}{\Omega_h}}\, F_X\left(\mathcal{G}(z, y)\right) dz\, dy,\,, \tag{32}$$

where:

$$F_X(x) = 1 + \sum_{j=1}^{M} (-1)^j \binom{M}{j} e^{-\frac{jx}{\Omega_{g2}}} . \tag{33}$$

3.3. Gaussian-Chebyshev Quadrature Approximation

Each inner integral

$$\mathcal{I}_{k,l} = \int_0^\infty \int_0^\infty e^{-\frac{ky}{\Omega_{g3}}} e^{-\frac{lz}{\Omega_h}} F_X\big(\mathcal{G}(z,y)\big)\, dz\, dy$$

is over $[0,\infty)$.

We map y, z to $[-1,1]$ via:

$$y = \frac{\Omega_{g3}(1+t)}{k(1-t)},\ z = \frac{\Omega_h(1+u)}{l(1-u)}, \quad t,u \in [-1,1] , \tag{34}$$

so that

$$dy = \frac{2(\Omega_{g3}/k)}{(1-t)^2}\, dt\ ,\ dy = \frac{2(\Omega_h/l)}{(1-t)^2}\, du\ , \tag{35}$$

and

$$e^{-ky/\Omega_{g3}} = e^{-\frac{1+t}{1-t}},$$
$$e^{-lz/\Omega_h} = e^{-\frac{1+u}{1-u}} .$$

Hence

$$\mathcal{I}_{k,l} = \int_{-1}^{1}\int_{-1}^{1} \mathcal{H}_{k,l}(t,u)\, dt\, du\ , \tag{36}$$

in which

$$\mathcal{H}_{k,l}(t,u) = \\ = 4F_X\big[\mathcal{G}\big(z(u), y(t)\big)\big]\ \frac{\Omega_{g3}\,\Omega_h\, e^{-\frac{1+t}{1-t}} e^{-\frac{1+u}{1-u}}}{k\,l\,(1-t)^2(1-u)^2}, \tag{37}$$

Applying a Gauss-Chebyshev quadrature of the second kind [16], we have:

$$t_q = \cos\left(\frac{2q-1}{2Q}\,\pi\right), \quad \omega'_q = \frac{\pi}{Q}\sqrt{1-t_q^2}, \quad q = 1,\dots,Q \tag{38}$$

and similarly $\{u_r, \omega'_r\}$. The double integral is approximated by the following formula:

$$\mathcal{I}_{k,l} \approx \sum_{q=1}^{Q}\sum_{r=1}^{Q} \omega'_q\, \omega'_r\, \mathcal{H}_{k,l}(t_q,u_r)\ , \tag{39}$$

where Q is a trade-off parameter between complexity and accuracy.

Putting everything together, the approximate closed-form expression of SOP for U_2 is given by:

$$\mathcal{S}_{U_2} \approx \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{l=1}^{L} (-1)^{m+k+l} \binom{M}{m}\binom{K}{k}\binom{L}{l} \\ \times \sum_{q=1}^{Q}\sum_{r=1}^{Q} \omega'_q\, \omega'_r\, \mathcal{H}_{k,l}(t_q,u_r) . \tag{40}$$

3.4. Secure Outage Probability of U_1

By direct analogy with the far-user case, one may formulate expressions for the near-user:

$$\mathcal{S}_{U_1} = \Pr\left(\mathcal{C}_1^{NOMA} < R_1\right) = \Pr\left(\frac{1+\gamma_{1,x_1}}{1+\gamma_{E,x_1}} < \theta_1\right) \\ = \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L} (-1)^{n+k+l} \binom{N}{n}\binom{K}{k}\binom{L}{l} , \\ \times \int_0^\infty\int_0^\infty e^{-\frac{ky}{\Omega_{g3}}}\, e^{-\frac{lz}{\Omega_h}}\, F_{X_1}(\Lambda(z,y))\, dz\, dy \tag{41}$$

where $\theta_1 = 2^{R_1}$ and:

$$\Lambda(z,y) = \frac{\theta_1\,\delta^2\, a_1\,\rho_E\, z\, y + \varsigma_1}{\delta^2\, a_1\,\rho_S\, z}, \tag{42}$$

$$F_{X_1}(x) = 1 + \sum_{j=1}^{N} (-1)^j \binom{N}{j} e^{-\frac{jx}{\Omega_{g1}}} . \tag{43}$$

In the above equations, $\varsigma_1 = \theta_1 - 1$.

We directly apply the Gauss-Chebyshev quadrature to avoid a separate derivation, as in Subsection 3.3, and define the same change of variables by:

$$y = \frac{\Omega_{g3}(1+t)}{k(1-t)}, \quad z = \frac{\Omega_h(1+u)}{l(1-u)}, \quad t,u \in [-1,1] , \tag{44}$$

and set

$$\mathcal{V}_{k,l}(t,u) = \\ = 4F_{X_1}\big[\Lambda\big(z(u), y(t)\big)\big]\ \frac{\Omega_{g3}\,\Omega_h\, e^{-\frac{1+t}{1-t}} e^{-\frac{1+u}{1-u}}}{k\,l(1-t)^2(1-u)^2}. \tag{45}$$

Using Q Chebyshev-II nodes $\{t_q, \omega'_q\}$ and $\{u_r, \omega'_r\}$, the double integral is approximated by the following:

$$\int_{-1}^{1}\int_{-1}^{1} \mathcal{V}_{k,l}(t,u)\, dt\, du \approx \sum_{q=1}^{Q}\sum_{r=1}^{Q} \omega'_q\, \omega'_r\, \mathcal{V}_{k,l}(t_q,u_r) . \tag{46}$$

Submitting Eq. (46) into Eq. (41), the approximate closed-form expression of SOP for U_1 is given as:

$$\mathcal{S}_{U_1} \approx \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L} (-1)^{n+k+l} \binom{N}{n}\binom{K}{k}\binom{L}{l} \\ \times \sum_{q=1}^{Q}\sum_{r=1}^{Q} \omega'_q\, \omega'_r\, \mathcal{V}_{k,l}(t_q,u_r) . \tag{47}$$

4. Numerical Results

In this section, we numerically evaluate our theoretical results concerning SOP performance. We now validate analytical formulas via Monte Carlo simulations with 10^7 independent trials and illustrate how key system parameters shape the secrecy-outage performance.

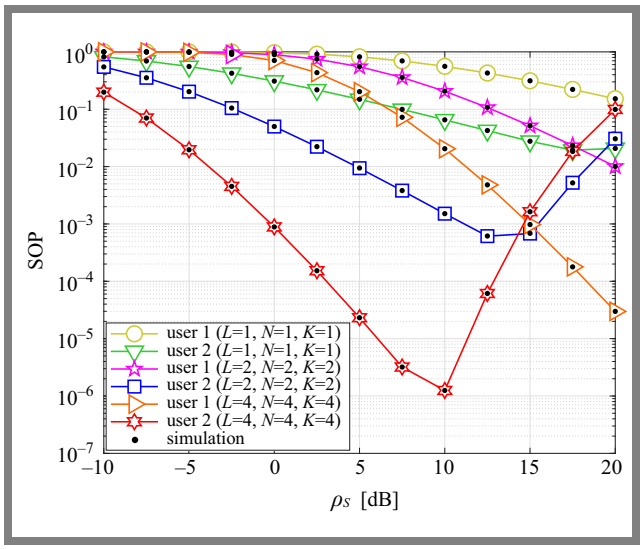

Fig. 2. SOP versus ρ_S for different L, N, M, and K values.

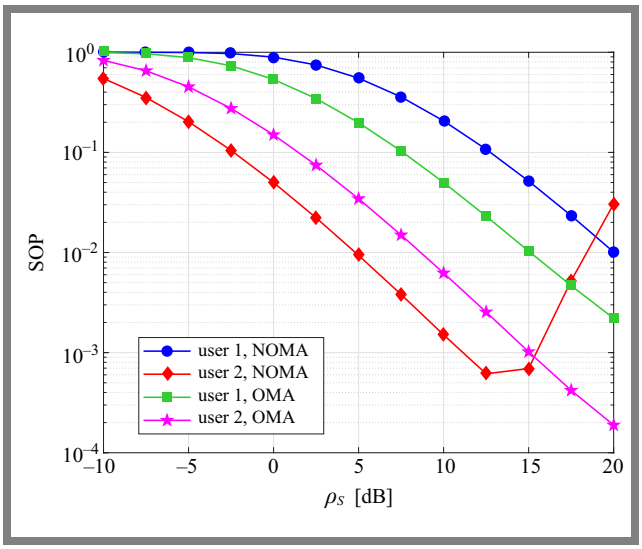

Fig. 3. Comparison between NOMA and OMA for SOP versus ρ_S with $L = N = M = K = 2$.

Unless otherwise noted, we set the power split $a_2 = 0.9$, $a_1 = 0.1$, target rates $R_1 = R_2 = 1$ bits per channel use (bpcu), scattering cross-section $\delta = 0.5$, and mean link gains $\Omega_h = 3$ dB, $\Omega_{g_1} = 6$ dB, $\Omega_{g_2} = 12$ dB, $\Omega_{g_3} = -20$ dB. For the Gauss-Chebyshev quadrature, we use $Q = 100$ nodes [17], which we found offers a near-perfect match to the simulation.

In Fig. 2, SOP is plotted against ρ_S in decibels for the different combinations of $L \in \{1, 2, 3\}$, $N \in \{1, 2, 3\}$, $M \in \{1, 2, 3\}$ and $K \in \{1, 2, 3\}$. As the number of source antennas L increases, both U_1 and U_2 experience a considerable reduction in SOP, showing better keyhole link diversity due to TAS.

Similarly, equipping the near-user with more antennas N achieves significant SOP gains through enhanced receive-side selection. The performance of far user M also rises noticeably with more antennas, though less steeply than for U_1, which manifests the NOMA power-split trade-off.

On the contrary, more antennas for eavesdropper K shift the SOP curves upward, which represents a more powerful adversary. All SOP curves drop off sharply in the low-to-mid-SNR range before saturating at some non-zero floor in the high-SNR regime. The nearer user U_1 always outperforms U_2 in the high SNR regime, as anticipated with SIC. Overall, Fig. 2 emphasizes that careful placement of antennas at legitimate nodes can greatly boost secrecy, even in the scenario in which a degenerate keyhole channel is exploited.

Figure 3 compares SOP performance of NOMA and OMA under identical antenna settings $L = N = M = K = 2$. As one may see, NOMA (blue and red curves) consistently outperforms OMA (pink and green curves) throughout the entire SNR range. For a given transmit power, the far user in NOMA achieves a lower SOP than its OMA counterpart, thanks to the superposition coding and SIC gains.

In the low-SNR regime ($\rho_S < 5$ dB), both access schemes suffer high outages, but the early slope is noticeably steeper in NOMA, indicating a faster improvement in secrecy as the power increases. Beyond the 10 dB level, NOMA's SOP declines to its asymptotic floor, whereas OMA lags by approximately one order of magnitude in outage. This gap persists even at high SNR, highlighting the enduring benefit of non-orthogonal resource sharing under keyhole fading.

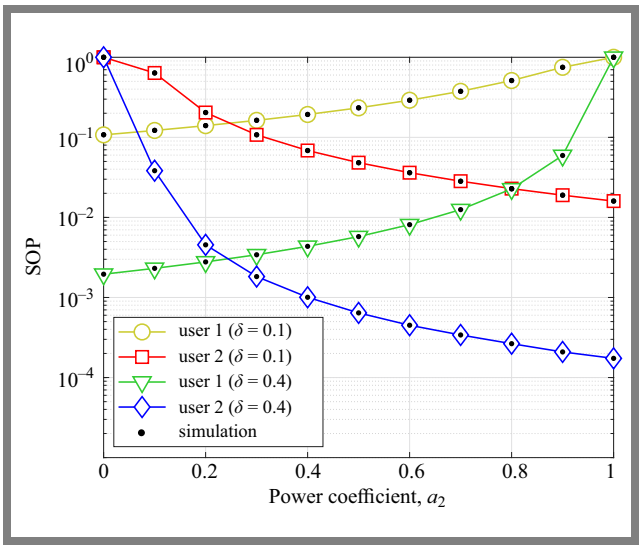

Fig. 4. SOP versus power coefficient a_2 with $L = N = M = K = 2$, $R_1 = R_2 = 0.1$ and $\rho_S = 5$ dB.

The near-user under NOMA further narrows the outage gap compared to the near-user curve, illustrating how power allocation favors the weaker link. Overall, Fig. 3 confirms that, when operating through a degenerate keyhole channel, NOMA not only boosts spectral efficiency, but also enhances physical-layer security relative to traditional OMA designs.

Figure 4 illustrates the impact of the NOMA power allocation coefficient a_2 on the SOP for both users when $L = N = M = K = 2$, $R_1 = R_2 = 0.1$ bpcu, and $\rho_S = 5$ dB. As a_2 increases from 0.1 to 0.9, the probability of outage of the far user initially decreases sharply, reflecting the stronger average power allocated to its signal, before flattening as a_2 approaches unity.

The simulation markers and Gauss-Chebyshev analytical curves once again overlap almost exactly, confirming the

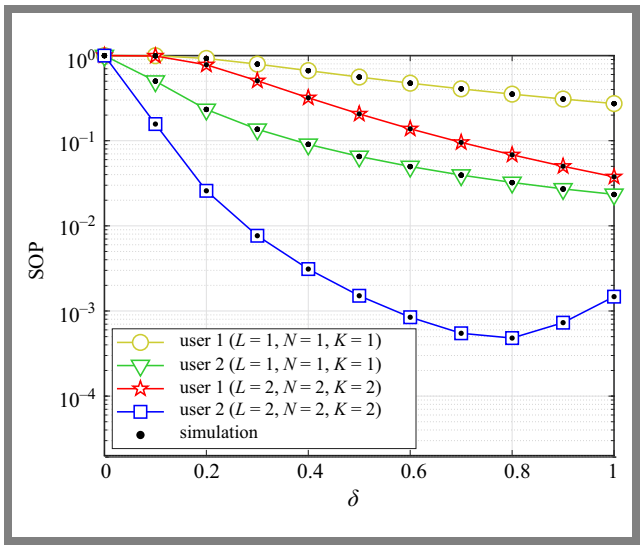

Fig. 5. SOP versus keyhole parameter δ, with $\rho_S = 10$ dB.

robustness of our approximation. Meanwhile, near-user U_1 exhibits a U-shaped SOP trend: too little power for U_2 (small a_2) forces U_1 to suffer large interference, while excessive a_2 starves U_1 of transmit power, also raising its outage.

Overall, Fig. 4 highlights that careful tuning of the NOMA power split can yield substantial secrecy gains in keyhole channels and that our quadrature-based analysis captures this behavior with a high degree of fidelity.

Finally, Fig. 5 examines how keyhole scattering cross-section δ influences SOP at a fixed SNR level of 10 dB with $L = N = M = K = 2$. As δ grows from 0.1 to 1.0, both U_1 and U_2 see their SOP curves drop steeply, reflecting the alleviation of multipath blockage through a "wider" keyhole. In particular, far user U_2 benefits more dramatically from increases in δ, while U_1 enjoys lower overall SOP due to its SIC advantage.

Beyond $\delta \approx 0.7$, SOP reduction levels off, approaching the asymptotic floor predicted by the high-SNR analysis. This plateau confirms that additional scattering offers diminishing secrecy gains once the keyhole ceases to be a severe bottleneck. Meanwhile, SOP mirrors the legitimate curves in reverse, improving as δ shrinks and worsening as it expands. Overall, Fig. 4 highlights that enhancing the effective aperture of the keyhole is a potent lever for bolstering physical layer security in NOMA systems.

5. Conclusions

We have presented a comprehensive study of keyhole-based NOMA downlink systems operating under the threat of a multi-antenna eavesdropper. By combining transmit and receive antenna selection with NOMA power splitting, we derived the closed-form approximate expression of SOP for both far and near users.

The proposed Gauss-Chebyshev quadrature method delivers fast and accurate approximations of these expressions, circumventing burdensome infinite integrals. Numerical results confirm that increases in source, near-user, or far-user antenna counts yield steep reductions in SOP, while a stronger eavesdropper degrades secrecy.

Additionally, the keyhole scattering cross-section and the NOMA power allocation balance play pivotal roles in determining outage floors. Overall, this work shows that even in highly constrained propagation scenarios, carefully designed antenna selections and non-orthogonal resource sharing are capable of significantly strengthening physical-layer security, paving the way for robust, low-complexity, secure communications in future IoT and 6G networks.

References

[1] A. Lozano and A.M. Tulino, "Capacity of Multiple-transmit Multiple-receive Antenna Architectures", *IEEE Transactions on Information Theory*, vol. 48, pp. 3117–3128, 2002 (https://doi.org/10.1109/TIT.2002.805084).

[2] W. Gao, X. Lu, C. Han, and Z. Chen, "On Multiple-antenna Techniques for Physical-layer Range Security in the Terahertz Band", *arXiv*, 2022 (https://doi.org/10.48550/arXiv.2201.06253).

[3] D. Chizhik, G.J. Foschini, M.J. Gans, and R.A. Valenzuela, "Keyholes, Correlations, and Capacities of Multielement Transmit and Receive Antennas", *IEEE Transactions on Wireless Communications*, vol. 1, pp. 361–368, 2002 (https://doi.org/10.1109/7693.994830).

[4] D. Chizhik, G.J. Foschini, and R.A. Valenzuela, "Capacities of Multielement Transmit and Receive Antennas: Correlations and Keyholes", *Electronics Letters*, vol. 36, pp. 1099–1100, 2000 (https://doi.org/10.1049/el:20000828).

[5] D. Gesbert, H. Bolcskei, D.A. Gore, and A.J. Paulraj, "Outdoor MIMO Wireless Channels: Models and Performance Prediction", *IEEE Transactions on Communications*, vol. 50, pp. 1926–1934, 2002 (https://doi.org/10.1109/TCOMM.2002.806555).

[6] S. Loyka and A. Kouki, "On MIMO Channel Capacity, Correlations, and Keyholes: Analysis of Degenerate Channels", *IEEE Transactions on Communications*, vol. 50, pp. 1886–1888, 2002 (https://doi.org/10.1109/TCOMM.2002.806543).

[7] H. Rahbari and M. Krunz, "Secrecy Beyond Encryption: Obfuscating Transmission Signatures in Wireless Communications", *IEEE Communications Magazine*, vol. 53, pp. 54–60, 2015 (https://doi.org/10.1109/MCOM.2015.7355566).

[8] M. Mitev, A. Chorti, H.V. Poor and G.P. Fettweis, "What Physical Layer Security Can Do for 6G Security", *IEEE Open Journal of Vehicular Technology*, vol. 4, pp. 375–388, 2023 (https://doi.org/10.1109/OJVT.2023.3245071).

[9] P. Almers, F. Tufvensson, and A.F. Molisch, "Keyhole Effect in MIMO Wireless Channels: Measurements and Theory", *IEEE Transactions on Wireless Communications*, vol. 5, pp. 359–3604, 2006 (https://doi.org/10.1109/TWC.2006.256982).

[10] H. Zhang *et al.*, "Performance Analysis of MIMO-HARQ Assisted V2V Communications with Keyhole Effect", *IEEE Transactions on Communications*, vol. 70, pp. 3034–3046, 2022 (https://doi.org/10.1109/TCOMM.2022.3163779).

[11] X. Zang, H. Xu, C. Ouyang, and H. Yang, "PHY Security in MIMOME Keyhole Channels", *IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, Beijing, China, 2023 (https://doi.org/10.1109/BMSB58369.2023.10211340).

[12] S.Q. Nguyen *et al.*, "Performance Evaluation of Downlink Multiple Users NOMA-able UAV-aided Communication Systems over Nakagami-m Fading Environments", *IEEE Access*, vol. 9, pp. 151641–151653, 2021 (https://doi.org/10.1109/ACCESS.2021.3124017).

[13] B.V. Minh *et al.*, "Performance Prediction in UAV-terrestrial Networks with Hardware Noise", *IEEE Access*, vol. 11, pp. 117562–117575, 2023 (https://doi.org/10.1109/ACCESS.2023.3325478).

[14] C.B. Le *et al.*, "Joint Design of Improved Spectrum and Energy Efficiency with Backscatter NOMA for IoT", *IEEE Access*, vol. 10, pp. 7504–7519, 2021 (https://doi.org/10.1109/ACCESS.2021.3139118).

[15] J. Chen, L. Yang, and M.-S. Alouini, "Physical Layer Security for Cooperative NOMA Systems", *IEEE Transactions on Vehicular Technology*, vol. 67, pp. 4645–4649, 2018 (https://doi.org/10.1109/TVT.2017.2789223).

[16] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York, USA: Dover, 1072 p., 1972 (ISBN: 9780486612720).

[17] Y. Cheng *et al.*, "Downlink and Uplink Intelligent Reflecting Surface Aided Networks: NOMA and OMA", *IEEE Transactions on Wireless Communications*, vol. 20, pp. 3988–4000, 2021 (https://doi.org/10.1109/TWC.2021.3054841).

Sang-Quang Nguyen, Ph.D.
https://orcid.org/0000-0002-1798-2296
E-mail: sangnq@ptit.edu.vn
Posts and Telecommunications Institute of Technology, Ho Chi Minh City, Vietnam
https://english.ptit.edu.vn

Chi-Bao Le, Student
https://orcid.org/0000-0002-3175-5698
E-mail: lechibao0411@gmail.com
Transcosmos Vietnam, Ho Chi Minh City, Vietnam
https://www.trans-cosmos.com.vn

Bio-inspired Routing Algorithms for UAV-based Networks: A Survey

Santosh Kumar[1], Amol Vasudeva[2], and Manu Sood[1]

[1]*Himachal Pradesh University, Shimla, India,*
[2]*Jaypee University of Information Technology, Solan, Himachal Pradesh, India*

https://doi.org/10.26636/jtit.2025.3.2101

Abstract — Rapid technological advancements, exponential growth, and unique characteristics are the key factors that enhance the usefulness of unmanned aerial vehicles (UAVs) in diverse applications, including military, agricultural, commercial, and communications-related fields. The use of UAVs for communication is a recent development that has become a topic of significant interest shown by researchers. A flying ad hoc network (FANET) made up of numerous UAVs cannot be developed without implementing an effective cooperative networking model that enables secure information sharing between UAVs. To achieve reliable and robust communication using FANETs, various design- and routing-related issues must be addressed in an appropriate manner. The use of bio-inspired algorithms for data routing in FANETs may be a promising direction, due to their ability to communicate efficiently in a swarm of devices. This work explores various bio-inspired routing algorithms proposed for transmitting data in UAV-based networks. Furthermore, their performance is evaluated and compared using routing metrics. All unresolved research concerns and prospective study avenues are examined based on the outcomes of the investigation conducted.

Keywords — bio-inspired routing protocols, FANET, routing protocols in FANETs, UAV

1. Introduction

The emergence of unmanned aerial vehicles (UAVs), ranging in size and capable of flying at any altitude, has become essential for a variety of existing and emerging applications. They are used in civilian and military settings, transport goods, take aerial photographs, manage traffic flows, participate in search and rescue operations, conduct surveillance missions and are parts of communication networks [1], [2]. Vehicles of this type may play multiple roles, e.g. those of aerial base stations, wireless relay systems or mobile sensors, owing to ease of their integration and flexibility [3], [4].

In a flying ad hoc network (FANET), drones can operate autonomously or cooperate as part of a specific mission [5], [6]. In stand-alone applications, UAVs can act as flying sensors or relay nodes communicating with ground control stations (GCS) or sink nodes.

However, their efficiency in performing various tasks is limited due to their mobility, interference, sparse coverage, the need for line of sight (LOS) communication, and limited energy supply [7].

Using multi-UAV or multi-drone systems can alleviate some of these limitations [8]. The need for enhanced protection and secure data transmission, together with improved overall network performance, is increasing the reliance on UAVs for coordination, collaboration and cooperation [9]. Adding more flying nodes enhances the range or size of the network used for transmission purposes in FANETs [10], which are more efficient than standalone UAV systems due to their scalability.

A FANET has few nodes directly linked to the infrastructure and each node operates in a mesh network [11]. A mesh network refers to a distributed communication network, where every node acts as both a transmitter and a receiver, forwarding information to other nodes. The nodes dynamically mesh together to create a dynamic, adaptive and self-organizing network that supports strong communication in dynamic mobile networks, such as FANETs, even when some nodes fail or change their location.

Nodes can use single-hop or multi-hop routing to connect to a GCS [12]. Multi-hop communication is a mechanism by which information is passed from a source node to a destination node via several intermediate nodes. Each node forwards the information to the next, facilitating communication across greater distances or beyond obstructions, thus increasing coverage and dependability within dynamic mobile networks. Efficient inter-node communication encourages collaboration and cooperation among flying nodes. In these ad hoc networks, the efficiency and credibility of the routing system determine how reliably communication may be conducted.

However, designing an effective communication system architecture for FANETs is a challenge. In FANET, every node participates in an ad hoc network, but only a few are connected to the infrastructure [13]. At the same time, the nodes can communicate with the GCS using single or multi-hop routing [14]. Due to their dynamic nature and deployment in demanding environmental conditions, FANETs require superior characterization compared to conventional ad hoc networks [15].

Reliable communication in an ad hoc network depends solely on the efficacy and trustworthiness of the routing protocol employed. A routing algorithm is a set of rules or procedures used by the nodes to determine the optimal path for transmitting data packets across the network to their destination. It accounts for factors such as node mobility, network topology changes, link quality, and energy constraints to ensure

efficient and reliable multi-hop communication in dynamic environments. Each ad hoc network has specific characteristics and challenges that need to be tackled when deploying it for in a particular application. Due to such unique characteristics as highly dynamic nature, resource limitations, uneven distribution, unpredictable movement, and deployment in complicated and harsh settings, networking protocols must be modified [16].

The design of FANETs mobility models and routing protocols is complicated [17]. Over the past few years, researchers have investigated several issues related to FANETs and have proposed many routing protocols and mobility models, especially for addressing frequent link interruptions triggered by dynamic and abrupt node mobility. Initially, some routing protocols from conventional ad hoc networks were modified to match the needs of FANETs.

1.1. Related Work

Bio-inspired routing protocols for FANETs have been the subject of increased research interest in the recent years due to the typical challenges posed by the high degree of mobility, dynamic topology, and energy limitations of UAVs. Bio-inspired algorithms, mimicking natural systems such as insect colonies, animal behavior, and evolutionary processes, have been investigated to solve these challenges by offering adaptive, scalable, and energy-efficient routing solutions. Furthermore, a limited number of scientific papers specifically address bio-inspired routing protocols for FANETs. Therefore, it is necessary to draw insight from the broader literature on general network-related issues, communication architectures, routing protocols, routing challenges specific to FANETs, and traditional ad hoc networks.

The authors of [18] performed an in-depth survey of bio-inspired routing techniques used in vehicular ad hoc networks (VANETs), comparing 48 algorithms across four categories. Their categorization framework demonstrated that bio-inspired techniques resulted in lower delays and higher packet delivery rates than those achieved with the use of legacy protocols in urban vehicular environments. Through large-scale simulations, the authors illustrated the adaptability of these protocols to dynamic topologies and their robustness under varying vehicular velocities. The paper is highly relevant to FANETs, as the shared challenges include high mobility, which VANET protocols naturally support through topology dynamics at high velocities; decentralization, where bio-inspired processes enable self-organizing routing and satisfy QoS demands, e.g., those more typical of delay-sensitive applications.

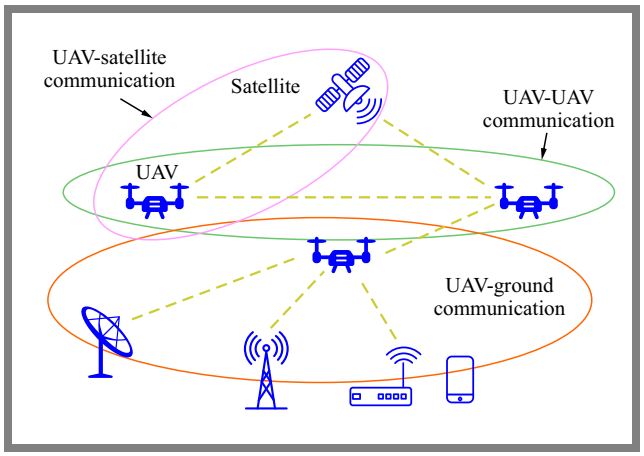

Fig. 1. FANET communication architecture.

In [19], a comprehensive review of new features, design challenges, and directions in the case of UAV-supported networks is presented. The paper is not explicitly focused on bio-inspired routing; however, its explanation of UAV communication constraints and remedies provides a starting point for researchers who plan to extrapolate VANET-inspired or biologically inspired solutions to FANETs.

Bio-inspired algorithms have been extensively researched for FANET routing, and valuable observations about their application have been presented in [20]. The research compares FANET with VANET and mobile ad hoc network (MANET) based on node mobility, density, and topology variation, and compares AntHocNet and BeeAdHoc with traditional protocols, like ad hoc on-demand distance vector (AODV) routing and dynamic source routing (DSR). It considers various parameters, e.g. scout ID, remaining energy for bee algorithms and pheromone-based parameters (e.g., α, β, ρ, η) for ant algorithms. NS-2.35 simulations were performed to illustrate performance under different scenarios, making the review of utmost importance for FANET routing studies.

In [21], a comprehensive review of FANET routing protocols is presented, dividing them into adaptive, proactive, reactive, and hybrid schemes. The paper includes swarm-based schemes, like those based on glowworm swarm optimization (GSO), that use bio-inspired ideas to cope with high dynamics and scarce energy. The authors contrast such schemes based on mobility parameters (e.g. speed, distance), link lifetime, location, available energy, and QoS parameters, such as packet delivery rate and delay, employing mobility models such as the random waypoint and random walk models. Their research emphasizes the flexibility of bio-inspired solutions for FANETs, although further studies are required on the details of specific protocols in order to solve 3D routing problems.

In their systematic review, the authors of [22] examine the use of bio-inspired algorithms for routing in FANETs. The research utilizes a systematic literature review to categorize bio-inspired routing techniques into hybrid and non-hybrid categories, focusing on their self-organizing and decentralized characteristics, which are particularly relevant to the high mobility and dynamic topology of UAV swarms. Compared to other existing algorithms, such as AntHocNet and BeeAdHoc, as well as traditional protocols, such as AODV and GPSR, the review also demonstrates their efficiency in improving packet delivery ratio (PDR), latency, and throughput.

Paper [23] introduces a comprehensive taxonomy of FANET routing protocols, such as topology-based, position-based, hierarchical, swarm-based and delay-tolerant networks (DTN) protocols. Swarm-based protocols, which utilize bio-inspired approaches, are contrasted based on their ability to handle

Tab. 1. List of acronyms and abbreviations used.

Acronym	Meaning	Acronym	Meaning
2D	Two-dimensional	HOPNET	Hybrid ant colony optimization routing
3D	Three-dimensional	HSCS	Hybrid self-organized clustering scheme
ACK	Acknowledgment	ID	Identification number
ACO	Ant colony optimization	IoT	Internet of Things
AI	Artificial intelligence	KH	Krill herd
AIS	Artificial immune system	LOS	Line of sight
AODV	Ad hoc on-demand distance vector	MAC	Media access control
APAR	Ant colony optimization-based polymorphism aware routing	MANETs	Mobile ad hoc networks
BAT-COOP	Bat algorithm using cooperation technique	PDR	Packet delivery ratio
BIA	Backward information ant	PeSOA	Penguin search optimization algorithm
BICSF	Bio-inspired clustering protocol for FANETs	PICA	Physarum-inspired clustering algorithm
BIR-SLB	Bio-inspired routing to support multimedia traffic in emergency conditions in FANET	PSO	Particle swarm optimization scheme
BR-AODV	On-demand routing using the Boids of Reynolds protocol	QoS	Quality of service
CH	Cluster head	RC	Route congestion
CH_DEC	Cluster head declaration	RREP	Route reply
CM	Cluster member	RREQ	Route request
CR	Cognitive radio	RS	Route stability
DF	Dragonfly scheme	RSF	Route selection function
DSR	Dynamic source routing	SI	Swarm intelligence
FANET	Flying ad hoc network	SIC	Swarm intelligence-based clustering
FEA	Forward exploration ant	SIL	Swarm intelligence-based localization scheme
GA	Genetic algorithm	SIL-SIC	Swarm intelligence-based localization and clustering
GCS	Ground control stations	SNRC	Signal-to-noise ratio combining
GSO	Glowworm swarm optimization	U2G	UAV-to-GCS
GW-COOP	Gray wolf algorithm using cooperative diversity technique	U2S	UAV-to-satellite
GWO	Grey wolf optimizer	U2U	UAV-to-UAV
HBA	Honey badger algorithm	UAVs	Unmanned air vehicles
HC	Hop count	VANETs	Vehicular ad hoc networks

high mobility and 3D routing. The paper introduces a novel taxonomy for reinforcement learning-based routing, contrasting the single-agent and multi-agent models. Although the lecture given does not mention specific parameters or simulators, the qualitative comparison with existing surveys highlights the prospect of bio-inspired protocols in FANETs, particularly for addressing dynamic topology issues.

The biologically inspired computation uses natural phenomena and available computer programs intended for solving various open research challenges. Therefore, this literature survey aims to study multiple bio-inspired routing protocols for FANETs and uncovers unresolved research issues. This work may serve as a reference point for those interested in solving open research challenges regarding efficient data routing in FANETs.

The paper is divided into six sections. Section 2 presents an outline of the FANET communication architecture, while Section 3 describes issues related to routing design. Section 4 discusses and compares various bio-inspired routing protocols. Section 5 provides a comparative analysis of bio-inspired routing protocols. Section 6 addresses several open research issues and future scopes. The article is concluded in Section 7. The acronyms and abbreviations used in this study are listed in Tab. 1.

2. FANET Communication Architecture

A consistent communication architecture in FANETs is essential because of the critical design requirements for internode

and intra-node communications. Each UAV or flying node participating in these networks can act as a source or destination. However, node-to-node communication exposes constraints such as frequent link disconnections, network fragmentation, and packet losses. Therefore, all FANET communications must follow a well-designed communication model to overcome these problems.

Furthermore, nodes can cooperate to cope with dynamic topology and networks can be reorganized using relay nodes. As illustrated in Fig. 1, the following three types of communication dominate in FANETs: UAV-to-UAV (U2U), UAV-to-GCS (U2G), and UAV-to-satellite (U2S).

In various application scenarios, UAVs communicate directly with one another by exchanging topology control messages. However, when the nodes are outside of each other's transmission range, multi-hop communication becomes essential to extend network coverage to specific areas of interest. LOS U2U communication is typically prevalent due to the minimal obstacles between UAVs flying in open-sky space.

However, there are areas for improvement associated with LOS U2U communication, such as the requirement for separate frequency bands for each pair and the challenges in critical remote missions where LOS communication is not guaranteed, mainly when UAVs fly in urban areas or mountainous regions [24].

A group of UAVs can be efficiently controlled through a fixed infrastructure on the ground, known as GCS, by exchanging topology updates and control messages using U2G communications [25]. LOS U2G communication remains dominant even at high elevations. However, UAVs flying at lower altitudes encounter numerous obstacles, decreasing the likelihood of LOS communication with GCS [26].

UAVs are frequently stationed in complex environmental settings, such as mountainous landscapes, water tanks, and forest areas, where GCS support is not feasible. Moreover, when a FANET suffers from severe network fragmentation and is unable to maintain connectivity with the GCS, centralized control is required to ensure permanent connectivity [27]. Allowing U2S communications can solve the issue by serving as a centralized controlled relay and providing LOS coverage [28].

Satellite communication is beneficial for critical data exchanges between drones and for transmitting information collected to the GCS positioned at a distant location on the ground [29]. However, this is costly, as maintaining communication with satellites requires additional hardware and expenses [30].

3. FANET Routing Objectives

Several issues must be overcome in a stable and reliable network before UAVs can be deployed efficiently. FANET communication protocols, especially routing protocols, have been explored to a comparatively lesser degree [31]. Although researchers have proposed many routing schemes for traditional ad hoc networks to address their characteristics and non-continuous connectivity, these protocols are not able to meet the communication needs of FANETs [32]. Therefore, effective routing schemes must be developed considering their distinctive characteristics. For efficient utilization of network resources, FANET routing schemes should be characterized by self-organizing and self-healing capabilities, robustness, and high proficiency.

The following routing goals must be considered when creating routing protocols for FANETs:

- Improving link stability. Link stability and lifespan are minimal due to the dynamic network topology, discontinuous connectivity, and highly fragmented networks [33]. Furthermore, FANET nodes are separated by comparatively longer distances, requiring frequent path discovery and link reestablishment processes [34]. Moreover, link stability depends on node density and the network's residual energy [35].
- Improving network coverage. In a FANET, the nodes are deployed at distant geographical locations, resulting in sparse networks [36]. Therefore, the nodes must have high transmission powers to provide efficient network coverage [37]. Otherwise, intermittent connectivity and a partitioned network will worsen coverage. Furthermore, making the network denser, for example in a multi-UAV communication scenario, improves network coverage [38].
- Improving quality of service (QoS) and routing performance. Lower link stability in a FANET requires frequent route discovery, which decreases routing performance [39]. Consequently, it is crucial to identify and select routes that have higher stability, longer lifespan, and fewer hops [40]. Furthermore, an increase in node density and transmission power can improve QoS [41].

4. Bio-inspired Routing Algorithms in FANETs

The network layer of the FANET communication model faces some critical challenges [42]. As stated earlier, FANET networking protocols should be developed and maintained with special attention paid to their specific characteristics and the complex environmental conditions in which they function.

Researchers proposed various routing protocols to adapt and satisfy conflicting design challenges, with high node mobility resulting in dynamic topology [43], efficient energy consumption [44], frequent link discontinuation [45], low security, poor scalability, and intelligent usage of allocated bandwidth and UAV resources [46].

However, fulfilling these requirements in a single protocol is nearly impossible. Hence, differentiating FANET routing protocols based on the deployment scenario is necessary. Based on the concepts, the central idea, the issues that need to be resolved and the techniques adopted, FANET routing protocols can be classified as network topology-aware, node position-aware, bio-inspired, stochastic, and beacon-less opportunistic. Given their versatility, adaptability, and potential

Tab. 2. Bio-inspired algorithms available.

Name	Description
Ant colony optimization (ACO) [55]	A probabilistic method in which graph-based solutions are utilized to solve issues related to route discovery. The pheromone-based information exchange mechanism of biological ants is used to find the optimal solution
AntNet routing [63], [64]	A routing scheme in which a group of mobile agents, sometimes known as artificial ants, attempts to establish connections between pairs of nodes by simultaneously scanning the network and exchanging data to update the routing tables
Artificial bee colony (ABC) [91]	It provides an optimal way to handle swarm-based communication by simulating the behavior of a honey bee swarm
Bacterial foraging optimization (BFO) [92]	A method of evolutionary optimization based on Escherichia coli bacteria's foraging behavior
Bees' wangle dance [76], [77]	A special dance by honey bees creates a precise, coded message to communicate the distance and direction of a new food source from the hive
Boids of Reynolds [68]	Boids are used to model the movement of an object in 3D and provide their geometric abilities
Bat echolocation [81]	The echolocation characteristics of bats are employed to sense the internode distance and surrounding obstacles. Also, they can differentiate between food/prey
Dragonfly scheme (DF) [79]	The DF algorithm makes use of dragonflies' static and dynamic swarming behaviors. Route-finding optimization is achieved through the application of both exploration and exploitation techniques
Glowworm swarm optimization (GSO) [52]	A technique for solving an optimization problem based on the luciferin value estimated by simulating the movement of a glowworm swarm around a luminescent quantity
Gray wolf optimizer (GWO) [56]	Gray wolves' leadership hierarchy is employed to implement cooperation in sparse networks
Honey badger algorithm (HBA) [89]	A metaheuristic optimization scheme modeled after the foraging habits of honey badgers, which are renowned for their tenacity and adaptive food search strategies
Krill herd (KH) [53], [54]	It mimics the herding behavior of individual krill and is capable of resolving a wide range of optimization issues across several domains
Moth flame optimization (MFO) [93]	A mechanism based on moths' innate transverse-orientation navigation method. Moths can fly fairly effectively at night, keeping a constant angle towards the moon
Penguin search optimization algorithm (PeSOA) [90]	An algorithm for metaheuristic optimization influenced by social behavior and cooperative hunting strategies of penguins, particularly their efficient foraging patterns in harsh environments
Physarum polycephalum (PP) [86]	An optimization technique derived from a single-celled fungus that demonstrates resource allocation efficiency and flexibility by self-organizing mycelium networks in complex natural environments
Particle swarm optimization (PSO) [59]	An optimal solution is obtained by iteratively improving the initial solution as per the required quality. The best local and global positions are explored by moving the particle swarm according to a mathematical formula in the search space
Red deer optimization (RDO) [94]	A metaheuristic optimization technique inspired by the strange mating habits of Scottish red deer in the breeding season
Swarm intelligence-based clustering (SIC) [60]	It offers an alternative method to achieve clustering in the absence of central control
Swarm intelligence-based localization (SIL) [58]	Provides a mechanism for the localization of highly mobile in a 3D environment

for efficient data routing, this study is limited to bio-inspired routing protocols within FANETs.

In recent years, bio-inspired technology has gained prominence as a research focus due to its capability to address various optimization challenges and deliver high performance [47]. With its unique features and bottom-up methodology, it primarily handles network challenges such as congestion control, security, and routing [48]. The various evolutionary algorithms inspired by nature are used to solve issues related to communication networks. These belong to two classes of

biologically inspired algorithms, namely swarm-based movement control algorithms and those that model the collective behavior of biological species [49].

Swarm intelligence (SI) refers to designing and implementing distributed problem-solving strategies inspired by social insect colonies and other animal societies that act collectively [50]. Researchers have proposed multiple bio-inspired algorithms based on swarm movement, mimicking honey bees, ant colonies, and bird flocks to provide an optimal solution to various networking problems [51].

Table 2 presents a range of bio-inspired algorithms that may effectively enhance routing efficiency in highly dynamic networks such as FANETs. The routing schemes based on bio-inspired algorithms proposed specifically for FANETs are mentioned in the upcoming subsection. These protocols are compared based on routing parameters such as adaptability, link stability, communication overhead, network coverage, efficiency, load balancing, security, and privacy to uncover open research issues.

4.1. Bio-Inspired Clustering Scheme for FANETs (BICSF)

BICSF is a clustering-based routing scheme for FANETs inspired by two biological algorithms: glowworm swarm optimization (GSO) [52] and krill herd (KH) [53], [54]. It comprises three phases: cluster formation, management, and maintenance. The GSO algorithm, aided by each UAV's residual energy and luciferin level, facilitates energy-efficient clustering and cluster head (CH) election.

Initially, as shown in Fig. 2a, each UAV autonomously calculates its fitness value F_i using:

$$L_i(t) = (1-\alpha)\, L_i(t-1) + \beta\, F\big(l_i(t)\big)\,, \tag{1}$$

$$E_{residual} = \sum_{i=1}^{n} \big(E_{initial}(i) - E_{current}(i)\big)\,, \tag{2}$$

$$F_i = \gamma\, E_{residual} + (1-\gamma)\, L_i(t)\,, \tag{3}$$

where for each glowworm i, $L_i(t)$ signifies its luciferin value, α represents the luciferin decay constant, β represents the luciferin perfection fraction, and $F\big(l_i(t)\big)$ is the glowworm goal function for the UAV i at current location l_i.

UAVs broadcast F_i, and the UAV with the maximum F_i becomes the CH. Low-energy UAVs (e.g. UAV 8 in Fig. 2a) are excluded to prioritize network longevity.

In the second phase, the KH algorithm ensures stability and efficient cluster management by modeling cluster members' communication and virtual positioning relative to the CH, inspired by krill swarming patterns. This guides alignment with the CH, ensuring cohesive and stable intra-cluster interactions. Genetic operators, including crossover and mutation, are utilized to optimize communication paths between members and enhance adaptability in response to dynamic topology changes.

Route selection employs a function that evaluates the count of adjacent UAVs, remaining energy, and the space between UAVs, ensuring efficient data transmission:

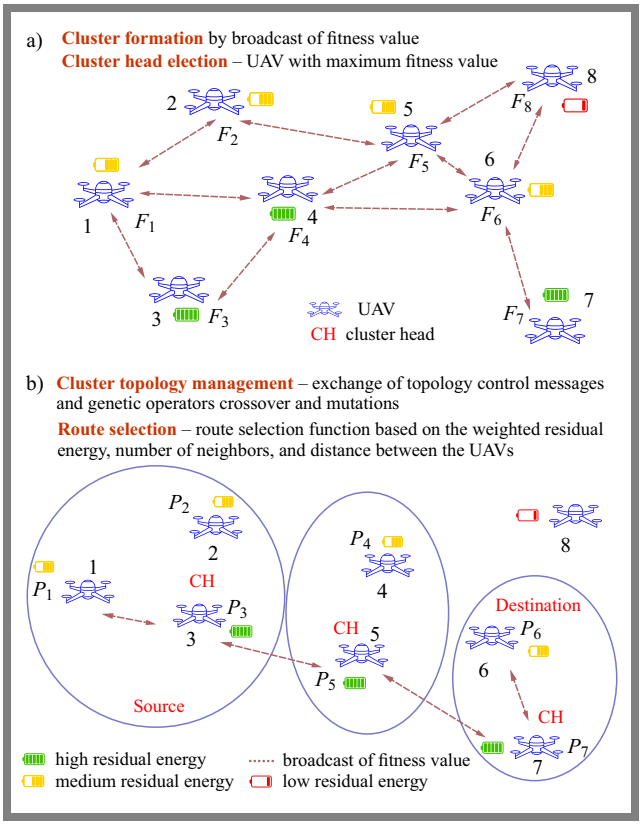

Fig. 2. BICSF algorithm demonstrating: a) cluster formation and cluster head election, and b) cluster topology management with route selection.

$$Route\ selection = \frac{\omega_1\, E_{residual}}{(\omega_2 \cdot N_i)(\omega_3 \cdot d)}\,, \tag{4}$$

where ω_1, ω_2, and ω_3 represent the weights associated with residual energy, N_i represents the count of UAVs in the neighborhood, and d is the distance between the UAVs.

After cluster formation, the CH manages intra-cluster communication by collecting and processing data from CMs and handles inter-cluster communication by forwarding data to neighboring CHs or the destination, as depicted in Fig. 2b. During the maintenance phase, UAV energy levels are periodically monitored. CMs with energy levels above a set threshold remain, while those below are removed as dead nodes, potentially necessitating re-clustering to sustain network performance.

The scheme described in [48] involves the following steps.

- UAVs initiate distributed clustering, each calculating fitness F_i based on luciferin and residual energy levels.
- Each UAV broadcasts a hello message with F_i.
- Upon receiving the messages, F_i builds, updates, and sorts a neighbor table. The table is continuously updated and sorted with the arrival of new hello messages.
- With three or more entries, the nodes check for the highest F_i. If it is the highest, they declare self-CH. Otherwise, they acknowledge the UAV with the highest F_i as CH, via a formation message.
- The remaining nodes join as CMs.

- The topology is periodically updated to ensure stability and energy efficiency.

The authors of [48] evaluated efficiency of BICSF against ant colony optimization (ACO) [55] and gray wolf optimizer (GWO) [56] approaches, demonstrating that BICSF requires less time for cluster formation compared to these alternatives. BICSF is highly adaptable, adjusting to dynamic node positions via bio-inspired algorithms. However, it exhibits low link stability due to the mobility of UAVs, which impacts reliability. It ensures a single stable route but lacks multiple routes, limiting fault tolerance. The scheme incurs high routing overhead resulting from frequent updates and computational complexity of GSO and KH. It requires substantial memory to route the data and is energy efficient during transmission, but lacks load balancing, leading to uneven energy depletion. BICSF does not prioritize privacy or security mechanisms, making the network susceptible to attacks such as data interception or unauthorized access.

The coverage of a BICSF network is based on clusters and can be expanded by CH using multi-hop communication. For the sake of enhancing cluster stability and reducing routing overhead, BICSF makes moderate assumptions about UAV mobility and selects highly reliable and connected UAVs as CHs. The UAVs are relocated to cover holes.

Area coverage is determined by clustering algorithms, making it computationally complex. By simulating krill movement to adapt to topological changes, the KH algorithm ensures stable coverage, and GSO iteratively optimizes the cluster assignments. Coverage is estimated by neighbor density using RSSI-based distance measurements. This approach incurs complexity of $O(n \log n)$ per cluster.

The advantages of BICSF are as follows:

- GSO helps in energy-efficient cluster management and makes the proposed protocol adaptable to topology changes.
- The KH-inspired clustering method ensures stable intra-cluster communication, even with limited node movement, thereby extending the lifespan of clusters.
- This scheme requires less cluster formation time and provides a longer cluster lifespan.

The limitations of BICSF can be summarized as:

- High routing overhead owing to frequent cluster head reelections.
- Limited scalability in dense UAV networks.
- Cluster instability under high mobility.
- Moderate computational complexity due to dynamic clustering using GSO and KH.
- No support for time-sensitive data delivery.

4.2. Swarm Intelligence-based Localization and Clustering (SIL-SIC)

SIL-SIC [57] is a hybrid of two bio-inspired techniques, namely the swarm intelligence-based localization (SIL) [58] and particle swarm optimization (PSO) [59] schemes. The SIL method analyzes the distance between existing connector nodes to estimate the destination's position that is randomly dispersed in a 3D search space. This scheme uses a bounding box approach to exploit the particle search space inside a restricted boundary.

Furthermore, the authors of [60] presented a clustering method based on PSO-based energy-proficient swarm intelligence (SIC). As shown in Fig. 3a, UAVs dynamically adjust their positions based on their local estimates and global information shared through the localization data within the swarm. UAVs evaluate their fitness value F_v based on multiple factors, including inter- and intra-cluster distances, the network's residual energy, and geographical locations. As depicted in Fig. 3b, UAVs with the highest F_v are selected as CHs.

The PSO method may introduce localization errors due to the random initial positions of particles, determined by their personal best p_{best} and information from neighboring particles, potentially leading to the exploration of unnecessary or irrelevant areas in the search space. Additionally, the g_{best} solution may converge prematurely, limiting exploration and resulting in suboptimal localization outcomes. The bounding box method enhances the efficiency in limiting the search space within a restricted boundary. The restricted search is marked by identifying multiple anchor UAVs within the communication range of an unknown UAV.

The objective of the SIC method is to maintain a uniform cluster size, as shown in Fig. 3c. UAVs iteratively update their distances and cluster centers to optimize node assignment using the following formula:

$$d = \sqrt{\left((x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2\right)}\,, \qquad (5)$$

where d is the distance between two UAVs and coordinates, x_1, y_1, z_1 and x_2, y_2, z_2 define the location of the first and second UAVs, respectively.

Clustering is done based on proximity to CH and inter-cluster distances. Additionally, maintaining the cluster size above a threshold value ensures a balanced cluster size. CMs are assigned to the nearest cluster. For instance, as shown in Fig. 3c, UAV 6 migrated to cluster 2 from cluster 1 to balance cluster sizes. After clustering, inter-CH data transmission is done using multi-hop routing [61], as depicted in Fig. 3d.

The scheme [57] involves the following steps:

- UAVs in the swarm start at random locations P_i with velocities V_i.
- Each UAV calculates its fitness value F_v using P_i and V_i. Optimal F_v = Personal best p_{best}.
- p_{best} is used to evaluate the best global position g_{best} for the entire swarm. Each UAV adjusts its trajectory according to its p_{best} and g_{best}.
- Center points are initialized randomly. These serve as the initial reference points for each cluster. Each UAV calculates its distance from different cluster centers using the Euclidean distance formula. The node is associated with a cluster with the nearest central point.

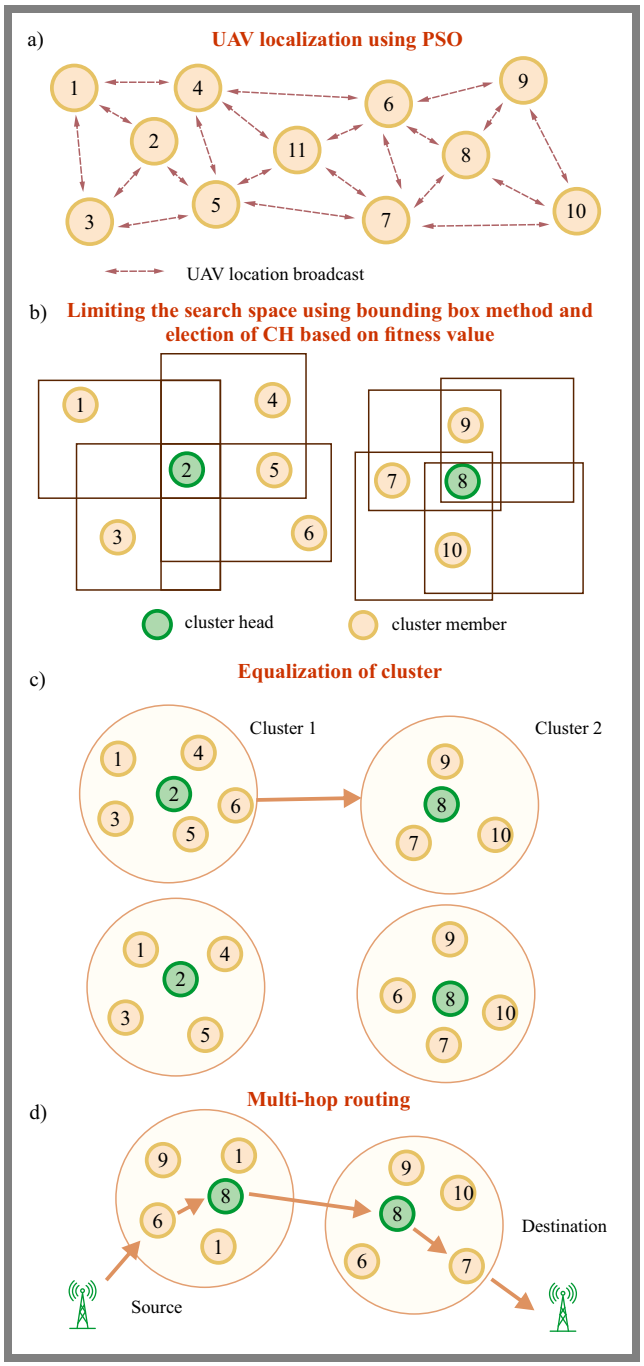

Fig. 3. Illustration of the SIL-SIC scheme for FANETs: a) dynamic position adjustment of UAVs in a 3D search space using SIL, b) CH selection based on F_v, c) cluster balancing with node migration, and d) multihop routing for data transmission.

- Inter-cluster distances are measured between all clusters and a distance matrix is constructed. This matrix stores the distances between each pair of clusters.
- Clusters are merged based on proximity until stable clusters are formed, ensuring efficient communication and swarm coordination.
- The optimal UAV is chosen as CH based on the UAV closest to the center of its cluster and the node with a higher residual energy.
- p_{best} and g_{best} are updated periodically.
- Data are transferred using multi-hop routing, where each CH communicates with the next CH to route data to GCS.

Based on simulation findings described in [57], due to its accurate localization, the SIC method outperforms conventional approaches in terms of routing overhead, packet delivery ratio (PDR), and average delay. SIL-SIC exhibits high adaptability and responds flexibly to network topology and UAV position modifications through its PSO-based clustering and routing approach, which enhances link stability and network coverage by optimizing UAV positions and communication paths. However, SIL-SIC focuses solely on single-path routing, limiting redundancy and fault tolerance.

Despite efforts to optimize overhead, SIL-SIC experiences considerable routing overhead due to the demands of the algorithms and frequent updates required to maintain dynamic clustering. The computational overhead of executing these algorithms and the storage requirements for particle positions and routing information contribute to space complexity. Frequent updates in positions and clusters can lead to significant energy consumption. Although SIL-SIC incorporates load-balancing features to evenly distribute the network load, it does not prioritize privacy or security mechanisms, leaving the network vulnerable to potential attacks, such as data intercepting or unauthorized access.

The 3D SIL algorithm, designed for emergency communication in highly mobile environments, utilizes PSO to determine the optimal location of the UAVs within a 3D bounded space, positioning them to cover the target area effectively. UAVs identify coverage holes through RSSI-based triangulation with anchor nodes. The SIC algorithm then clusters UAVs based on geolocation, inter- and intracluster distances, and RE. Clustering is utilized to partition the network into manageable clusters, with each cluster covering a portion of the operational area.

A pheromone-based coverage map is used to compute the coverage area. UAVs deposit virtual pheromones in visited areas, and low-pheromone regions cause member redistribution. Complexity $O(n^2)$ arises due to a pairwise comparison, which restricts deployment in emergency cases involving large sizes, where computational latency is higher.

The advantages of SIL-SIC can be summarized as:

- Use of the PSO scheme in SIL-SIC enables precise localization of UAVs.
- Effective clustering improves energy efficiency and extends network's life.

The limitations of such an algorithm are:

- High routing overhead due to swarm coordination and localization.
- Performance degrades with the addition of UAVs.
- Cluster instability due to high mobility.
- High computational overhead is incurred due to the PSO algorithm used for position estimation.
- Due to frequent clustering, time-sensitive data delivery is not supported.

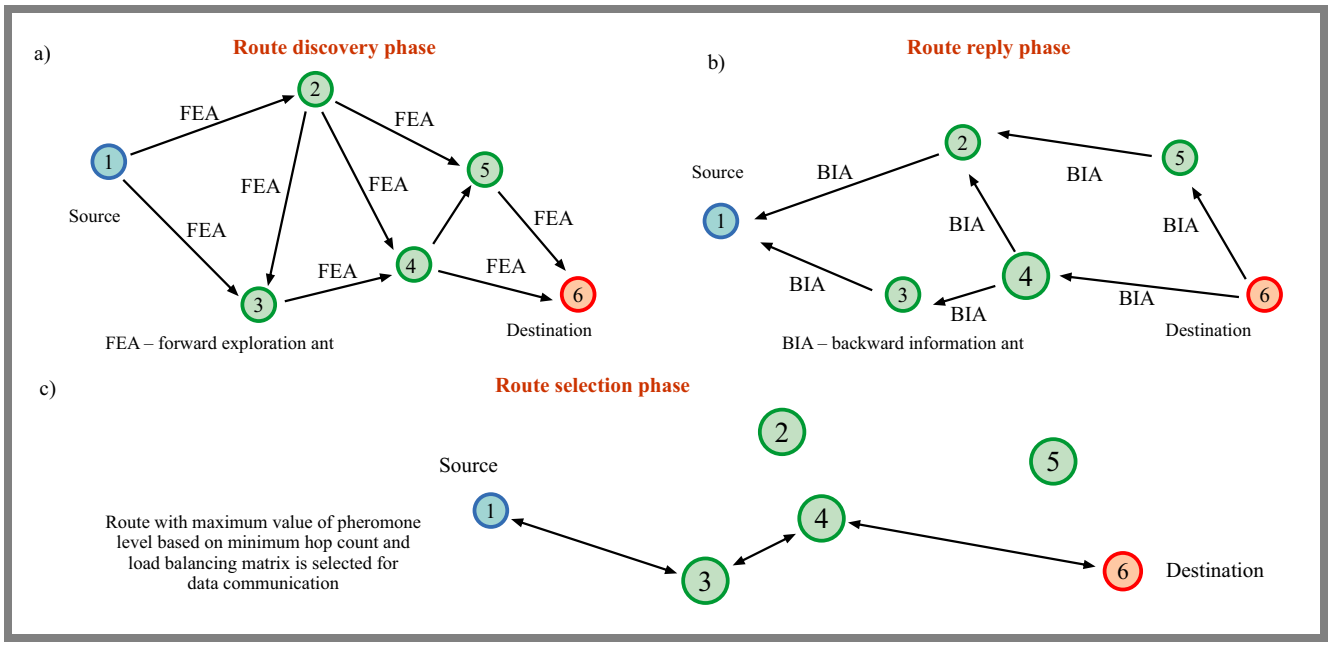

Fig. 4. BIR-SLB scheme for FANETs demonstrating: a) route discovery phase, b) route reply phase, and c) route selection phase.

4.3. Bio-inspired Routing to Support Multimedia Traffic under Emergency Conditions in FANETs (BIR-SLB)

BIR-SLB is a bio-inspired routing scheme that employs multi-objective optimization metrics, focusing on shortest path and load-balancing [62]. These metrics evaluate potential routes and select the most optimal one for routing multimedia traffic. The proposed approach inherits certain behaviors from the AntNet routing protocol [63], [64].

Specifically, in very dynamic networks, such as UAV-based mesh networks, identifying the optimal path is challenging due to constant changes in topology. To address this, BIR-SLB optimizes the routing process by decreasing path length while considering the traffic load across the routes. This guarantees that the network will continue to function effectively even in high-traffic situations.

The scheme comprises three phases:

Route exploration. In the first phase, the source drone initiates probabilistic forward exploration ant (FEA) communication to explore potential paths toward the destination (as shown in Fig. 4a). FEA packets traverse various routes, collecting vital network information such as hop count (HC), link quality, and available bandwidth. This phase allows drones to probe the network and identify viable routes.

Route reply. In the second phase, after an FEA packet reaches its destination, the destination generates a backward information ant (BIA) packet that retraces the explored route back to the source (as depicted in Fig. 4b). During this reverse traversal, BIA collects and relays critical performance metrics, such as latency, energy consumption, and network congestion at each node. This phase ensures that the source drone receives valuable feedback to evaluate the quality and efficiency of each path.

Route reinforcement and load balancing. In the final phase, the BIA packet incrementally updates the pheromone levels on each link as it follows the reverse route to the source (as illustrated in Fig. 4c). The pheromone concentration reflects the reliability, energy efficiency, and traffic load of each route. Routes with higher pheromone levels are reinforced and are more likely to be selected for future transmissions. This phase dynamically balances network load by distributing traffic across multiple paths, promoting routes that are not only shorter but also energy efficient and less congested, and thus optimizing overall network performance.

The scheme described in [62] proceeds as follows:

- UAVs are deployed in the emergency area. Each source UAV broadcasts an FEA message towards the destination.
- FEA packets travel probabilistically, exploring potential routes to the destination.
- When the destination receives the FEA message, it confirms this by generating a BIA message.
- The BIA backtracks to the source following the reverse path of the original FEA message.
- The source node chooses the most efficient path to route its multimedia data based on the optimal values of link quality, pheromone value, and HC.
- This process is repeated periodically, ensuring that the network can adjust to changes in topology caused by UAV movement or link alteration failures.

The authors of [62] evaluated the performance of BIR-SLB by comparing it with link-state routing protocols [65]. They observed that the proposed scheme is more adaptable and requires less overhead than link-state routing, which has the maximum flow and a single shortest path.

BIR-SLB is highly adaptable and efficiently handles changes in network topology and node position. It benefits from high

link stability because it focuses on path length optimization, leading to more reliable connections. However, BIR-SLB does not support multiple routes, which limits redundancy and fault tolerance within the network. Due to the minimal use of control messages, the scheme keeps a low routing overhead, leading to a balanced path length and traffic load.

Furthermore, the availability of optimal routes leads to low computational and communication overheads. This scheme does not emphasize energy efficiency, privacy, and security, making networks vulnerable to security threats.

Designed for multimedia support in dynamic environments, BIR-SLB balances load distribution through an ant-inspired mechanism. BIR-SLB implements pheromone-based coverage estimation where "coverage ants" deposit virtual pheromones, proportionally to RSSI measurements during path exploration. The buffered data is forwarded to uncovered regions via ants.

The ACO algorithm is utilized to identify coverage holes, i.e. regions with pheromone values below adaptive thresholds. The algorithm executes $O(k \cdot n)$ operations per agent (k – number of iterations, n – number of agents) for the map convergence with bandwidth overhead due to periodic rebroadcasts of the pheromone table in mobile environments.

The advantages of BIR-SLB are:

- Due to the availability of many routes to distribute calls, BIR-SLB can handle many calls and efficiently distribute call requests among the UAV-based network.
- Due to the low cluster build time, this protocol shows better route exploration and scalability.

The limitations of BIR-SLB can be summarized as follows:

- Less effective in dense networks.
- Route instability due to high mobility of the UAV.
- No support for real-time packet delivery, limiting its applicability to emergency scenarios.

4.4. On-demand Routing Using Boids of Reynolds Protocol (BR-AODV)

BR-AODV is a biologically inspired upgrade of the traditional AODV routing scheme [66] for UAV-based networks [67]. The proposed scheme merges AODV, a conventional reactive routing protocol, with the Reynolds Boids [68] algorithm to ensure dynamic path connectivity and route maintenance during data transmission.

The protocol operates an on-demand basis, as routes are discovered only when needed rather than maintaining pre-established routes. This approach helps reduce unnecessary overhead, making the protocol more efficient for UAV networks, where nodes frequently move and change their positions.

The node activity monitoring process is illustrated in Fig. 5a. A timer is connected to each UAV inside the network for monitoring purposes. After a silence period T, UAVs become active or inactive. This is the period in which an UAV neither transmits nor receives any data. The remaining rules are as follows:

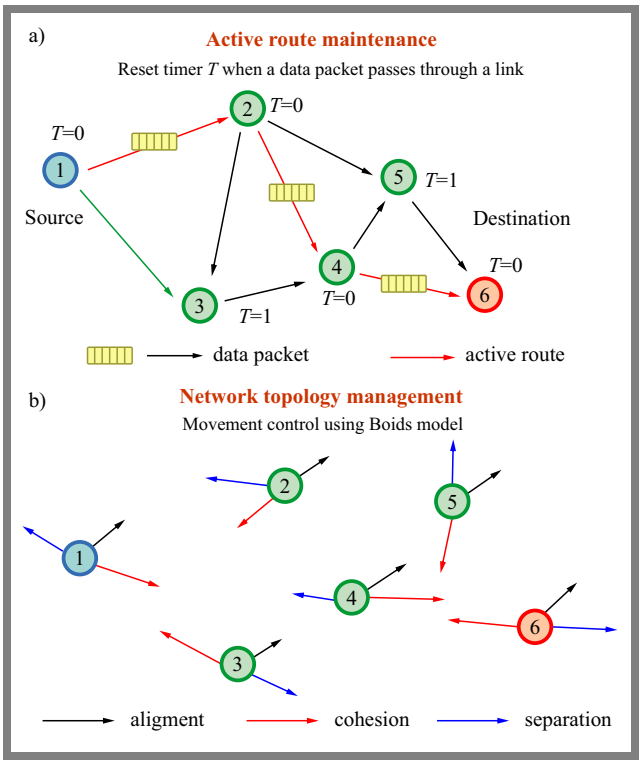

Fig. 5. Illustration of the BR-AODV scheme for FANETs: a) active route maintenance and b) network topology management.

- If silence period T exceeds threshold T_{max}, the UAV is marked as inactive.
- If $T \leqslant T_{max}$, the UAV remains active.
- The timer value is reset to 0 whenever the UAV under consideration sends or receives a message. Otherwise, the value is incremented by 1.

The status of UAVs is checked regularly. Routes through active UAVs are selected for data forwarding. During data transmission for active links, the movement of UAVs is controlled using the Reynolds Boids. The Boids of Reynolds rules are based on distance, cohesion (tendency to remain closer to neighbors), and alignment (velocity and direction toward the neighbors) between the communicating UAVs in the neighborhood, as described in Fig. 5b.

The key objective of the BR-AODV scheme is to maintain connectivity for a route during data transfer [69]. The AODV protocol is augmented with a control module based on Reynolds Boids to track UAVs' movement and counter issues observed in AODV. UAVs can determine their participation in multiple active paths. If so, they must decide their future movements by avoiding any route disconnection.

The scheme [67] involves the following sequence of steps:

- In a swarm, each UAV initializes its parameters such as its unique identifier (MAC or IP address), position P_i, and velocity V_i.
- The UAVs broadcast a hello message, signal strength, V_i, P_i.

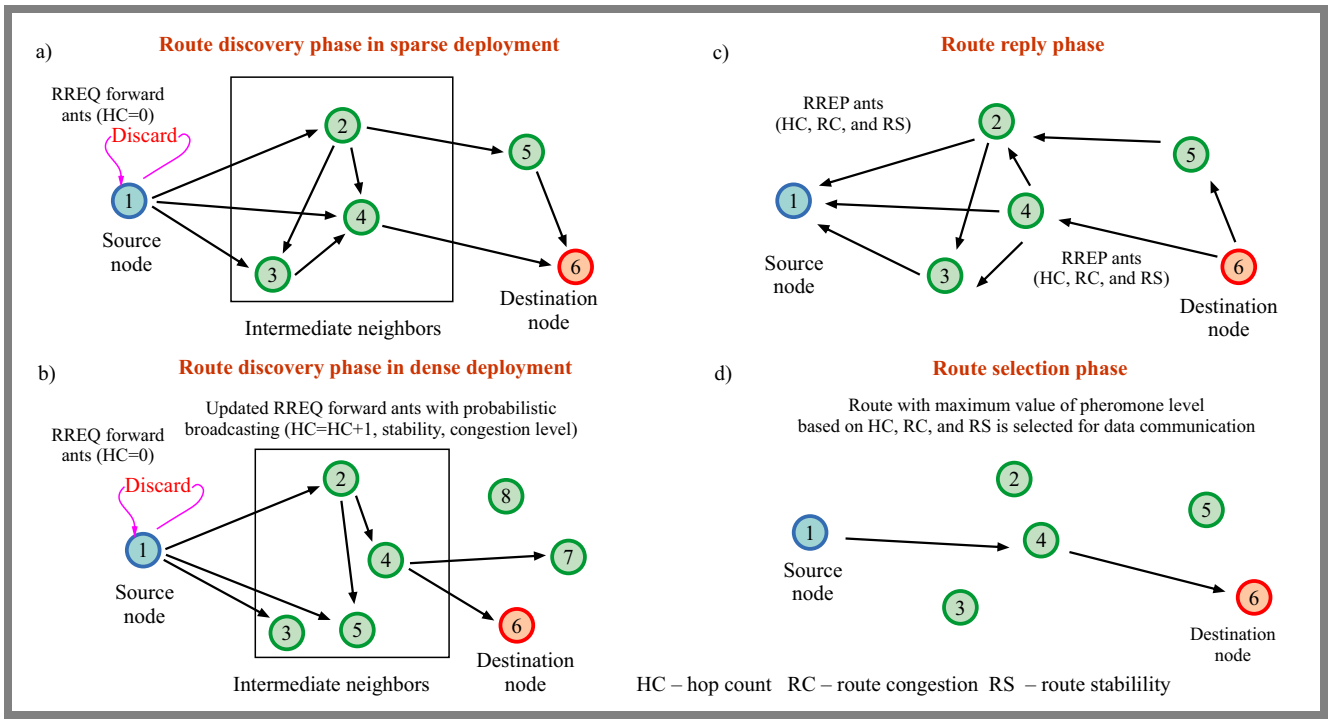

Fig. 6. APAR routing process for FANETs demonstrating the following: a) route discovery phase in sparse deployment, b) route discovery phase in dense deployment, and c-d) route reply and selection phase.

- The UAV identifies nearby UAVs based on signal strength and distance. If a UAV is within radio range, it is added to the list of neighbors.
- The source UAV transmits an RREQ packet to find a path to the destination. If intermediate UAVs do not have a recorded path, they forward the message.
- Once the destination is reached or an intermediate UAV with a known route to the destination is found, it responds with an RREP.
- Each UAV keeps track of whether it is active or inactive based on a timer, denoted by T. If a UAV has been transmitting or receiving packets within time $T \leqslant T_{max}$, it is marked active. If the UAV has been idle for more than T_{max}, it is marked inactive.
- Active UAVs behave like Boids and adjust their movements, applying alignment, cohesion, and separation.
- The route is checked for being active. If all nodes in the path are active, the route is marked as active. If any node in the path becomes inactive, the path is considered inactive. The process is started over to identify an alternative route.
- The UAVs periodically exchange hello messages and update neighbor tables to maintain active routes.
- The UAVs communicate multihop, forwarding data from one UAV to another until they reach the destination or gateway.

BR-AODV is highly adaptable, as the Boids model efficiently manages UAV movements and network routing decisions. The scheme also achieves high route stability, leading to reliable connections. Nevertheless, support for multipath routing is still missing, resulting in low fault tolerance. Owing to the poor scalability of the distance-based approach, the scheme suffers from low network coverage in the case of dispersed networks. BR-AODV has a significant routing overhead due to the combination of the Boids of Reynolds scheme with the traditional AODV routing protocol.

However, the scheme is energy efficient because of route optimization and low redundancy, as shown by the Boids model. BR-AODV also incorporates load-balancing mechanisms by dispersing traffic flow among the routes equally. Unfortunately, BR-AODV does not provide any protection against common security threats.

The scheme suffers from low network coverage, particularly in the case of dispersed networks. Tailored for on-demand routing in dynamically positioned UAVs, BR-AODV integrates RSSI-based coverage mapping during route discovery. This method mimics bird flocks, in which UAVs prevent coverage gaps by realigning to remain within range, thereby preserving group cohesion, separation, and alignment through connectivity while transmitting data.

BR-AODV ensures the coverage of the operational area, since the UAVs maintain their connections in order to form an integrated network. A fitness function evaluates coverage using neighbor counts and path loss models. The computational burden includes $O(m \cdot p)$ (m – node count, p – generations) for neighbor checks during route requests, resulting in latency in low-density networks when updating coverage maps. The computational complexity arises from the need to simulate flocking behavior and dynamically adapt routes, ensuring robust coverage in FANETs.

The advantages of BR-AODV include the following:

- The Boids of Reynolds scheme provides active path connectivity and efficient route maintenance, increasing throughput and reducing packet loss.
- This scheme shows low end-to-end delay, especially for high network traffic loads.

The known limitations of BR-AODV are as follows:

- High routing overhead due to reactive route discovery.
- Poor scalability due to latency spikes in dense networks.
- Poor link stability due to high UAV movements.
- Does not support time-sensitive communication.

4.5. Ant Colony Optimization-based Polymorphism Aware Routing (APAR)

The highly dynamic network topology of FANETs makes standard routing strategies unsuitable for meeting their unique requirements and complex application scenarios. Furthermore, using the hop count as the only route metric makes it impossible to guarantee a consistent PDR in real-time battlefield scenarios. To overcome these challenges, the authors of [70] proposed APAR, which integrates the ant colony optimization (ACO) scheme with the dynamic source routing (DSR) protocol [71]. The algorithm comprises three phases.

Route exploration phase. The authors of [70] proposed two route exploration techniques, with both of them based on network node density, to minimize the broadcast storm during the route creation phase. In sparse deployments, as illustrated in Fig. 6a, a source node without route information in its cache forwards a route request (RREQ) to all nearby stable nodes. Stable nodes are defined as those with high link quality and low mobility, which ensures reliable communication.

Upon receiving the RREQ, intermediate nodes update critical information such as HC, congestion, and stability, and then forward the updated RREQ to their stable neighbors. This process continues until the RREQ arrives at its destination or an intermediary node with a known path, which then transmits a route reply (RREP) back to the source. A probabilistic broadcasting strategy is employed in dense deployments, as illustrated in Fig. 6b.

Upon receiving an RREQ, each intermediate node calculates probability p of forwarding the RREQ based on the number of mutual and unique stable neighbors compared to the previous hop, reducing redundant broadcasts and minimizing the broadcast storm.

Route response and selection phase. Upon receiving the RREQ, the destination responds with an RREP source containing information on the HC, route congestion (RC) and route stability (RS) (see Fig. 6c). After receiving the RREP, the source node calculates the pheromone value for each path according to these metrics and chooses the path with the highest pheromone value for data transmission (see Fig. 6d).

Route maintenance phase. For route maintenance, the DSR topology control packets are modified to include extra information about the route, such as the UAVs' speed, antenna gain, transmitting power, address, and load level. The signal intensity of the received control packet is used to assess stability of the connection between two UAVs.

Channel load is determined as the percentage of time the medium is busy and is calculated as:

$$Channel\ load = \frac{busy\ time}{busy\ time + idle\ time} \times 100\% \,. \quad (6)$$

Buffer utilization is determined by the ratio of the current queue length to the maximum queue length in the media access control (MAC) layer. The level of congestion is evaluated based on both channel load and average buffer occupancy. The ACO algorithm manages routes – those with declining pheromone levels or increased congestion are discarded. This dynamic path management ensures efficient data transfer routes across the network.

The step-by-step breakdown of the routing procedure for the scheme from [70] is as follows:

- In a UAV swarm, for data communication, a UAV checks its route cache for any existing routes that can be used for transmission.
- If such routes exist, the optimal path is selected for information transmission. If no valid route exists, the route discovery process is continued.
- The source node broadcasts RREQ packets to all stable neighboring nodes in sparse formations.
  - Intermediate nodes receiving RREQ check routing loops and update the control packet with HC, stability, and congestion level.
  - If an intermediate node possesses a route to the destination, it initiates the RREP process.
  - If not, it forwards the RREQ to its stable neighbors.
  - Upon receiving RREQ, the destination node updates the route information and returns the RREP to the source.
- Each intermediate node calculates broadcast probability p for dense deployment based on the number of mutual and unique stable neighbors and forwards RREQ accordingly. Upon receiving RREQ, the destination updates the route information and returns the RREP to the source.
- The destination generates an RREP containing HC, RC, and RS and returns it to the source along the same route.
- Intermediate nodes receiving an RREP forward it to the source. If an intermediate node has a path upon receiving the RREQ, it converts it to an RREP and sends it to the source.
- After receiving the RREP, the source node extracts HC, RC, and RS and calculates the pheromone level for each route.
- The route with the highest pheromone level is selected for data transmission.
- The protocol maintains multiple routes with varying levels of pheromones. If the primary route fails due to inactive nodes, backup routes are utilized without restarting the route discovery process.

The authors of [70] compared the performance of APAR with DSR, hybrid ant colony optimization routing (HOPNET) [72],

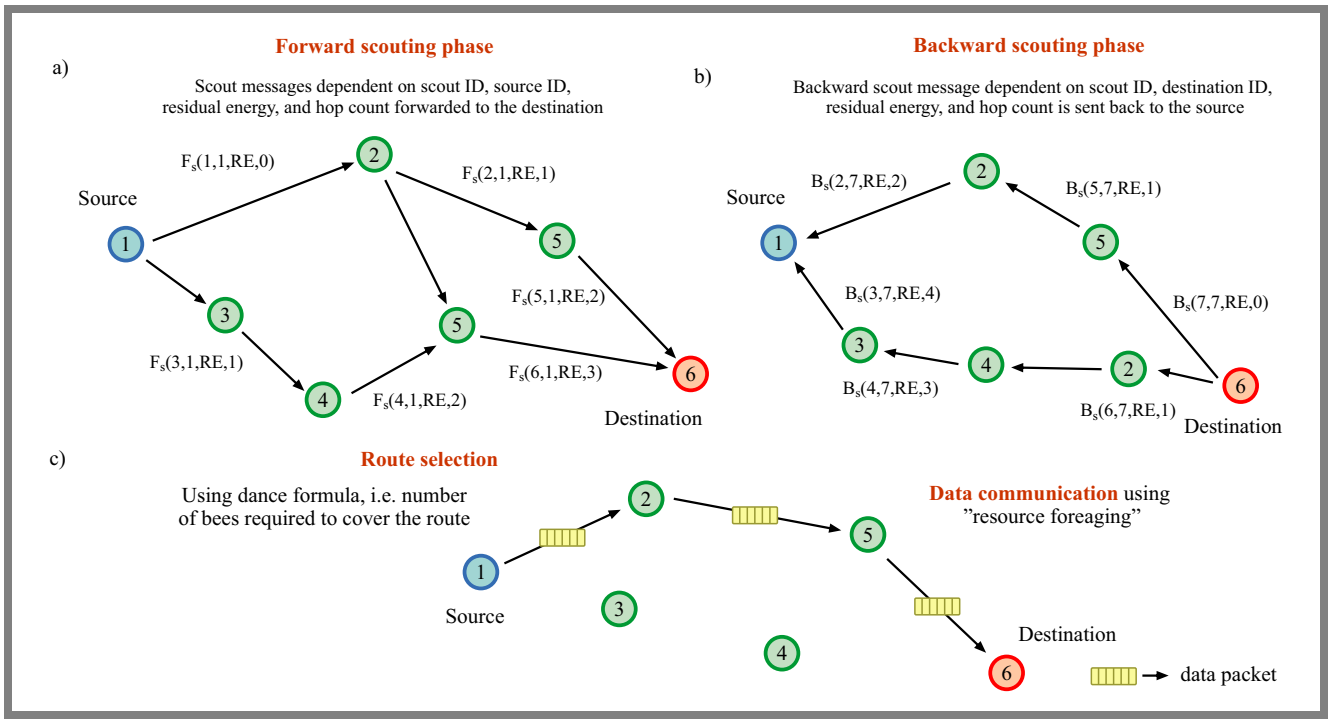

Fig. 7. Illustration of the BeeAdhoc scheme for FANETs: a) forward scouting phase, b) backward scouting phase, and c) route selection phase.

and Ant-DSR [73] algorithms. APAR exhibits high PDR, low average end-to-end delay, and low routing overhead compared to these schemes. Due to the quick topology adjustments of ACOs, APAR is highly adaptable. It also provides a high degree of link stability and scalability by using pheromone values to reinforce successful routes.

However, it suffers from high routing overhead and moderate energy efficiency due to the computational complexity and extra communication overhead required for pheromone updates. APAR effectively distributes traffic over the network, resulting in improved efficiency and reliability. Nevertheless, the protocol does not take into account privacy and security.

Solving polymorphic formations with ACO, APAR utilizes pheromone trails as a route stability method, where ants discover high-coverage routes. With the inclusion of DF and GSO, HSCS improves flexibility, allowing the network to efficiently adapt to dynamic node locations and topologies.

A polymorphic ACO algorithm computes the probability of coverage through Bayesian inference [74] on an ant exploration data. Computational complexity $O(m \cdot n)$ for (m – paths and n – ants) arises from the iterative optimization required to maintain robust coverage.

The advantages of the APAR protocol can be summarized as follows.

- The ACO algorithm provides a stable route by sensing pheromone levels in routes.
- It is applicable in both sparse and dense application scenarios, as different routing schemes were employed accordingly.
- Network performance degradation can be reduced by utilizing the pheromone volatilization property in routes.

The limitations of APAR include:

- High computational overhead of multipath exploration.
- Slow convergence in dense networks.
- Poor link stability due to high UAV movements.
- Lack of support for low-latency applications.

4.6. BeeAdhoc Routing Protocol

The BeeAdhoc routing protocol, introduced in [75], is a bio-inspired mechanism designed to manage the movement in FANETs. Grounded in swarm intelligence (SI), BeeAdhoc draws inspiration from honey bees' collective movement and foraging behavior to enable efficient data routing.

The protocol relies on frequent message exchanges utilizing two distinct message types: scouts and foragers. Scouts operate on demand to identify new paths to the destination. In the forward scouting phase, depicted in Fig. 7a, the source node broadcasts a forward scout message F_s containing the scout identification number (ID), sender ID, residual energy (RE) and hop count (HC).

Intermediate nodes update this information and relay it to neighboring nodes toward the destination. When F_s are received, the destination node responds with a backward scout message B_s comprising the scout ID, destination ID, HC, and RE, which retraces the route to the source, as depicted in Fig. 7b. The final route is selected based on a dance-inspired formula modeled after the bee waggle dance [76], [77], which evaluates the number of UAVs required to cover the path (see Fig. 7c).

Data transmission employs foragers, mirroring the resource-transportation behavior of bees. The routing procedure, as detailed in [77], unfolds as follows:

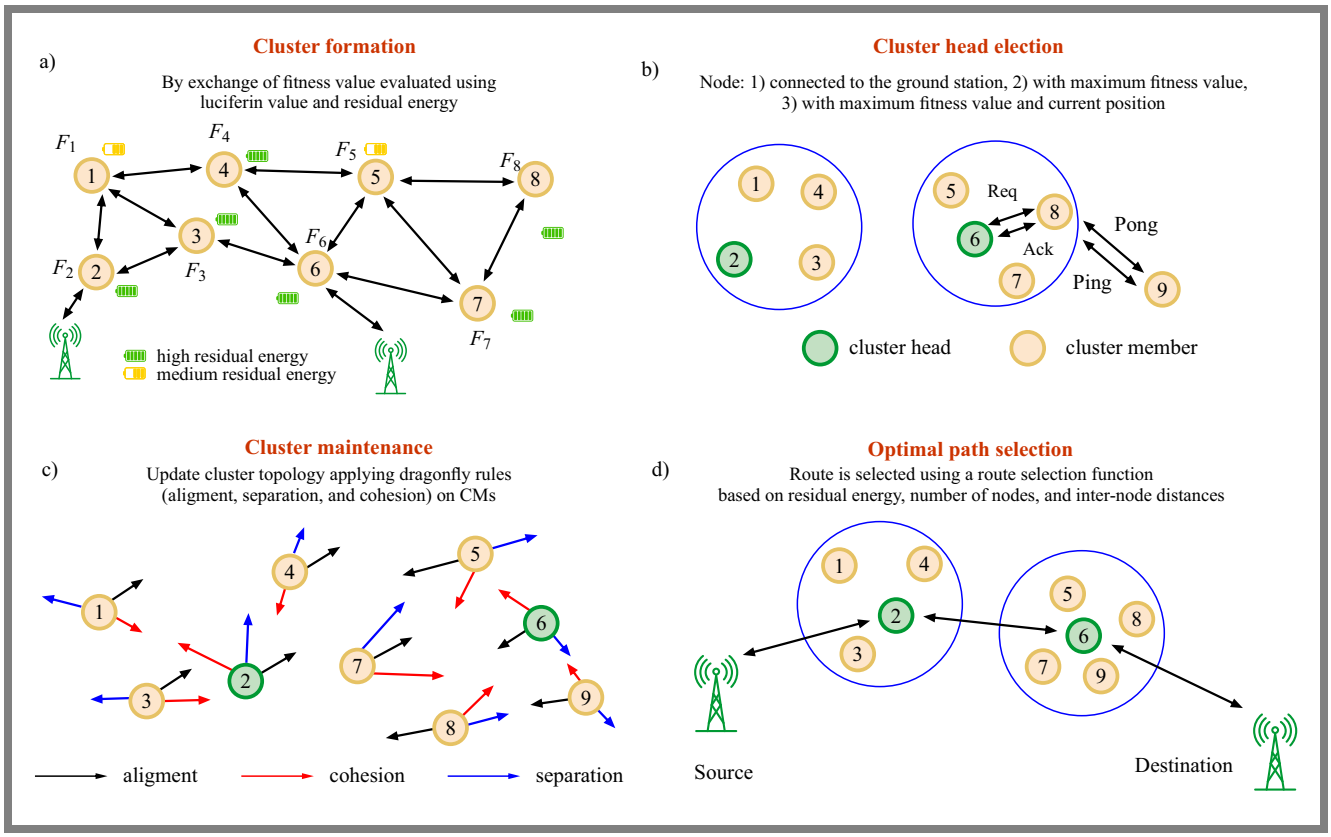

Fig. 8. HSCS scheme for FANETs: a) cluster formation, b) cluster head election, c) cluster maintenance, and d) optimal path selection.

- The sender node in a swarm transmits an F_s (scout ID, Sender ID, RE, HC) toward the destination.
- Intermediate nodes receive the F_s, update its contents, and forward it to neighboring nodes closer to the destination.
- Upon receiving the F_s, the destination node generates a backward scout message B_s (scout ID, destination ID, HC, RE), which backtracks the route to the sender.
- During data transmission, a packet from the transport layer arrives at the packing floor.
- A packer agent is instantiated to store the data packet temporarily. The packer searches the "dance floor" for an available forager to transport the packet to the destination. If a suitable forager is found, the packet is transferred and the packer terminates. If no forager is available, the packer waits briefly for a returning forager; if none arrives, a scout is deployed to discover a new route.
- Foragers acquire the entire route from a scout or another forager and transmit data relying on the multihop process.
- For TCP transmissions, the forager at the destination waits to be piggybacked.
- For UDP, the forager may become stranded at the destination and unable to return. To mitigate this, the swarm embeds multiple foragers, some in the header and others in the payload, to replenish resources at the destination.
- Upon returning to the source, the forager performs a "dance" to recruit additional foragers, reflecting the quality of the traversed path.

This scheme is highly adaptable, and the foraging behavior improves link stability. Furthermore, BeeAdhoc incurs a high routing overhead due to its high computation and communication requirements. Privacy and security are not prioritized, making the network vulnerable.

Network coverage is reduced due to the emphasis on local clustering. Optimized for energy efficiency in moderate mobility, BeeAdHoc uses bee scouts to discover routes to uncovered areas. Scout bees divide the terrain, and foragers evaluate coverage through path redundancy checks.

A flower pollination model maps coverage using scout bee trajectories. This incurs $O(s \cdot d)$ messaging overhead (s – scouts, d – destinations), resulting in severe accuracy degradation in urban canyons where sector misalignment causes coverage overestimation. Computational load increases exponentially with terrain difficulty and dense networks.

The advantages of the BeeAdhoc protocol include:

- It is energy efficient due to imitation of honey bees' foraging behavior and low transmission of control packets.
- Honey bees' foraging behavior results in efficient and consistent cluster formation and a low re-clustering rate.
- Due to its simple implementation, this technique is appropriate for many network situations.

The limitations of BeeAdhoc are as follows:

- High computational overhead from multipath exploration.
- Poor scalability due to proliferation of control packets.

- Poor link quality in large networks, as it suffers from a longer end-to-end delay.
- No priority scheduling for real-time data.

4.7. Hybrid Self-organized Clustering Scheme (HSCS)

To counteract the dynamic topology and limited resources of FANETs, the authors of [78] proposed the HSCS routing protocol. HCHS is a hybrid of the dragonfly scheme (DF) [79] and GSO algorithms, consisting of three essential phases: cluster formation, cluster management, and cluster maintenance. Such an approach facilitates effective communication among drones within the network. The GSO method is used for clustering and CH election process, while the DF scheme monitors CMs' mobility.

During clustering, as shown in Fig. 8a, each node evaluates its fitness based on luciferin values and residual energy levels. A cluster is formed by broadcasting the fitness value in the neighborhood. The fitness values and connectivity of the GCS are evaluated for the purpose of CH selection. If a single node within a cluster maintains a connection to the GCS, that node is selected as the CH, while the remaining nodes become CMs.

On the contrary, if multiple nodes are connected to the GCS or no nodes are connected, the CH is elected based on the maximum fitness value, as depicted in Fig. 8b. A node must broadcast a PING control message to the CH if it wants to join a cluster. A CM must forward an REQ control message to its CH if it receives the control message. Upon receiving the REQ message, the CH must send an acknowledgment (ACK) to the CM, and the CM will respond with a PONG message to the new node.

Furthermore, as explained in Fig. 8c, during the cluster management phase, the protocol facilitates effective coordination between CM and the CH, ensuring adherence to swarm behavior within the cluster. This coordination is achieved through a next-hop selection function that fosters overall network stability. The process is governed by DF rules, prioritizing alignment, cohesion, and separation.

These principles direct the movement of CMs, ensuring that they remain connected to the CH while maintaining optimal distances between themselves. This approach prevents collisions and fosters effective communication within the cluster. During the cluster maintenance phase, the status of CMs is periodically evaluated to ensure network stability. Nodes with energy levels that are lower than a threshold value are identified and removed from the cluster. Additionally, an alternative route is determined.

Finally, the optimal route is selected using a route selection function (RSF) based on key parameters such as node count, residual energy, and inter-node distances. The optimal value of RSF ensures low energy consumption. The route selection process depicted in Fig. 8d frees the protocol from multiple routes. The process outlined in scheme [80] is as follows:

- Each UAV calculates its fitness F_v (connectivity, residual energy, and luciferin value).
- UAVs exchange hello messages to discover their neighbors. If no UAVs are connected to the BS, the node with $\max(F_v) = CH$. If multiple UAVs are connected to the BS, the node with $\max(F_v) = CH$.
- The CH initializes the formation of the cluster, and neighboring nodes join as CMs.
- The CH applies DA rules (alignment, cohesion, and separation) and updates the cluster topology table based on the positions of its members.
- For data transfer, UAVs employ the RSF algorithm to determine the best route to the target. RSF analyzes aspects such as connection quality, HC, and RE.
- The data are communicated using the optimal route.

HSCS integrates bio-inspired techniques to optimize network organization and routing. CHs monitor member dropout rates to detect holes. By combining DF and GSO, HSCS improves adaptability, enabling the network to adjust efficiently to dynamic node positions and topology changes.

The protocol achieves high link stability due to its hybrid solution, which optimizes both clustering and routing paths in dynamic topologies to improve connections. Cognitive IoT-optimized HSCS supports a PSO coverage optimizer that attempts to cover the minimum number of uncovered grids using fitness functions. Minimizing communication, but with $O(n \cdot i)$ (n – nodes, i – iterations) complexity, it drains batteries on long-duration missions, resulting in reconstruction errors that are greater than those in high-mobility cases.

However, HSCS supports the establishment of multiple routes by leveraging its clustering approach. Multiple paths can be maintained within and between clusters, enhancing redundancy and fault tolerance. The network coverage ratio is high, as the hybrid clustering approach effectively manages more extensive and dispersed network areas.

HSCS suffers from high routing overhead due to high computational and communication complexity. Owing to its high computational complexity, energy efficiency performance is moderate. HSCS effectively manages network traffic but does not provide privacy and security, leaving the network open to security threats.

The advantages of the HSCS routing technique may be summarized in the following manner:

- Due to the fewer topology control messages, this scheme provides low routing overhead and high energy efficiency.
- The proposed scheme shows efficient cluster management due to the use of the DA algorithm for tracking CMs.
- Due to the nature of self-optimization, HSCS improves link stability.

HSCS suffers from the following limitations:

- High routing overhead due to complex computations for hybrid clustering and cognitive decision-making.
- Cluster formation latency with high mobility.
- Not optimized for multimedia traffic.

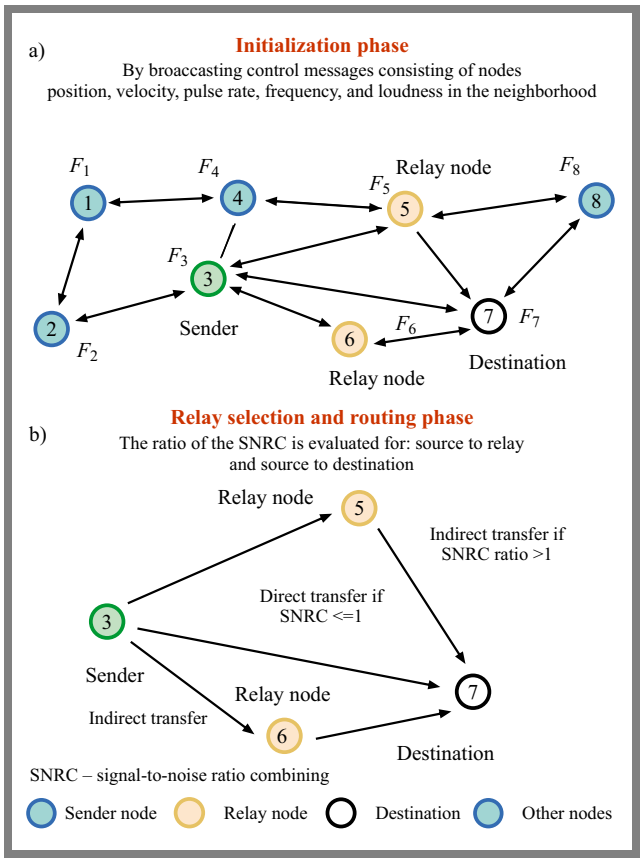

Fig. 9. BAT-COOP scheme for FANETs: a) initialization phase and b) relay selection and routing phase.

4.8. Bat Algorithm Using Cooperation Technique (BAT-COOP)

The authors of [81] proposed the BAT-COOP technique for FANETs to minimize transmission losses by decreasing end-to-end delay. To achieve this objective, bat echolocation [81] characteristics have been mimicked to induce cooperation features in the network. The algorithm comprises initialization, cooperation, as well as relay selection and routing. During the initialization phase, the nodes share their information by exchanging control messages consisting of the position, distance d, velocity v, pulse rate PR, frequency f, and loudness l in the neighborhood. Each node in the network evaluates its local optimal position, which is beneficial for intelligent selection of the next hop by the sender, as depicted in Fig. 9a. In the cooperation phase, the sender simultaneously forwards data to the destination and relay nodes.

Relay nodes forward the received data to the destination. Upon receipt of the data packets, the sink node merges the received signals and the power of the sender directly with the relay node [82]. Finally, during the relay selection and routing phase, a node with a higher cost value based on distances between nodes, signal-to-noise ratio, cost factor, and instantaneous link condition is selected for the relay. This is shown in Fig. 9b.

Following the selection of the relay node, the information is forwarded thereto. Data forwarding depends on signal-to-noise ratio combining (SNRC), which is evaluated for routes from the sender to the relay and from the sender to the destination. If the $ratio <= 1$, the data is transferred directly to the destination; otherwise, for a $ratio > 1$, the data is transferred to the relay node. During the indirect transfer, the relay node waits for a predefined hold time, during which redundant data packets are rejected at the destination, which is received from the sender directly or through the relay. When two relay nodes are available on the destination path, the relay node will not trigger cooperation. The needless forwarding of data packets has been successfully countered using the bat cooperation technique. The procedure described in scheme [81] includes the following steps:

- The system is initialized, with each node being aware of its position, velocity, frequency, pulse emission rate, and the best local position.
- The source node transmits its current position.
- Solutions for each node are generated by adjusting the pulse emission rate, while velocities and positions are updated based on the current best solutions discovered so far.
- The source generates a random number R and compares it with pulse emission rate r_i. If $R > r_i$, a solution is selected from the current best solutions; otherwise, a local solution is generated using random flight. Next, a fitness value is evaluated for each node; if it is worse than its local best, it is replaced with a new solution generated by random flight. If the new solution is better than the current one, it is accepted and the local best update is provided.
- After selecting the local best, a cost function $C_f\ (d, f, l)$ is evaluated, and the feasibility of a direct transfer is checked. If the direct transfer path is feasible and satisfies the cost function, the data is transferred directly from the source to the destination. Otherwise, the system selects the best relay node for the data transfer.
- If the destination receives multiple signals from the relay nodes, the SNRC scheme combines these signals, selecting the one with the highest SNR.

The BAT-COOP protocol employs a multi-hop routing strategy inspired by the cooperative behavior of bats to enhance network performance. BAT-COOP demonstrates high adaptability through its cooperative framework and efficiently handles dynamic network topology. Selecting optimal paths improves link stability and diminishes the chances of link failures. The protocol supports the establishment of multiple routs, thus improving redundancy and fault tolerance. Its multi-hop strategy and cooperative beam-focusing technique ensure extensive network coverage, making it suitable for larger operational areas.

However, BAT-COOP incurs significant routing overhead due to the computational demands of cooperative diversity techniques and frequent routing updates. Continuous message exchanges contribute to high communication overhead, while the storage and management of cooperative data structures result in elevated space complexity.

Energy efficiency remains moderate, as the cooperative approach increases energy consumption during multi-hop transmissions. The protocol integrates effective load-balancing

mechanisms to distribute traffic evenly across the network. Despite its advantages, BAT-COOP prioritizes cooperative communication and network localization over privacy and security, potentially leaving the network vulnerable to security issues.

With cooperative diversity as its target in mobile FANETs, BAT-COOP utilizes cooperative relay probing, in which UAVs mimic signal pulses to identify coverage holes and node sparsity based on echo delays.

BAT-COOP employs a pulse-echo simulator model of coverage based on delay-sensitive sonar equations. Urban environments require $O(n \log n)$ complexity (n – number of bats), which increases in high-density areas. Unmodeled weather effects introduce accuracy errors in rain and fog conditions.

The advantages of the protocol are as follows.

- The approach suits dynamic networks, since the node count is regularly updated and shared.
- The scheme overcomes the issue of unnecessary forwarding of packets and reduces transmission losses.
- The scheme achieves better load balancing due to the use of forwarding nodes.
- The scheme suits sparse and dense networks in real-time application scenarios.

The limitations can be summarized as below:

- High routing overhead due to the computational complexity of cooperative routing.
- Limited scalability due to distributed cooperation.
- Not specifically designed to handle real-time packet delivery constraints.

4.9. Gray Wolf Algorithm Using Cooperative Diversity Technique (GW-COOP)

For efficient routing in FANETs, the authors of [83] proposed the GW-COOP approach that is based on the gray wolf optimizer (GWO) and uses the collaboration technique and the leadership hierarchy of gray wolves [84] to find the optimum path to the destination. The proposed scheme comprises two phases, namely initialization and routing.

As depicted in Fig. 10a, in the initialization phase, every node iteratively evaluates its fitness value based on energy level and distance from the destination. The nodes are ranked according to their position and fitness value. Next, the first three best options are identified during each iteration. The first best UAV is considered alpha, which is the sender node. Beta and delta nodes are the second and third best UAVs, respectively. The remaining nodes are regarded omega.

The sender can directly forward critical data to the sink during the routing phase. In the event of less critical data or a change in topology, the data can be sent through beta or delta relay nodes. Upon receiving a data packet from relay nodes, the enhanced signal-to-noise ratio combining (ESNRC) approach is used at the destination to consider higher strength signals, as shown in Fig. 10b.

The steps involved in scheme [83] are described below:

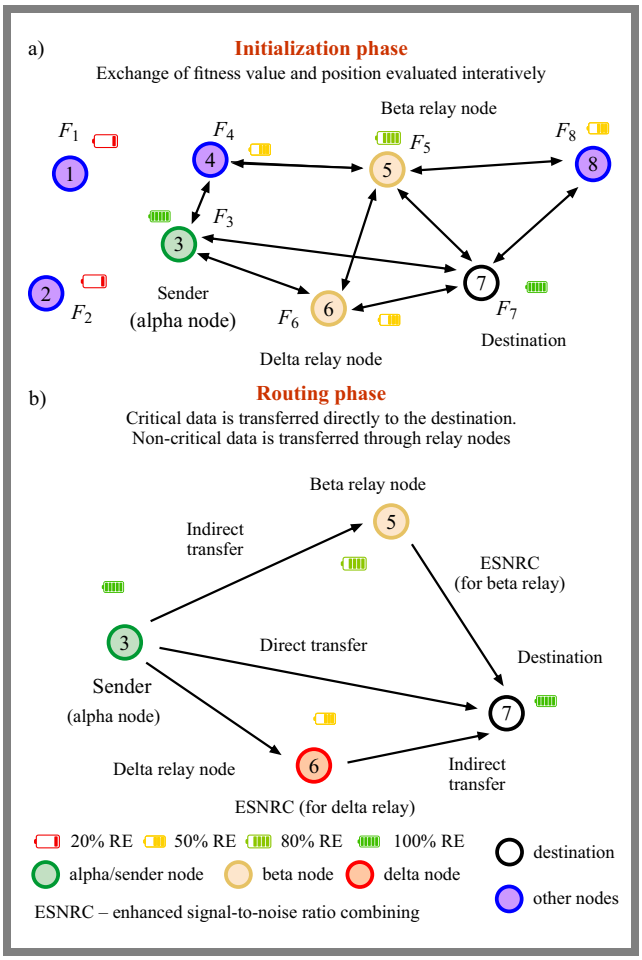

Fig. 10. Illustration of the GW-COOP scheme for FANETs: a) initialization phase and b) routing phase.

- UAVs are initialized randomly and each UAV evaluates its fitness value F_v iteratively, based on energy level.
- Based on F_v, the nodes are categorized as: α – first best solution, β – second best solution, δ – third best solution, and ω – the remaining solutions.
- The system iteratively searches for the best route based on the cost function.
- If critical data are detected, α is used as the direct path for communication. Otherwise, the cooperative phase is activated.
- Residual energy level is estimated for the source and potential relay nodes R_1 and R_2.
- A direct path is chosen if the source node has a higher residual energy than R_1 or R_2.
- Otherwise, a relay path is selected using the amplify-and-forward (AF) relay strategy at R_1 and R_2.
- Enhanced SNR combining is applied at the destination to improve reliability.

The performance of the proposed scheme was compared to BAT-COOP to evaluate its efficiency. Due to its incorporation of social hierarchy and cooperative diversity, the GW-COOP scheme outperforms BAT-COOP by showing lower transmission losses, improved adaptability, and more

reliable communication. GW-COOP is highly adaptable, efficiently adjusting to changes in node mobility through its cooperative diversity approach.

The protocol supports multiple routes, enhancing redundancy and fault tolerance within the network. Network coverage is also high due to its multi-hop strategy, which facilitates extensive reach across larger areas. However, GW-COOP incurs relatively low routing overhead due to efficient cooperation mechanisms that minimize redundant data transmissions and optimize path selection, making the process of scaling up for large networks easy. The cooperative approach may result in higher energy consumption during multi-hop communications. GW-COOP effectively manages network load. However, it does not provide privacy and security.

Designed to target link-aware routing in high mobility, GW-COOP utilizes a gray wolf hierarchy for loss prediction during routing. "Alpha wolves" direct pack movement using fitness functions that evaluate coverage entropy from neighbor reports. Link quality thresholds trigger rerouting.

A social hierarchy-based search is used to minimize path loss. The $O(p \cdot n)$ complexity (p – pack size, n – generations) causes longer decision cycles in moderate-density networks.

The advantages of GW-COOP can be summarized in the following manner:

- The scheme employs fewer parameters and is easy to implement.
- It considerably minimizes transmission losses, resulting in reduced energy consumption, packet loss ratio, and link delay.
- The protocol offers better local exploration, as it employs a hierarchical pattern.
- It handles FANETs effectively due to the use of ESNRC.
- It offers reliable network communication, even in sparse deployment scenarios.

The limitations of GW-COOP include:

- Moderate routing overhead from gateway selection.
- Gateway congestion in large networks.
- Link instability during gateway failure.
- Time-window or latency-sensitive communication is not prioritized, limiting real-time performance.

4.10. Physarum-inspired Clustering Algorithm (PICA)

The authors of [85] proposed PICA – an innovative clustering algorithm designed explicitly for FANETs. The proposed scheme is based on the foraging behavior of physarum polycephalum (PP) silme mold [86]. PICA is known for its remarkable ability to form efficient networks and solve complex optimization problems. It utilizes a distributed multi-hop clustering approach that significantly improves forming and maintaining clusters within FANETs.

PICA is a two-stage process that includes cluster formation and maintenance. All nodes start in the initial IN state during cluster formation. The clustering process is depicted in Fig. 11a. Nodes establish their neighbor sets by exchanging hello

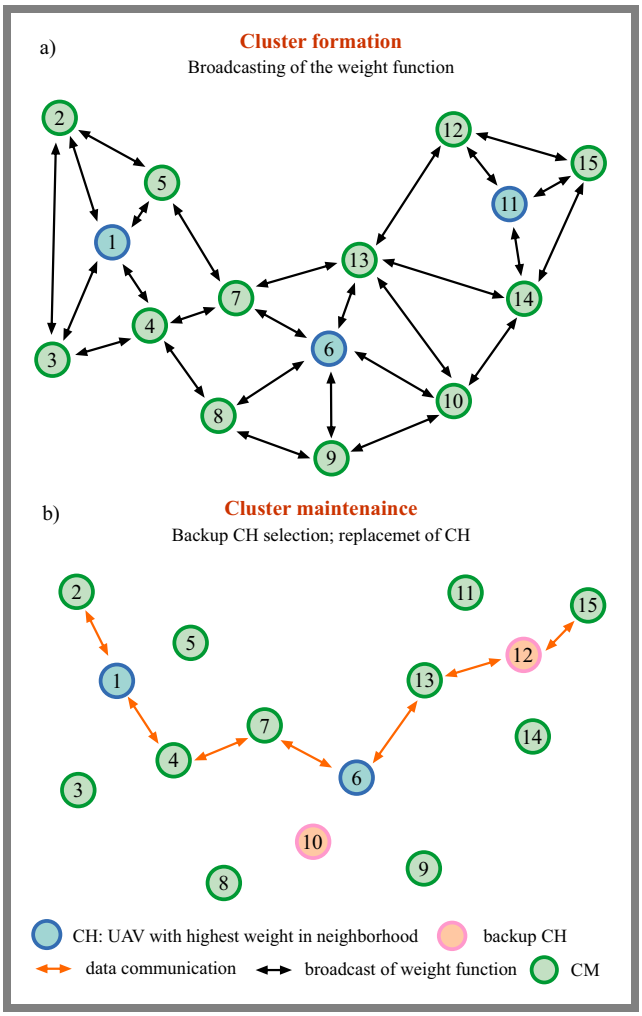

Fig. 11. Purpose of operation of the PICA scheme for FANETs demonstrating: a) cluster formation and b) cluster maintenance.

control packets and determining their weights using a W_f function that depends on the likelihood of the current node becoming a CH. These weights are subsequently shared with neighboring nodes in a second control message exchange round. If a node's neighbor set is populated and its probability of becoming a CH exceeds that of neighboring nodes, it is elected as the CH. The CH transmits a cluster head declaration (CH_DEC) to the surrounding nodes.

Nodes issue a join request (JOIN_REQ) to join the cluster after receiving a CH_DEC message, designating the node with the best virtual communication quality as their CH. It switches to CM after receiving a join response (JOIN_RESP). This multi-hop clustering structure optimizes cluster coverage by minimizing the frequency of switching between CMs and clusters.

During the cluster maintenance phase, procedures for detecting damage and merging clusters work to maintain the cluster's stability structure, while backup CHs are updated regularly, as illustrated in Fig. 11b. The CH periodically broadcasts the list of CMs. When this broadcast is received, any member with more than 60% of the CMs in its neighbor set determines its priority for possible election as a backup CH.

This priority is determined by factors such as the average virtual communication load with other nodes in the cluster and the member's remaining energy. The CH selects the member with the highest priority as the backup CH, and any updates regarding changes in the backup CH status are immediately broadcast to the cluster. To avoid selecting edge nodes, only nodes with strong connectivity to other cluster nodes are considered for backup CH selection. The step-by-step breakdown of the routing procedure for the scheme is as follows [85]:

- UAVs initiate the clustering process in a distributed manner. The source node shares hello control messages with its neighboring nodes.
- Each node assesses its weight function W_f, which relies on BeCH, and its likelihood of becoming a CH (BeCH).
- W_f is shared with the neighboring nodes during the second exchange of hello messages.
- If BeCH is the highest, the node changes its status to CH and invites nearby nodes to join the cluster; otherwise, it tries to connect to a nearby CH. If the node can connect to a CH, its state changes to CM. JOIN_REQ is forwarded and the cluster is joined if JOIN_RESP is received from CH. If the node cannot connect to any CH, it uses a multi-hop connection through another CM.
- A backup cluster head (BCH) is selected to take over if the current CH fails.
- If the clusters become too small, they merge with other clusters.

The algorithm considers several critical factors during the cluster formation process, including link stability, residual energy, and communication quality. PICA offers high adaptability and link stability levels by dynamically adjusting to network topology changes and mimicking PP's foraging behavior. This results in extensive network coverage and effective communication within clusters. However, PICA faces high computational and communication overhead challenges due to the complex computations and frequent updates required for accurate clustering.

In addition, storing pheromone levels and routing information contributes to increased memory overhead. While the algorithm aims for energy-efficient clustering and offers effective load balancing by distributing network load evenly, it does not inherently support multiple routing paths. Additionally, it does not prioritize privacy and security, leaving the network vulnerable to attacks.

Optimized for scalable clustering in highly mobile FANETs, PICA implements physarum-inspired multi-hop coverage trees. CHs utilize Levy flight patterns [87] to adjust their flights while exploring previously unexplored areas. CMs relay data across clusters to fill coverage gaps. The protocol prioritizes zones with low node density by triggering the CH redeployment when coverage holes are detected via RSSI decay.

A Levy flight coverage analyzer estimates coverage gaps, and CHs merge coverage maps via Levy flight, reducing

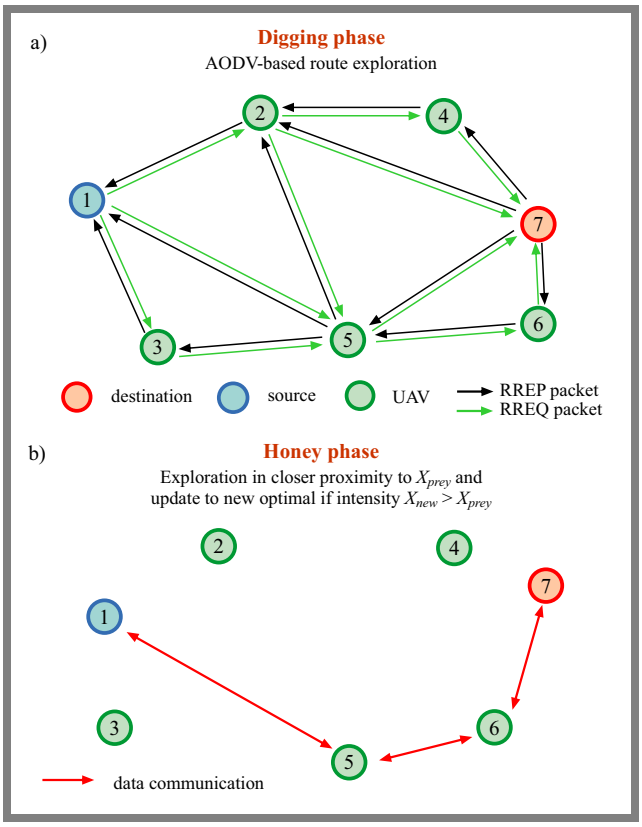

Fig. 12. HB-AODV scheme for FANETs: a) digging phase and b) honey phase.

communication overhead but requiring $O(m \cdot n)$ computations (m – clusters, n – waypoints).

The advantages of the PICA method include the following:

- PICA ensures the shortest paths between nodes, reducing communication latency and improving network performance.
- The algorithm can dynamically reconfigure itself in response to environmental changes, ensuring robust and flexible network connectivity.
- By optimizing the routing paths, PICA helps reduce energy consumption.

The limitations of PICA are presented below:

- High routing overhead due to computational complexity from multilevel clustering.
- Poor link stability in the case of highly dynamic networks.
- There is no dedicated mechanism for data delivery, limiting emergency communication support.

4.11. Hybrid Optimization of the 2-opt Heuristic and Honey Badger Algorithm (HB-AODV)

To counteract the dynamic topology of FANETs, the authors of [88] proposed HB-AODV – a hybrid of the 2-opt heuristic and the honey-badger algorithm (HBA) [89]. The proposed scheme comprises two phases: initial solution construction and route optimization using HBA. During the initialization phase, routes are determined using the 2-opt heuristic, as represented in Fig. 12a. To explore the route to the destination,

the source node initiates a route discovery process using RREQ and RREP control messages, in a way that is similarly to that employed in the traditional AODV scheme.

A source node broadcasts RREQ messages when it needs a path to the destination. Each RREQ packet is modified to include a path field, accumulating the node IDs of all intermediate nodes traversed. The destination node collects multiple RREQ packets from the same source-destination pair. Each unique record path forms an initial solution in the population. Then, all these records are represented as a graph, in which edges denote communication links. Path length is reduced by iteratively swapping edges. During the route optimization phase, as shown in Fig. 12b, a metaheuristic approach is applied to refine the initial solutions obtained from phase 1. The scheme continues exploring new paths while exploiting the available routes.

The goal is to identify the path with the fewest hops. This phase also includes route maintenance, which is used to repair inconsistent paths. If a link is detected, the route repair process is triggered to rapidly adapt to topology changes. This continuous optimization and maintenance ensure reliable communication and robust performance. The steps taken in scheme [88] are described below:

- Randomly initialize the positions of honey badgers (potential solutions) within the search space.
- Assess the fitness of each honey badger position based on HC, delay, and energy consumption.
- Store the honey badger position with the best fitness (i.e. the most promising route) as x_{prey} and its F_v as f_{prey}.
- Find the optimal solution. When stopping criteria such as reaching the maximum number of iterations or achieving the desired fitness are met, the algorithm yields the current best-results solution; otherwise, it continues to update variables and searches for a better solution.
- Calculate each honey badger intensity level I based on its F_v and P.
- Generate a random number r between 0 and 1. If $r < 0.5$, update the honey badger's position using an exploration equation. If $r \geqslant 0.5$, update the honey badger's position using an exploitation equation.
- Evaluate the fitness of the updated solution.
- If the new fitness f_{new} is better than the current best fitness f_{prey}, update it.

The authors of [88] assessed the performance of HB-AODV using a network simulator. They concluded that HB-AODV is superior to DSDV and AntHocNet in PDR and outperforms traditional protocols such as AODV, DSDV, and AntHocNet in terms of QoS metrics, including PDR, average end-to-end delay, and throughput.

The metaheuristic approach makes this scheme adaptable to dynamic topology, ensuring reliable communication. However, the scheme suffers from high routing overhead due to high computational and communication requirements. Route optimization makes the scheme energy efficient, but security and privacy issues are neglected.

Combining 2-opt heuristics and HBA for dynamic networks, HB-AODV employs genetic coverage optimization to prevent coverage holes. UAVs continuously monitor neighbor density via hello packets, and GA-optimized paths avoid low-connectivity zones by penalizing routes with low neighbor counts.

GA-driven Dijkstra's algorithm evaluates coverage using a fitness function combining path loss (Friis free-space model), node degree, and link stability index. The $O(p \cdot g)$ complexity (p – population, g – genes) incurs more computational overhead when optimizing 3D coverage with altitude variations.

The advantages of the scheme are as follows:

- The scheme shows better performance than traditional routing approaches.
- This scheme is highly adaptable and balances exploration and exploitation.

The limitations of HB-AODV are:

- High routing overhead due to the increased complexity of the hybrid mechanism.
- Limited scalability.
- Reactive routing delays persist, affecting responsiveness in dynamic scenarios.
- Not optimized for latency-sensitive applications, limiting real-time performance.

4.12. Adaptive Secure and Efficient Bio-inspired Routing Protocol (Penguin-AIS)

To provide secure routing in FANETs, the authors of [90] proposed Penguin-AIS, a hybrid penguin search optimization algorithm (PeSOA), and artificial immune systems (AIS). PeSOA provides optimal routes using penguin collaborative hunting strategies. The PeSOA routing process is initialized by clustering UAVs that are randomly dispersed in the search space, mimicking the spatial distribution of penguins. UAVs gradually refine their paths, converging towards an optimized solution, as depicted in Fig. 13a. In the next phase, the location is estimated through a fitness function and then adjusted using penguin foraging behavior. Additionally, UAVs can optimize their current location by exploring new regions.

PeSOA prioritizes security and privacy and employs AIS, a solution similar to the human immune system, capable of detecting and countering malicious activities, as shown in Fig. 13b. The step-by-step breakdown of the routing procedure for scheme [90] is as follows:

- The search area is initialized and velocity- and position-related restrictions imposed on the nodes are defined.
- The UAVs are divided into equal-sized groups and a group head (GH) is assigned to each group.
- The beacon coordinates are sent to GH to guide them to the search area. They then further search within their assigned sub-area, updating UAV positions as needed.
- Information is exchanged between the GH and the central command.

Tab. 3. Comparative analysis of bio-inspired routing systems.

Routing protocol	Goal of optimization	Network coverage context	Network coverage	Protocol-specific constraints	Mapped universal constraints	Timing constraints
BICSF	Network coverage enhancement	Assume moderate-density UAV swarms in open areas	Partial	High re-clustering overhead	C1↑, C3↑	Cluster maintenance window
SIL-SIC	Efficient localization and clustering	Focuses on emergency scenarios with dynamic node distributions	High	Localization delay in sparse networks	C1↑, C4↓	Emergency data delivery
BIR-SLB	Multimedia routing	Targets multimedia traffic in mobile emergency networks	High	QoS degradation under mobility	C2↓, C4↓	Multimedia frame delivery
BR-AODV	On-demand routing	Assumes uniform node distribution	High	Slow route recovery	C2↓, C4↓	Route recovery deadline
APAR	Stability-aware routing	Medium-density airspace	High	Slow convergence in large networks	C3↑, C4↓	Convergence time for node heterogeneity
BEEAdhoc	Bee foraging-inspired routing	MANET-adapted	Low	Energy-intensive path discovery	C1↑	Scout path discovery latency
HSCS	Self-organized clustering	Cognitive IoT (static-sensor interaction)	High	Cluster instability in dynamic environments	C2↓, C5↓	Cluster stabilization
BAT-COOP	Enhances cooperative diversity	Suburban environments	High	Weather-sensitive links	C5↓	Relay handshake window
GW-COOP	Link/loss-aware routing	Moderate mobility	High	High computational load	C1↑	Link-failure response time
PICA	Multi-hop clustering	Uniform FANETs	Partial	Multi-hop latency	C4↓	Multi-hop forwarding
HB-AODV	Hybrid optimization	Generic UAV networks	High	Complex parameter tuning	C3↑	Hybrid metric tuning latency
Penguin-AIS	Adaptive secure routing	Security-focused FANETs	High	Security-induced overhead	C1↑, C3↑	Security validation delay
C1 – energy, C2 – mobility, C3 – scalability, C4 – QoS, C5 – link stability, constraint violation (↑) constraint compliance (↓)						

- Each group checks if its assigned sub-area has been searched thoroughly.
- The central command checks that the entire search area has been covered. Otherwise, the search continues.
- The UAV nearest to the destination is updated and stored as the best global solution g_{best}.
- UAVs from completed groups can be redistributed to unfinished groups.
- Low-power UAVs are identified and sent for recharge.

The authors of [90] evaluated PeSOA's performance through simulations and concluded that the scheme exhibits adaptability, high PDR, and reduced average end-to-end delay due to optimized route discovery. The scheme also focuses on load balancing and energy efficiency.

However, the overall routing overhead is moderate due to efficient route discovery and low computation overhead. The scheme continuously monitors packets due to the integrated AIS, enabling the easy detection of any abnormal activity.

Additionally, the scheme provides high route stability and optimized single-route communication. UAVs emulate penguin hunting behavior, cooperatively "diving" into coverage gaps detected via encrypted neighbor discovery. The protocol prioritizes the coverage of the secure zone by isolating untrusted nodes and dynamically positioning UAVs to fill the holes identified through packet loss thresholds.

Providing secure routing for mobile UAVs, PeSOA uses penguin search-inspired thermal coverage mapping, where "huddles" collaboratively detect cold spots (coverage holes). Computational complexity is $O(p \cdot f + k)$ per iteration (p – penguins, f – fitness evaluations, k – grid cells).

The advantages of the scheme are:

- It is highly adaptable to dynamic topologies.
- It is secure and easily counterattacks due to the built-in AIS security mechanisms.

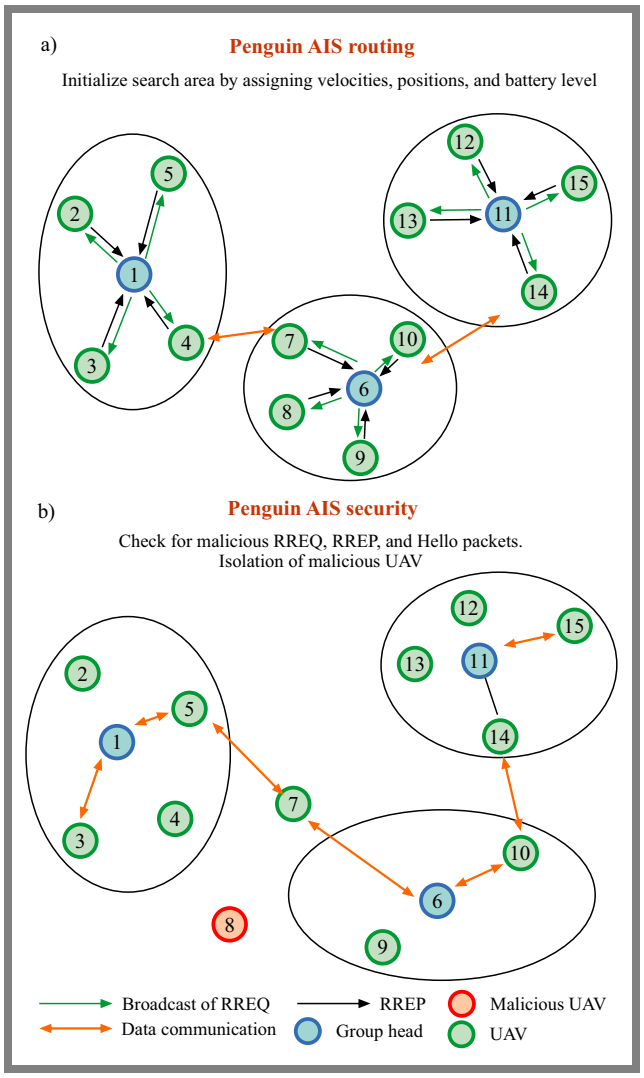

Fig. 13. PeSOA scheme for FANETs: a) routing phase and b) security phase.

The limitations can be summarized as:

- It suffers from moderate computational overhead due to adaptive and secure route selection.
- Poor scalability in dense networks.
- There is no explicit support for real-time traffic handling, limiting emergency communication support.

5. Comparative Analysis

The operational efficacy of bio-inspired FANET routing schemes is inherently related to network coverage characteristics, a dependency that is critically underexplored in the existing literature. Coverage heterogeneity (e.g. urban canyons, sparse deployments, altitude variations) induces profound performance divergences across protocols under identical mission scenarios. This section conducts a systematic comparative analysis of 12 state-of-the-art bio-inspired protocols, evaluating their resilience against: coverage-driven topological constraints (NLoS propagation, node density extremes, 3D mobility) and universal FANET limitations (C1 – energy, C2 – mobility, C3 – scalability, C4 – QoS, C5 – link stability), with its results presented in Tab. 3. By benchmarking protocol behaviors against these dual axes, we expose critical trade-offs, resilience gaps, and domain-specific suitability to guide protocol selection and future research. Table 3 analysis reveals intrinsic coverage-constraint couplings:

- BICSF incurs higher communication overhead than SIL-SIC in large swarms to counter mobility (C2) but achieves only partial coverage due to instability in ultra-dense/sparse deployments.
- BeeAdhoc reduces the delay in rerouting compared to BR-AODV at high speeds (C2) but exhibits low coverage efficacy due to non-adapted MANET mechanisms.
- BAT-COOP maximizes link stability (C5) by increasing retransmissions during weather disruptions. However, it ignores urban multipath fading.
- PeSOA's security measures increase resource requirements compared to lightweight protocols, but they are lower than those in clustering-intensive methods such as BICSF.

The routing protocols explicitly target dynamic FANET environments where UAVs exhibit significant positional changes. Optimization goals focus on maintaining routing stability despite 3D mobility, and all of the protocols address node displacement as a core challenge. Crucially, no protocol assumes static deployments, even GW-COOP's link-aware routing and PICA's multi-hop clustering prioritize adaptability to variations in velocity. This mobility-centric design imperative is evidenced by:

- **Mobility-specific mechanisms**. BeeAdhoc's scout bees relied upon for route rediscovery (high mobility), and BAT-COOP's echolocation used for relay handovers.
- **Universal constraint violations**. The 9/12 protocols violate C2 (mobility) due to delayed response to abrupt maneuvers.
- **Timing constraints**. Emergency data deadlines (SIL-SIC) and route recovery windows (BR-AODV) reflect real-time mobility adaptation needs.

Thus, all of the reviewed protocols fundamentally optimize data routing in networks with frequent UAV repositioning, not quasistatic deployments. Most importantly, altitude-related coverage variance – a characteristic of FANETs – is not addressed by 75% of the protocols (e.g. hybrid BR-AODV, PeSOA), which exacerbates urban shadowing attacks. These cross-protocol trade-offs, as reported, attest to the lack of Pareto-optimal solutions for all coverage-context constraints. Table 4 provides a comparison of the metrics harnessed in the analysis.

6. Open Challenges and Areas for Future Research

Bio-inspired routing protocols offer great potential to cater to the needs of FANET routing. FANETs have distinctive characteristics, including highly dynamic topologies, higher

Tab. 4. Comparison of metrics used in bio-inspired routing systems.

Routing protocol	Routing strategy	Bio-inspired algorithm	Routing function	Mobility model	Routing metrics								
					Adaptability	Link stability	Multiple routes	Computational overhead	Communication overhead	Memory overhead	Energy efficiency	Load balancing	Privacy and security
BICSF	Clustering	GSO, KH	Luciferin value	3D Gauss-Markov	Yes	Low	No	High	High	High	Yes	No	No
SIL-SIC	Clustering	PSO	Fitness function	Swarm waypoint	Yes	High	No	High	Mod.	High	Yes	Yes	No
BIR-SLB	Clustering	AntNet	Path length	Swarm based	Yes	Low	No	Low	Low	Low	No	Yes	No
BR-AODV	Distance-based	Boids of Reynolds	Velocity alignment	3D random walk	Yes	Low	No	High	High	High	Yes	Yes	No
APAR	On-demand reactive	ACO	Pheromone concentration	Random waypoint	Yes	Low	Yes	High	High	High	Mod.	Yes	No
BEE Ad hoc	Swarm intelligence	Bee's wangle dance	Swarm intelligence	Swarm based	Yes	Low	Yes	High	Mod.	High	Yes	Yes	No
HSCS	Clustering	DF, GSO	Luciferin energy level	Swarm based	Yes	High	No	High	High	High	Mod.	Yes	No
BAT-COOP	Multi-hop	Bats' cooperation	Pulse emission/reception	Swarm based	Yes	High	Yes	High	Mod.	High	Mod.	Yes	No
GW-COOP	Multi-hop	GWO	Fitness function	Swarm based	Yes	Low	Yes	Low	Low	Low	Mod.	Yes	No
PICA	Clustering	PP	Cluster head function	Swarm based	Yes	Low	No	High	High	High	Yes	Yes	No
HB-AODV	Multi-objective optimization	HBA	Fitness function	Swarm based	Yes	High	No	High	High	High	Yes	Yes	No
Penguin-AIS	Meta-heuristic optimization	PeSOA, AIS	Membership function	Swarm based	Yes	High	No	Mod.	Mod.	Low	Yes	Yes	Yes

node mobility, energy constraints, sparse deployment, limited processing capabilities, and complex application scenarios that require specific routing protocols. Various routing protocols have been proposed to fulfill these critical requirements, each with its advantages and limitations that ultimately lead to their failure.

Most bio-inspired protocols (10/12) (such as BICSF, BR-AODV, and APAR) have considerable computational burdens that stem from intricate biomimetic functions. These impose severe energy-scalability compromises (violating C1/C3 constraints).

BICSF's dynamic clustering shows significantly greater processing requirements than SIL-SIC, and BeeAdHoc's path discovery shows modest communication overhead despite significant memory needs. Most importantly, BIR-SLB achieves a uniformly low overhead. GW-COOP excels in communication and memory efficiency but with moderate energy costs, resulting in a significantly prolonged network lifetime in large swarms.

Verification-based optimizations validate the high memory overhead from localization caching in SIL-SIC, while BAT-COOP maintains moderate communication overhead using focused probing.

Some of the constraint-specific gaps are listed below:

- Cluster-based protocols (BICSF, SIL-SIC) optimize energy efficiency (C1) in terms of coverage limitation.
- Cooperative designs (BAT-COOP, GW-COOP) optimize link stability (C5), but GW-COOP shows instability during disruption (C5).
- Penguin-AIS alone implements adaptive encryption with moderate computation (C1) cost without sacrificing QoS (C4).
- Biohybrids, such as HB-AODV, demonstrate good QoS (C4) but lack multipath support and scalability (C3).

These overhead properties place inherent limitations on real-world deployment of resource-constrained UAV platforms. In addition, bio-inspired schemes, e.g. artificial bee colony (ABC) [91], bacterial forage optimization (BFO) [92], moth flame optimization (MFO) [93], and red deer optimization (RDO) [94] are still unexplored.

Therefore, more research and innovation are necessary to explore the use of these algorithms and effectively address the complexities of FANET routing [95].

These trade-offs emphasize the importance of choice of an environment-sensitive protocol in FANET deployments. Some of the open research issues will be discussed in the next subsection.

6.1. Dynamic Topology and Sparse Deployment

High mobility of nodes leads to frequent topological changes in the network, resulting in sporadic connectivity [96]. Additionally, nodes can enter and exit sparsely deployed networks when necessary. Several routing protocols have attempted to counter these challenges [97], but most protocols, including bio-inspired routing schemes, fail to ensure proper network coverage. Routing protocols must be designed with increased adaptability and better topology management to counter intermittent connectivity issues. High PDR, reduced end-to-end delay, quick route recovery, reduced overhead of control messages, and high reliability must be supported as well.

6.2. Security Attacks and Data Encryption

Most FANET routing schemes have been designed considering intermittent connectivity, without taking into consideration security threats. Flying nodes are deployed in very harsh or complex situations where they remain unattended, making them easy to capture or be victimized by attackers [98]. Most bio-inspired protocols do not have the provision to provide security and privacy. Under the military application scenario, data transmission must be performed securely and must be protected by suitable encryption algorithms.

Consequently, in the event of a node capture attack, the attacker will need some time to decode the secret information. Therefore, FANET routing protocols must be designed with data encryption in mind, to provide relevant security levels.

6.3. Realistic Mobility Models and Simulation Environment

Although several node mobility models have been presented in the literature, further improvements are still required to simulate nodes in a real-world scenario, with a variable degree of movement in 3D space and multiple mobility models [99]. Furthermore, simulation tools can be designed to predict the node's future locations precisely.

6.4. Energy-efficient Routing with Energy Harvesting

The flying nodes in FANETs are battery powered, with the energy utilized for node mobility, hovering, data processing, payloads, and transmission to intermediate nodes or sink nodes [100]. The amount of residual energy affects each of these operations.

Furthermore, in complex environmental application scenarios, UAVs remain unattended for a long time [101]. Due to the non-availability of any power recharging systems, efficient energy management becomes the most critical routing design issue for future research. Therefore, residual energy of UAVs must be considered when designing routing protocols, especially for FANETs. Additionally, various power transfer schemes and energy harvesting techniques using renewable energy sources should be considered.

6.5. Networking Protocols and the Use of AI

The useful life of FANETs depends on the use of the appropriate networking protocols [102]. Depending on the application scenario and the density of the nodes, the communication protocols must be adaptable and dynamic for improved network throughput, such as selecting wireless technology or the cognitive radio (CR) [103] technique.

Furthermore, principles of artificial intelligence (AI), such as reinforcement learning and deep learning, can enhance system performance by learning from experience [104]. Through reinforcement learning, AI methods can find the optimal node positions and movements. Consequently, UAVs can self-organize to choose optimal routes during flight.

6.6. Cross-layer Architecture

Adopting cross-layer architecture strategies has garnered substantial research interest in improving data communication and network performance, especially while addressing the challenges posed by erroneous link state information caused by the drones' 3D movements occurring in highly dynamic environments [105].

Multiple studies indicate that routing protocols have been designed for conventional layered or hierarchical network architectures. Although the layered network architecture has achieved acceptable performance in a traditional wired network model, it is not on par with the unique requirements of FANETs. Therefore, a cross-layer architecture design will be a fascinating choice allowing to meet FANETs' unique requirements and efficient data routing. Additionally, such a system allows data sharing and feedback among its different layers.

6.7. Load Balancing

In FANET routing protocols, including bio-inspired routing schemes, the source node forwards data packets to intermediate nodes using greedy forwarding or through optimal algorithms, such as those identifying the shortest or lowest cost path, without considering the load of the next receiver, thus reducing network performance due to generating congestion over a particular route.

Therefore, forwarding data packets according to varying traffic loads can increase network throughput. In view of the above, the ability to design load-sensitive routing protocols [106] for FANETs is another critical issue for future research.

6.8. QoS and Standards

QoS is necessary for specific FANET applications where different types of data, such as on-demand, real-time audio and video, as well as video streaming, can be communicated over the network [107]. Therefore, the network must ensure

the essential QoS to meet predetermined service constraints, such as delay, bandwidth, and packet loss. In addition, FANET communication must be standardized to make it universally acceptable.

7. Conclusions

This study explored the efficiency of bio-inspired routing algorithms in overcoming communication obstacles in FANETs. Its main objective was to identify bio-inspired routing algorithms that balance adaptability, scalability, and security for reliable data transmission in UAV-based networks.

We examined several FANET routing protocols utilizing bio-inspired algorithms, drawing inspiration from swarm behaviors such as ACO, ABC, BFO, DF, GSO, GWO, HBA, KH, MFO, PeSOA, PP, PSO, and RDO. Our analysis evaluated their efficiency in addressing critical FANET routing design issues, encompassing adaptability, energy efficiency, link stability, and security considerations.

While many bio-inspired routing protocols exist, our findings reveal that most exhibit trade-offs affecting critical characteristics that are essential for effective FANET routing. The PeSOA scheme is the most promising routing protocol, covering all routing metrics. In contrast, many available schemes suffer from high routing overhead and lack focus on load balancing and network security.

Furthermore, the application of bio-inspired schemes such as ABC, BFO, MFO, and RDO remains unexplored. Finally, several open research issues were identified based on the advantages, limitations, and comparative studies; these issues must be considered before developing robust, reliable, and application-specific FANET routing protocols.

References

[1] H. Zhao, H. Wang, W. Wu, and J. Wei, "Deployment Algorithms for UAV Airborne Networks Toward On-demand Coverage", *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2015–2031, 2018 (https://doi.org/10.1109/JSAC.2018.2864376).

[2] J. Kwak, J.H. Park, and Y. Sung, "Emerging ICT UAV Applications and Services: Design of Surveillance UAVs", *International Journal of Communication Systems*, vol. 34, no. 2, article no. e4023, 2021 (https://doi.org/10.1002/dac.4023).

[3] S.-Y. Park, C.S. Shin, D. Jeong, and H. Lee, "DroneNetX: Network Reconstruction Through Connectivity Probing and Relay Deployment by Multiple UAVs in Ad Hoc Networks", *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 11192–11207, 2018 (https://doi.org/10.1109/TVT.2018.2870397).

[4] L. Ferranti, L. Bonati, S. D'Oro, and T. Melodia, "SkyCell: A Prototyping Platform for 5G Aerial Base Stations", *2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, Cork, Ireland, 2020 (https://doi.org/10.1109/WoWMoM49955.2020.00062).

[5] I.U. Khan *et al.*, "Smart IoT Control-based Nature Inspired Energy Efficient Routing Protocol for Flying Ad Hoc Network (FANET)", *IEEE Access*, vol. 8, pp. 56371–56378, 2020 (https://doi.org/10.1109/ACCESS.2020.2981531).

[6] J. Zhang, J. Yan, and P. Zhang, "Multi-UAV Formation Control Based on a Novel Back-stepping Approach", *IEEE Transactions on Vehicular Technology*, vol. 69, no. 3, pp. 2437–2448, 2020 (https://doi.org/10.1109/TVT.2020.2964847).

[7] A. Srivastava and J. Prakash, "Future FANET with Application and Enabling Techniques: Anatomization and Sustainability Issues", *Computer Science Review*, vol. 39, art. no. 100359, 2021 (https://doi.org/10.1016/j.cosrev.2020.100359).

[8] G. Skorobogatov, C. Barrado, and E. Salamí, "Multiple UAV Systems: A Survey", *Unmanned Systems*, vol. 8, no. 02, pp. 149–169, 2020 (https://doi.org/10.1142/S2301385020500090).

[9] I. Valiulahi and C. Masouros, "Multi-UAV Deployment for Throughput Maximization in the Presence of Co-channel Interference", *IEEE Internet of Things Journal*, vol. 8, no. 5, pp. 3605–3618, 2020 (https://doi.org/10.1109/JIOT.2020.3023010).

[10] A. Pandey, P.K. Shukla, and R. Agrawal, "An Adaptive Flying Ad-hoc Network (FANET) for Disaster Response Operations to Improve Quality of Service (QoS)", *Modern Physics Letters B*, vol. 34, no. 10, art. no. 2050010, 2020 (https://doi.org/10.1142/S0217984920500104).

[11] M.A. Al-Absi, A.A. Al-Absi, M. Sain, and H. Lee, "Moving Ad Hoc Networks – A Comparative Study", *Sustainability*, vol. 13, no. 11, article no. 6187, 2021 (https://doi.org/10.3390/su13116187).

[12] Y. Cheriguene *et al.*, "COCOMA: a Resource-optimized Cooperative UAVs Communication Protocol for Surveillance and Monitoring Applications", *Wireless Networks*, pp. 4429–4445, 2022 (https://doi.org/10.1007/s11276-022-03031-8).

[13] X. Zhu, F. Vanegas, and F. Gonzalez, "An Approach for Multi-UAV System Navigation and Target Finding in Cluttered Environments", *2020 International Conference on Unmanned Aircraft Systems (ICUAS)*, Athens, Greece, pp. 1113–1120, 2020 (https://doi.org/10.1109/ICUAS48674.2020.9214062).

[14] A. Guillen-Perez, A.-M. Montoya, J.-C. Sanchez-Aarnoutse, and M.-D. Cano, "A Comparative Performance Evaluation of Routing Protocols for Flying Ad-hoc Networks in Real Conditions", *Applied Sciences*., vol. 11, no. 10, art. no. 4363, 2021 (https://doi.org/10.3390/app11104363).

[15] M. Nemati *et al.*, "Non-terrestrial Networks with UAVs: A Projection on Flying Ad-hoc Networks", *Drones*, vol. 6, no. 11, art. no. 334, 2022 (https://doi.org/10.3390/drones6110334).

[16] G. Amponis *et al.,* "A Survey on FANET Routing from a Cross-layer Design Perspective", *Journal of Systems Architecture*, vol. 120, p. 102281, 2021 (https://doi.org/10.1016/j.sysarc.2021.102281).

[17] A.H. Sawalmeh and N.S. Othman, "An Overview of Collision Avoidance Approaches and Network Architecture of Unmanned Aerial Vehicles (UAVs)", *International Journal of Engineering and Technology*, vol. 7, 2018, (https://doi.org/10.14419/ijet.v7i4.35.27395).

[18] S. Bitam, A. Mellouk, and S. Zeadally, "Bio-inspired Routing Algorithms Survey for Vehicular Ad Hoc Networks", *IEEE Communications Surveys & Tutorials*, vol. 17, no. 2, pp. 843–867, 2014 (https://doi.org/10.1109/COMST.2014.2371828).

[19] Y. Zeng, R. Zhang, and T.J. Lim, "Wireless Communications with Unmanned Aerial Vehicles: Opportunities and Challenges", *IEEE Communications Magazine*, vol. 54, no. 5, pp. 36–42, 2016 (https://doi.org/10.1109/MCOM.2016.7470933).

[20] A.V. Leonov, "Applying Bio-inspired Algorithms to Routing Problem Solution in FANET", *Bulletin of the South Ural State University. Series Computer Technologies, Automatic Control, Radio Electronics*, vol. 17, no. 2, pp. 5–23, 2017 (https://doi.org/10.14529/ctcr170201).

[21] M.F. Khan, K.-L.A. Yau, R.M. Noor, and M. A. Imran, "Routing Schemes in FANETs: A Survey", *Sensors*, vol. 20, no. 1, art. no. 38, 2020 (https://doi.org/10.3390/s20010038).

[22] T.R. Beegum, M.Y.I. Idris, M.N.B. Ayub, and H.A. Shehadeh, "Optimized Routing of UAVs Using Bio-inspired Algorithm in FANET: A Systematic Review", *IEEE Access*, vol. 11, pp. 15588–15622, 2023 (https://doi.org/10.1109/ACCESS.2023.3244067).

[23] M.J. Almansor *et al.*, "Routing Protocols Strategies for Flying Ad-hoc Network (FANET): Review, Taxonomy, and Open Research

Issues", *Alexandria Engineering Journal*., vol. 109, pp. 553–577, 2024 (https://doi.org/10.1016/j.aej.2024.09.032).

[24] N. Kumar, D. Puthal, T. Theocharides, and S.P. Mohanty, "Unmanned Aerial Vehicles in Consumer Applications: New Applications in Current and Future Smart Environments", *IEEE Consumer Electronics Magazine*, vol. 8, no. 3, pp. 66–67, 2019 (https://doi.org/10.1109/MCE.2019.2892278).

[25] A. Orsino *et al.*, "Effects of Heterogeneous Mobility on D2D- and Drone-assisted Mission-critical MTC in 5G", *IEEE Communications Magazine*, vol. 55, no. 2, pp. 79–87, 2017 (https://doi.org/10.1109/MCOM.2017.1600443CM).

[26] A.A. Khuwaja *et al.*, "A Survey of Channel Modeling for UAV Communications", *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp. 2804–2821, 2018 (https://doi.org/10.1109/COMST.2018.2856587).

[27] N.K. Tomar *et al.*, "FANet: A Feedback Attention Network for Improved Biomedical Image Segmentation", *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 11, pp. 9375–9388, 2022 (https://doi.org/10.1109/TNNLS.2022.3159394).

[28] N.H. Motlagh, T. Taleb, and O. Arouk, "Low-altitude Unmanned Aerial Vehicles-based Internet of Things Services: Comprehensive Survey and Future Perspectives", *IEEE Internet of Things Journal*, vol. 3, no. 6, pp. 899-922, 2016 (https://doi.org/10.1109/JIOT.2016.2612119).

[29] M. Bacco *et al.*, "Reliable M2M/IoT Data Delivery from FANETs via Satellite", *International Journal of Satellite Communications and Networking*, vol. 37, no. 4, pp. 331–342, 2019 (https://doi.org/10.1002/sat.1274).

[30] G. Sun, D. Qin, T. Lan, and L. Ma, "Research on Clustering Routing Protocol Based on Improved PSO in FANET", *IEEE Sensors Journal*, vol. 21, no. 23, pp. 27168–27185, 2021 (https://doi.org/10.1109/JSEN.2021.3117496).

[31] G.A. Kakamoukas, P.G. Sarigiannidis, and A.A. Economides, "FANETs in Agriculture – A Routing Protocol Survey", *Internet Things*, vol. 18, art. no. 100183, 2022 (https://doi.org/10.1016/j.iot.2020.100183).

[32] L. Gupta, R. Jain, and G. Vaszkun, "Survey of Important Issues in UAV Communication Networks", *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1123–1152, 2015 (https://doi.org/10.1109/COMST.2015.2495297).

[33] X. Li and J. Yan, "LEPR: Link Stability Estimation-based Preemptive Routing protocol for Flying Ad Hoc Networks", *2017 IEEE Symposium on Computers and Communications (ISCC)*, Heraklion, Greece, 2017 (https://doi.org/10.1109/ISCC.2017.8024669).

[34] J. Agrawal, M. Kapoor, and R. Tomar, "A Novel Unmanned Aerial Vehicle-Sink Enabled Mobility Model for Military Operations in Sparse Flying Ad-hoc Network", *Transaction on Emerging Telecommunication Technologies*, vol. 33, no. 5, art. no. e4466, 2022 (https://doi.org/10.1002/ett.4466).

[35] M. Xu *et al.*, "Improving Traditional Routing Protocols for Flying Ad Hoc Networks: A Survey", *2020 IEEE 6th International Conference on Computer and Communications (ICCC)*, Chengdu, China, 2020 (https://doi.org/10.1109/ICCC51575.2020.9345206).

[36] S.N. Karam *et al.*, "Inspection of Unmanned Aerial Vehicles in Oil and Gas Industry: Critical Analysis of Platforms, Sensors, Networking Architecture, and Path Planning", *Journal of Electronic Imaging*, vol. 32, no. 1, art. no. 011006, 2022 (https://doi.org/10.1117/1.JEI.32.1.011006).

[37] M.H. Siddiqi *et al.*, "FANET: Smart City Mobility off to a Flying Start with Self-organized Drone-based Networks", *IET Communications*, vol. 16, no. 10, pp. 1209–1217, 2022 (https://doi.org/10.1049/cmu2.12291).

[38] A. Eroğlu and E. Onur, "Revisiting Slotted ALOHA: Density Adaptation in FANETs", *Wireless Personal Communication*, vol. 124, no. 2, pp. 1711–1740, 2022 (https://doi.org/10.1007/s11277-021-09428-6).

[39] S.K. Maakar *et al.* "Performance Evaluation of AODV and DSR Routing Protocols for Flying Ad hoc Network Using Highway Mobility Model", *Journal of Circuits, Systems and Computers*, vol. 31, no. 1, art. no. 2250008, 2022 (https://doi.org/10.1142/S0218126622500086).

[40] J. Souza *et al.*, "A Proposal for Routing Protocol for FANET: A Fuzzy System Approach with QoE/QoS Guarantee", *Wireless Communications and Mobile Computing*, vol. 2019, art. no. 8709249, 2019 (https://doi.org/10.1155/2019/8709249).

[41] A. Chriki, H. Touati, H. Snoussi, and F. Kamoun, "FANET: Communication, Mobility Models and Security Issues", *Computer Networks*, vol. 163, art. no. 106877, 2019 (https://doi.org/10.1016/j.comnet.2019.106877).

[42] G. Secinti, P.B. Darian, B. Canberk, and K.R. Chowdhury, "SDNs in the Sky: Robust End-to-end Connectivity for Aerial Vehicular Networks", *IEEE Communications Magazine*, vol. 56, no. 1, pp. 16–21, 2018 (https://doi.org/10.1109/MCOM.2017.1700456).

[43] H.R. Hussen, S.-C. Choi, J.-H. Park, and J. Kim, "Performance Analysis of MANET Routing Protocols for UAV Communications", *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)*, pp. 70–72, Prague, Czech Republic, 2018 (https://doi.org/10.1109/ICUFN.2018.8436694).

[44] I. Mahmud and Y.-Z. Cho, "Adaptive Hello Interval in FANET Routing Protocols for Green UAVs", *IEEE Access*, vol. 7, pp. 63004–63015, 2019 (https://doi.org/10.1109/ACCESS.2019.2917075).

[45] C. Li, L. Zheng, W. Xie, and P. Yang, "Ad Hoc Network Routing Protocol Based on Location and Neighbor Sensing", *2018 IEEE International Conference on Computer and Communication Engineering Technology (CCET)*, pp. 1–5, Beijing, China, 2018 (https://doi.org/10.1109/CCET.2018.8542225).

[46] R. Fan *et al.*, "Optimal Node Placement and Resource Allocation for UAV Relaying Network", *IEEE Communications Letters*, vol. 22, no. 4, pp. 808–811, 2018 (https://doi.org/10.1109/LCOMM.2018.2800737).

[47] N. Palmieri and A. Serianni, "Overhead Analysis of a Bio-inspired Routing over FANET", *2022 30th Telecommunications Forum (TELFOR)*, Belgrade, Serbia, 2022 (https://doi.org/10.1109/TELFOR56187.2022.9983785).

[48] A. Khan, F. Aftab, and Z. Zhang, "BICSF: Bio-inspired Clustering Scheme for FANETs", *IEEE Access*, vol. 7, pp. 31446–31456, 2019 (https://doi.org/10.1109/ACCESS.2019.2902940).

[49] A. Yadav, A. Shastri, and S. Verma, "Experimental Analysis of ACO with Modified Firefly and Modified Genetic Algorithm for Routing in FANETs", in *Optical and Wireless Technologies: Proceedings of OWT 2021*, Springer, pp. 81–87, 2022 (https://doi.org/10.1007/978-981-19-1645-8_9).

[50] Z. Zhang, K. Long, J. Wang, and F. Dressler, "On Swarm Intelligence Inspired Self-organized Networking: Its Bionic Mechanisms, Designing Principles and Optimization Approaches", *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 513–537, 2013 (https://doi.org/10.1109/SURV.2013.062613.00014).

[51] A.K. Kar, "Bio Inspired Computing – A Review of Algorithms and Scope of Applications", *Expert Systems with Applications*, vol. 59, pp. 20–32, 2016 (https://doi.org/10.1016/j.eswa.2016.04.018).

[52] K.N. Kaipa and D. Ghose, "Glowworm Swarm Optimization: Algorithm Development", in *Glowworm Swarm Optimization*, pp. 21–56, 2017 (https://doi.org/10.1007/978-3-319-51595-3_2).

[53] A.H. Gandomi and A.H. Alavi, "Krill Herd: A New Bio-inspired Optimization Algorithm", *Communications in Nonlinear Science and Numerical Simulation*, vol. 17, pp. 4831–4845, 2012 (https://doi.org/10.1016/j.cnsns.2012.05.010).

[54] G.-G. Wang, A.H. Gandomi, A.H. Alavi, and D. Gong, "A Comprehensive Review of Krill Herd Algorithm: Variants, Hybrids and Applications", *Artificial Intelligence Review*, vol. 51, pp. 119–148, 2019 (https://doi.org/10.1007/s10462-017-9559-1).

[55] M. Dorigo, M. Birattari, and T. Stutzle, "Ant Colony Optimization", *IEEE Computational Intelligence Magazine*, vol. 1, pp. 28–39, 2006 (https://doi.org/10.1109/MCI.2006.329691).

[56] N.A. Al-Aboody and H.S. Al-Raweshidy, "Grey Wolf Optimization-based Energy-efficient Routing Protocol for Heterogeneous Wireless Sensor Networks", *2016 4th International Symposium on Computa-*

tional and Business Intelligence (ISCBI), Olten, Switzerland, 2016 (https://doi.org/10.1109/ISCBI.2016.7743266).

[57] M.Y. Arafat and S. Moh, "Localization and Clustering Based on Swarm Intelligence in UAV Networks for Emergency Communications", *IEEE Internet of Things Journal*, vol. 6, pp. 8958–8976, 2019 (https://doi.org/10.1109/JIOT.2019.2925567).

[58] G.A. Amran *et al.*, "Efficient and Secure WiFi Signal Booster via Unmanned Aerial Vehicles WiFi Repeater Based on Intelligence Based Localization Swarm and Blockchain", *Micromachines*, vol. 13, art. no. 1924, 2022 (https://doi.org/10.3390/mi13111924).

[59] C. Zhou, H.B. Gao, L. Gao, and W.G. Zhang, "Particle Swarm Optimization (PSO) Algorithm J", *Application Research of Computers*, vol. 12, pp. 7–11, 2003.

[60] W. Bin and S. Zhongzhi, "A Clustering Algorithm Based on Swarm Intelligence", *2001 International Conferences on Info-Tech and Info-Net. Proceedings (Cat. No. 01EX479)*, Beijing, China, 2001 (https://doi.org/10.1109/ICII.2001.983036).

[61] O.T. Abdulhae, J.S. Mandeep, and M. Islam, "Cluster-based Routing Protocols for Flying Ad Hoc Networks (FANETs)", *IEEE Access*, vol. 10, pp. 32981–33004, 2022 (https://doi.org/10.1109/ACCESS.2022.3161446).

[62] F. De Rango, M. Tropea, and P. Fazio, "Bio-inspired Routing over FANET in Emergency Situations to Support Multimedia Traffic", *Proc. of the ACM MobiHoc Workshop on Innovative Aerial Communication Solutions for FIrst REsponders Network in Emergency Scenarios*, pp. 12–17, 2019 (https://doi.org/10.1145/3331053.3335033).

[63] B. Baran and R. Sosa, "A New Approach for AntNet Routing", *Proc. of Ninth International Conference on Computer Communications and Networks (Cat. No. 00EX440)*, pp. 303–308, 2000 (https://doi.org/10.1109/ICCCN.2000.885506).

[64] R. Sosa and B. Baran, "AntNet routing algorithm for data networks based on mobile agents", *Revista Iberoamericana de Inteligencia Artificial*, vol. 5, pp. 75–84, 2001 [Online] Available: (https://www.redalyc.org/pdf/925/92551210.pdf).

[65] P. Jacquet *et al.*, "Optimized Link State Routing Protocol for Ad Hoc Networks", *Proc. of IEEE International Multi Topic Conference, 2001. IEEE INMIC 2001. Technology for the 21st Century*, pp. 62–68, 2001 (https://doi.org/10.1109/INMIC.2001.995315).

[66] I.D. Chakeres and E.M. Belding-Royer, "AODV Routing Protocol Implementation Design", *Proc. of 24th International Conference on Distributed Computing Systems Workshops*, pp. 698–703, 2004 (https://doi.org/10.1109/ICDCSW.2004.1284108).

[67] N.E.H. Bahloul, S. Boudjit, M. Abdennebi, and D.E. Boubiche, "Bio-inspired on Demand Routing Protocol for Unmanned Aerial Vehicles", *2017 26th International Conference on Computer Communication and Networks (ICCCN)*, Vancouver, Canada, 2017 (https://doi.org/10.1109/ICCCN.2017.8038487).

[68] C.W. Reynolds, "Flocks, Herds and Schools: A Distributed Behavioral Model", *Proc. of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, pp. 25–34, 1987 (https://doi.org/10.1145/37401.37406).

[69] R.G. Braga, R.C. Da Silva, A.C. Ramos, and F. Mora-Camino, "Collision Avoidance Based on Reynolds Rules: A Case Study Using Quadrotors", *Information Technology-New Generations*, vol. 558, pp. 773–780, 2018 (https://doi.org/10.1007/978-3-319-54978-1_96).

[70] Y. Yu *et al.*, "Ant Colony Optimization Based Polymorphism-aware Routing Algorithm for Ad Hoc UAV Network", *Multimedia Tools and Applications*, vol. 75, pp. 14451–14476, 2016 (https://doi.org/10.1007/s11042-015-3240-y).

[71] D.B. Johnson, D.A. Maltz, and J. Broch, "DSR: The Dynamic Source Routing Protocol for Multi-hop Wireless Ad Hoc Networks", *Monarch Project at Carnegie Mellon University*, 2001 [Online] Available: (https://www.cse.iitb.ac.in/~mythili/teaching/cs653_spring2014/references/dsr.pdf).

[72] J. Wang, E. Osagie, P. Thulasiraman, and R.K. Thulasiram, "HOPNET: A Hybrid Ant Colony Optimization Routing Algorithm for Mobile Ad Hoc Network", *Ad Hoc Networks*, vol. 7, pp. 690–705, 2009 (https://doi.org/10.1016/j.adhoc.2008.06.001).

[73] M. Aissani, M. Fenouche, H. Sadour, and A. Mellouk, "Ant-DSR: Cache Maintenance Based Routing Protocol for Mobile Ad Hoc Networks", *The Third Advanced International Conference on Telecommunications (AICT'07)*, Morne, Mauritius, 2017 (https://doi.org/10.1109/AICT.2007.14).

[74] A.M. Ellison, "Bayesian Inference in Ecology", *Ecology Letters*, vol. 7, pp. 509–520, 2004 (https://doi.org/10.1111/j.1461-0248.2004.00603.x).

[75] H.F. Wedde *et al.*, "BeeAdHoc: An Energy Efficient Routing Algorithm for Mobile Ad Hoc Networks Inspired by Bee Behavior", *Proc. of the 7th Annual Conference on Genetic and Evolutionary Computation*, pp. 153–160, 2005 (https://doi.org/10.1145/1068009.1068034).

[76] J.R. Riley *et al.*, "The Flight Paths of Honeybees Recruited by the Waggle Dance", *Nature*, vol. 435, pp. 205–207, 2005 (https://doi.org/10.1038/nature03526).

[77] S. Dong, T. Lin, J.C. Nieh, and K. Tan, "Social Signal Learning of the Waggle Dance in Honey Bees", *Science*, vol. 379, pp. 1015–1018, 2023 (https://doi.org/10.1126/science.ade1702).

[78] F. Aftab, A. Khan, and Z. Zhang, "Hybrid Self-organized Clustering Scheme for Drone Based Cognitive Internet of Things", *IEEE Access*, vol. 7, pp. 56217–56227, 2019 (https://doi.org/10.1109/ACCESS.2019.2913912).

[79] Y. Meraihi, A. Ramdane-Cherif, D. Acheli, and M. Mahseur, "Dragonfly Algorithm: A Comprehensive Review and Applications", *Neural Computing and Applications*, vol. 32, pp. 16625–16646, 2020 (https://doi.org/10.1007/s00521-020-04866-y).

[80] S. Mirjalili, "Dragonfly Algorithm: A New Meta-heuristic Optimization Technique for Solving Single-objective, Discrete, and Multi-objective Problems", *Neural Computing and Applications*, vol. 27, pp. 1053–1073, 2016 (https://doi.org/10.1007/s00521-015-1920-1).

[81] G. Jones and B.M. Siemers, "The Communicative Potential of Bat Echolocation Pulses", *Journal of Comparative Physiology A*, vol. 197, pp. 447–457, 2011 (https://doi.org/10.1007/s00359-010-0565-x).

[82] S. Hameed *et al.*, "An Improved iBAT-COOP Protocol for Cooperative Diversity in FANETs", *Computers, Materials & Continua*, vol. 67, pp. 2527–2546, 2021 (https://doi.org/10.32604/cmc.2021.013896).

[83] S. Hameed *et al.*, "Link and Loss Aware GW-COOP Routing Protocol for FANETs", *IEEE Access*, vol. 9, pp. 110544–110557, 2021 (https://doi.org/10.1109/ACCESS.2021.3101361).

[84] R.O. Peterson *et al.*, "Leadership Behavior in Relation to Dominance and Reproductive Status in Gray Wolves, Canis Lupus", *Canadian Journal of Zoology*, vol. 80, pp. 1405–1412, 2002 (https://doi.org/10.1139/z02-124).

[85] S. Yang *et al.*, "Bio-inspired Multi-hop Clustering Algorithm for FANET", *Ad Hoc Networks*, vol. 154, art. no. 103355, 2024 (https://doi.org/10.1016/j.adhoc.2023.103355).

[86] X. Zhang *et al.*, "An Improved Physarum Polycephalum Algorithm for the Shortest Path Problem", *The Scientific World Journal*, pp. 1–9, 2014 (https://doi.org/10.1155/2014/487069).

[87] A.M. Reynolds and C.J. Rhodes, "The Lévy Flight Paradigm: Random Search Patterns and Mechanisms", *Ecology*, vol. 90, pp. 877–887, 2009 (https://doi.org/10.1890/08-0153.1).

[88] A. Kout *et al.*, "A Hybrid Optimization Solution for UAV Network Routing", *Engineering, Technology & Applied Science Research*, vol. 13, pp. 10270–10278, 2023 (https://doi.org/10.48084/etasr.5661).

[89] F. A. Hashim *et al.*, "Honey Badger Algorithm: New Metaheuristic Algorithm for Solving Optimization Problems", *Mathematics and Computers in Simulation*, vol. 192, pp. 84–110, 2022 (https://doi.org/10.1016/j.matcom.2021.08.013).

[90] A. Beghriche, "An Adaptive Secure and Efficient Bio-inspired Routing Protocol for Effective Cooperation in FANETs", *Ingénierie des Systèmes d'Information*, vol. 28, pp. 49–66, 2023 (https://doi.org/10.18280/isi.280106).

[91] F.S. Abu-Mouti and M.E. El-Hawary, "Overview of Artificial Bee Colony (ABC) Algorithm and its Applications", *2012 IEEE Interna-*

tional Systems Conference SysCon 2012, Vancouver, Canada, 2012 (https://doi.org/10.1109/SysCon.2012.6189539)

[92] K.M. Passino, "Bacterial Foraging Optimization", in: *Innovations and Developments of Swarm Intelligence Applications*, pp. 219–234, 2010 (https://doi.org/10.4018/978-1-4666-1592-2.ch013).

[93] S. Mirjalili, "Moth-flame Optimization Algorithm: A Novel Nature-inspired Heuristic Paradigm", *Knowledge-Based Systems*, vol. 89, pp. 228–249, 2015 (https://doi.org/10.1016/j.knosys.2015.07.006).

[94] A.F. Fard and M. Hajiaghaei-Keshteli, "Red Deer Algorithm (RDA); A New Optimization Algorithm Inspired by Red Deer's Mating", *International Conference on Industrial Engineering*, Tehran, Iran, 2016.

[95] A. Meier and J.S. Thompson, "Cooperative Diversity in Wireless Networks", *6th IEE International Conference on 3G and Beyond,* Stevenage, UK, 2005.

[96] V.A. Maistrenko, L.V. Alexey, and V.A. Danil, "Experimental Estimate of Using the Ant Colony Optimization Algorithm to Solve the Routing Problem in FANET", *2016 International Siberian Conference on Control and Communications (SIBCON)*, Moscow, Russia, 2016 (https://doi.org/10.1109/SIBCON.2016.7491805).

[97] A. Heidari, N.J. Navimipour, M. Unal, and G. Zhang, "Machine Learning Applications in Internet-of-drones: Systematic Review, Recent Deployments, and Open Issues", *ACM Computing Surveys*, vol. 55, pp. 1–45, 2023 (https://doi.org/10.1145/3571728).

[98] O. Ceviz, S. Sen, and P. Sadioglu, "A Survey of Security in UAVs and FANETs: Issues, Threats, Analysis of Attacks, and Solutions", *IEEE Communications Surveys & Tutorials*, 2024 (https://doi.org/10.1109/COMST.2024.3515051).

[99] H. Yang and Z. Liu, "An optimization routing protocol for FANETs", *EURASIP J. on Wireless Communications and Networking*, art. no. 120, 2019 (https://doi.org/10.1186/s13638-019-1442-0).

[100] S. Bharany *et al.*, "Wildfire Monitoring Based on Energy Efficient Clustering Approach for FANETS", *Drones*, vol. 6, art. no. 193, 2022 (https://doi.org/10.3390/drones6080193).

[101] D. Bein, W. Bein, A. Karki, and B.B. Madan, "Optimizing Border Patrol Operations Using Unmanned Aerial Vehicles", *2015 12th International Conference on Information Technology-New Generations*, Las Vegas, USA, 2015 (https://doi.org/10.1109/ITNG.2015.83).

[102] K. Mariyappan, M.S. Christo, and R. Khilar, "Implementation of FANET Energy Efficient AODV Routing Protocols for Flying Ad Hoc Networks FEEAODV", *Materialstoday Proceeding*, 2021 (https://doi.org/10.1016/j.matpr.2021.02.673).

[103] N. Mansoor *et al.*, "Cognitive Radio Ad Hoc Network Architectures: A Survey", *Wireless Personal Communications*, vol. 81, pp. 1117–1142, 2015 (https://doi.org/10.1007/s11277-014-2175-3).

[104] A. Hussain *et al.*, "Taking FANET to Next Level: The Contrast Evaluation of Moth-and-ant with Bee Ad Hoc Routing Protocols for Flying Ad Hoc Networks", *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal*, vol. 10, pp. 321–337, 2021 (https://doi.org/10.14201/ADCAIJ2021104321337).

[105] H.S. Mansour *et al.*, "Cross-layer and Energy-aware AODV Routing Protocol for Flying Ad Hoc Networks", *Sustainability*, vol. 14, art. no. 8980, 2022 (https://doi.org/10.3390/su14158980).

[106] R. Khaleghnasab *et al.*, "An Energy and Load Aware Multipath Routing Protocol in the Internet of Things", *Management Strategies and Engineering Sciences*, vol. 7, pp. 1–15, 2020 (https://doi.org/10.61838/msesj.7.3.1).

[107] P. Kaur, A. Singh, and S.S. Gill, "RGIM: an Integrated Approach to Improve QoS in AODV, DSR and DSDV Routing Protocols for FANETS Using the Chain Mobility Model", *The Computer Journal*, vol. 63, pp. 1500–1512, 2020 (https://doi.org/10.1093/comjnl/bxaa040).

Santosh Kumar, M.Tech., Research Scholar
Computer Science Department
https://orcid.org/0000-0002-0556-7351
E-mail: santosh4103@gmail.com
Himachal Pradesh University, Shimla, India
https://www.hpuniv.ac.in

Amol Vasudeva, Ph.D., Assistant Professor
Computer Science and Engineering Department
https://orcid.org/0000-0002-6909-0820
E-mail: amol.vasudeva@juit.ac.in
Jaypee University of Information Technology, Solan, Himachal Pradesh, India
https://www.juit.ac.in

Manu Sood, Ph.D., Professor
Computer Science Department
https://orcid.org/0000-0002-4946-9550
E-mail: soodm_67@yahoo.com
Himachal Pradesh University, Shimla, India
https://www.hpuniv.ac.in

Enhancing Leaf Area Segmentation by Using Attention Gates and Knowledge Distillation in UNet Architecture

A. Shamim Banu[1,2] and S. Deivalakshmi[1]

[1]*National Institute of Technology, Tiruchirappalli, Tamilnadu, India,*
[2]*Government Polytechnic College, Tiruchirappalli, Tamilnadu, India*

https://doi.org/10.26636/jtit.2025.3.2079

Abstract — Accurate segmentation of leaf regions plays a vital role in plant phenotyping and agricultural analysis. This paper presents AKDUNet, a lightweight UNet-based architecture that integrates attention gates and knowledge distillation to improve segmentation performance while minimizing computational complexity. The architecture replaces traditional skip connections with attention gates to focus on salient spatial features and employs a two-stage training pipeline, where a compact student model learns from a deeper teacher model using a tailored distillation loss function. AKDUNet is evaluated on two benchmark datasets (CWFID and Sunflower) and outperforms a range of state-of-the-art models, including UNet++, Inception UNet, VGG-based UNets, SDUNet, INSCA UNet, and SegFormer. Ablation studies confirm the advantages of attention modules, and qualitative analyses using Grad-CAM visualizations reveal the model's ability to effectively focus on crucial leaf structures. The results demonstrate that AKDUNet is not only computationally efficient but also highly accurate, making it suitable for real-time deployment in resource-constrained agricultural environments.

Keywords — attention gate, knowledge distillation, modified light weight UNet, semantic segmentation

1. Introduction

Plant phenotyping has recently gained more attention from researchers due to its potential to enhance high-yield plant capabilities and augment food security. It is a key tool for understanding plant genetics, plant-environment interactions, and various traits [1]. It also involves creating new technologies to improve plant yields and address the aforementioned issues. Furthermore, plant phenotyping is crucial for examining plant growth, yield, and internal structure.

Plant image analysis is a key technique for plant phenotyping, facilitating the evaluation of plant traits, growth forecasts, and spatial details of plants. Manual measurements of visual characteristics are costly. Hence, there is a need for automated solutions. Recent studies suggest that deep learning (DL) techniques, a contemporary AI approach, are becoming increasingly important in plant phenotyping due to their advanced features [2].

In plant phenotyping, leaf area segmentation is crucial for analyzing plant growth. However, the task becomes challenging when dealing with small leaves or when many leaves overlap. Furthermore, the effectiveness of leaf area segmentation can be significantly influenced by factors such as image capture angles and lighting conditions [3].

The proposed work introduces an AKDUNet model, designed to enhance leaf area segmentation. It boosts segmentation accuracy while keeping the model simple, by relying on a smaller number of parameters. It is based on the UNet architecture, known for its effective leaf segmentation, and is compared with other known segmentation models, in example UNet++, Inception UNet, SDUNet, and INCSA UNet, as well as pre-trained deep learning models such as VGG16-UNet, VGG19-UNet, ResNet-UNet, and SegFormer. The performance of the AKDUNet model is assessed using the crop weed field image dataset (CWFID) and the Sunflower data set for plant phenotyping. Both qualitative and quantitative analyses demonstrate that AKDUNet outperforms current leaf segmentation techniques.

The key contributions of this work are as follows.

- **Knowledge distillation.** This method trains a smaller student model P^s to replicate the performance of a larger teacher model P^t. By transferring the core knowledge from the teacher to the student, this approach enables the smaller model to achieve similar performance with fewer parameters, thus reducing the computational burden and memory requirements.
- **Attention gate mechanism.** Traditional skip connections are replaced with an attention gate mechanism. This technique allows the model to focus on the most relevant parts of the input data and ignore less important information. By emphasizing significant features and filtering out irrelevant details, attention gates enhance the model's efficiency and accuracy without significantly increasing computational demands.
- **Performance evaluation.** The effectiveness of the AKDUNet model is evaluated in comparison to advanced models using metrics such as IoU score, F1 score, and Dice coefficient loss, providing a thorough evaluation of its performance.
- **Assessment of computational complexity.** This research evaluates the computational efficiency of the proposed ap-

proach by analyzing floating-point operations per second (FLOPs), the number of model parameters, and the inference time, and thus providing a comprehensive assessment of both performance and resource utilization.

The structure of the remaining sections is as follows. Section 2 reviews related work on semantic segmentation, attention modules, and knowledge distillation learning. Section 3 provides a detailed description of the key components of the AKDUNet model, including its architecture, implementation, loss functions, and training algorithm. Section 4 presents an in-depth analysis of the experimental setup and results. Section 5 discusses the major advancements and limitations of the approach and provides the conclusion.

2. Related Works

The field of knowledge distillation continues to evolve with several advances that enhance the efficiency and effectiveness of computer vision models. Various techniques such as channel attention, transformers, self-attention, and novel feature distillation modules mature to optimize performance by reducing model size and complexity. Each of these contributions reflects ongoing efforts to balance the trade-off between model performance and computational resources, which is crucial for deploying advanced computer vision technologies in mobile and embedded systems.

The authors of [4] introduced a method that uses channel attention and feature maps within a transformer framework to facilitate knowledge transfer between a large teacher model P^t and a smaller student model P^s. Paper [5] pioneered the application of KD frameworks, specifically in the field of salient object detection (SOD), which involves identifying the most important objects in an image. In [6], a method for efficient semantic segmentation was proposed that combines self-attention mechanisms with self-distillation. The authors of [7] developed a common feature distillation module designed to consolidate multi-stream information into a spatially coherent single-stream representation.

Semantic segmentation involves assigning each pixel in an image to a specific category. Modern RGB-D semantic segmentation methods often rely on deep learning techniques which have significantly advanced the field due to their strong automatic learning and feature extraction capabilities. In such a context, the authors of [8] improved deep learning models by incorporating channel attention mechanisms to enhance features, while in [9], self-attention modules were introduced to refine the features extracted by the encoder.

In [10], an RFNet was developed which balances performance and speed for real-time applications. [11] proposed an ESANet which dynamically adjusts feature space representations by weighting the outputs of each encoding block. In [12], high-level features were applied to dynamically adjust the decoding structure of deep learning networks.

Several adaptations of the traditional UNet architecture have been developed to improve segmentation tasks. The authors of [13] presented an improved UNet++ design which adds deconvolution blocks to the skip connections. This modification enriches the semantic information in the decoder, enabling deeper supervision. In [14], the UNet architecture was introduced, incorporating inception layers and combining binary cross-entropy, the Dice coefficient, and intersection over union to boost performance.

The authors of [15] developed a SDUNet which features structured dropout in all UNet layers. This approach helps prevent overfitting by eliminating some semantic details from the network. [16] explored the INCSA UNet architecture which integrates inception blocks with spatial attention mechanisms. This model uses parallel and sequential layers to effectively extract key features.

In [17], a compressed version of a UNet customized for plant disease segmentation was presented. This streamlined model is more storage-efficient and performs faster than the original UNet. Paper [18] proposed SegFormer, a transformer-based semantic segmentation model that combines hierarchical encoding with a lightweight MLP decoder. The model eliminates the need for positional encoding, improving robustness to varying input resolutions.

Traditional semantic segmentation methods frequently face challenges that can result in inaccurate predictions. To address these issues, this paper proposes an AKDUNet model that integrates attention mechanisms and knowledge distillation with the UNet architecture. This approach effectively reduces the number of parameters compared to that used by conventional UNet models, thereby reducing computational complexity. Despite the reduction in the number of parameters, the AKDUNet model aims to enhance performance by utilizing attention mechanisms to focus on important features and employing knowledge distillation to preserve essential knowledge from more complex models. This customized solution specifically addresses the limitations of existing methods.

3. Proposed Methodology

This section presents a detailed overview of the AKDUNet model, including its architecture, implementation, loss functions, and training algorithm for knowledge distillation.

3.1. Implementation of Framework

Knowledge distillation is a deep learning technique that enables the transfer of knowledge from a large to a smaller model. It is particularly useful for deploying models in resource-constrained environments, where computational power and memory are limited. To enhance segmentation accuracy, the proposed method transfers knowledge from a larger teacher model P^t to a smaller student model P^s. Figure 1 illustrates the model's architecture, which includes two deep convolutional networks such as the student model P^s with weights θ_s and the teacher model P^t with weights θ_t.

The work uses a UNet backbone [19] for both the teacher model P^t and the student model P^s, which helps to effectively capture and maintain detailed features. The network configurations are as follows: teacher model P^t (depth, width);

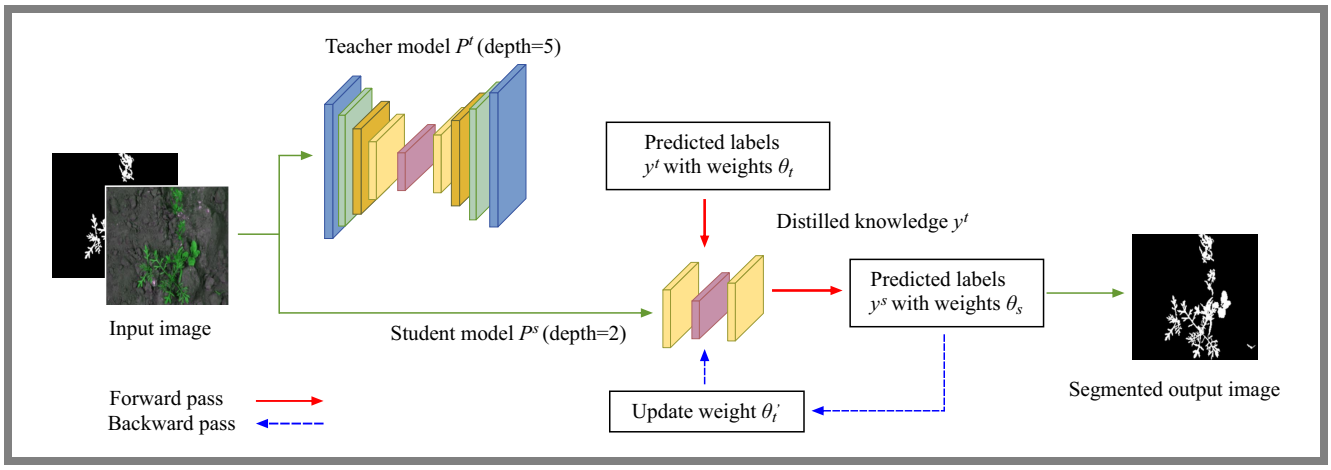

Fig. 1. Architecture of the AKDUNet model for leaf segmentation.

$n \in \{5, 3\}$ and student UNet P^s (depth, width); $n \in \{2, 3\}$, where depth denotes the number of layers and width denotes the number of convolutions in each layer which are depicted in Fig. 1. These variations are intended to explore the trade-offs between detailed feature extraction, the number of parameters, and overall model complexity.

Table 1 presents a detailed architectural comparison between the teacher UNet (depth 5) and the student UNet (depth 2). The teacher model follows a deeper architecture with more convolutional blocks and a higher parameter counts of 2 354 785 to ensure rich feature extraction and accurate segmentation.

On the contrary, the student model is a lightweight version with significantly fewer layers and 13 681 parameters, designed for faster inference and deployment in resource-constrained environments. Despite its simplicity, the Student UNet retains the core structural elements of UNet, including convolution, pooling, up-sampling, and skip connections, allowing it to perform efficient segmentation with reduced computational cost.

The number of parameters in a 2D convolutional layer is given by the formula:

$$CONV_{Par} = (K_H \times K_W \times C_{IN} + 1) \times C_{OUT} \,, \qquad (1)$$

where: K_H – kernel height, K_W – kernel width, C_{IN} – number of input channels, C_{OUT} – number of output filters.

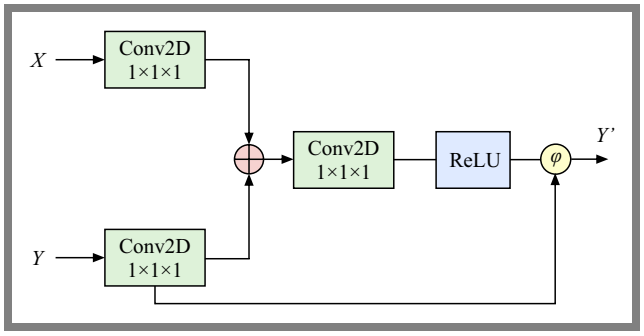

Fig. 2. Attention gate module.

Instead of directly passing features from the encoder to the decoder through conventional skip connections, the proposed method incorporates attention gates within the skip connections to enhance feature selection, as illustrated in Fig. 2. In this design, the encoder feature map is used as one input X to the attention gate, while the gating signal from the decoder's previous stage is used as the second input Y. The attention gate computes attention coefficients that selectively highlight relevant regions in the feature maps while suppressing less informative activations. This is mathematically represented as:

$$Y' = ReLU(Y) \odot Y \,, \qquad (2)$$

where: $\odot$ – element-wise multiplication and Y' – output after applying the attention.

The multiplication with Y ensures that only the most significant features of the decoder contribute to the subsequent layers. By adaptively focusing on important spatial regions, these attention gates help improve segmentation accuracy by refining the information passed from the encoder to the decoder. The training process is carried out in two stages. In the first stage, teacher model P^t is designed with an encoder decoder architecture of depth 5. The model is trained using the Dice coefficient loss function, with the objective of optimizing its performance based on F1 and IoU scores. Once training has been completed, the parameters of the teacher model, including weights θ_t, are frozen to ensure that they remain fixed during the subsequent distillation phase. The trained teacher model P^t generates fixed predictions y^t, which will be used as ground truth in the second stage.

In the second stage, a smaller student model P^s with a depth of 2 is used. The student model has fewer parameters than the teacher model, allowing for a more compact architecture. During this stage, the outputs of teacher model P^t serve as ground truth y^t for calculating the knowledge distillation loss L_{KD}, which quantifies the discrepancy between the teacher's ground truth predictions y^t and the student model's predictions y^s. It is also advantageous to train the student model with ground truth labels y to get predictions y^s produced by the student model itself to calculate standard loss L_s. This loss is derived using the Dice coefficient loss, which quanti-

Tab. 1. Architecture and parameters of teacher and student UNet models.

Stage	Teacher UNet (depth 5)	Student UNet (depth 2)
Input	128 × 128 × 3	128 × 128 ×1
Block 1	3 × Conv2D (16 filters ReLU) MaxPool (2 × 2) Dropout (0.5)	3 × Conv2D (16 filters ReLU) MaxPool (2 × 2) Dropout (0.5)
Block 2	3 × Conv2D (32 filters ReLU) MaxPool (2 × 2) Dropout (0.5)	–
Block 3	3 × Conv2D (64 filters ReLU) MaxPool (2 × 2) Dropout (0.5)	–
Block 4	3 × Conv2D (128 filters ReLU) MaxPool (2 × 2) Dropout (0.5)	–
Bottleneck	2 × Conv2D (256 filters ReLU)	2 × Conv2D (32 filters ReLU)
Up block 4	UpSample (2 × 2) Conv2D (128) Concat 2 × Conv2D Dropout (0.5)	–
Up block 3	UpSample (2 × 2) Conv2D (64) Concat 2 × Conv2D Dropout (0.5)	–
Up block 2	UpSample (2 × 2) Conv2D (32) Concat 2 × Conv2D Dropout (0.5)	–
Up block 1	UpSample (2 × 2) Conv2D (16) Concat 2 × Conv2D Dropout (0.5)	UpSample (2 × 2) Conv2D (16) Concat 2 × Conv2D Dropout (0.5)
Output	3 × Conv2D (labels 1 × 1) Activation (Sigmoid/Softmax)	3 × Conv2D (labels 1 × 1) Activation (Sigmoid/Softmax)
Total parameters	2 354 785	13 681

fies the predictions of similarity between student model y^s and ground truth y.

This total distillation loss combines both standard loss L_s and knowledge distillation loss L_{KD}, ensuring that the student model learns to approximate both teacher predictions y^t and ground truth y effectively. The combination of these losses is used to update the weights of student model θ_s, resulting in an optimized student model with updated parameters θ'_s.

Loss functions play a vital role in image segmentation as they guide the model to segment images into distinct regions. The choice of loss function depends on the specific requirements of the segmentation task and the nature of the segmentation process. To evaluate the proposed knowledge distillation approach, three key loss functions such as knowledge distillation loss L_{KD}, student loss L_s, and overall distillation training loss L_T are considered.

3.2. Knowledge Distillation Loss

In knowledge distillation, knowledge is transferred from teacher model P^t to student model P^s by minimizing the discrepancy between the predictions of teacher labels $y^s \in R^{h\times w}$ and student labels $y^t \in R^{h\times w}$, where y^s and y^t represent predictions of student model P^s and teacher model P^t respectively, with h and w referring to the height and width of the input image. The equations used to derive knowledge distillation loss L_{KD} are as follows:

$$p^t = \text{softmax}\left(\frac{y^t}{\tau}\right), \tag{3}$$

$$p^s = \text{softmax}\left(\frac{y^s}{\tau}\right), \tag{4}$$

$$L_{KD} = \frac{2}{\tau}\left(\sum_{i=1}^{B}\sum_{j=1}^{C} p_{ij}^t \log\frac{p_{ij}^t}{p_{ij}^s}\right), \tag{5}$$

where B is the batch size, C is the number of classes, and τ is the temperature factor that typically ranges between 1 and 5 and controls the smoothness of the probability distributions p^t and p^s.

The purpose of Eqs. (3) and (4) is to convert the raw logits y^t and y^s from the teacher model and student model into probability distributions. The softmax function is applied to the logits to normalize them, transforming them into values between 0 and 1 that sum up to 1 across all classes. This ensures that both teacher model P^t and student model P^s outputs are in the form of probabilities, allowing for a meaningful comparison of their predictions.

The equation for L_{KD} shown in Eq. (5) quantifies how much the student model probability distribution differs from the teacher model P^t distribution. In Eq. (5), p^t and p^s are probabilistic prediction values of teacher model P^t and student model P^s, respectively, while $\log\frac{p_{ij}^t}{p_{ij}^s}$ is the Kullback-Leibler divergence which measures the difference between the teacher and the student probability distributions. The logarithm computes how far the student model P^s distribution is from the teacher model P^t distribution.

This loss function can also be interpreted as the cross-entropy between the distributions and encourage the student to match the teacher distribution. By doing this, student model P^s learns not only the prediction of the final class, but also the relative likelihoods of different classes, enabling it to better approximate the decision-making process.

3.3. Standard Loss

It is beneficial to train the student together with ground truth labels y and y^s labels predicted by student model P^s to obtain standard loss L_s. The Dice coefficient loss is utilized to get the standard loss which is:

$$L_S = 1 - 2\,\frac{|y \cap y^s|}{|y + y^s|}, \tag{6}$$

where $|y \cap y^s|$ is the number of pixels that are correctly predicted by student model P^s with respect to ground truth y, $|y + y^s|$ represent the total number of pixels in ground truth y and predicted labels y^s of student model P^s.

3.4. Total Distillation Loss

Equation (7) represents the overall training loss as a weighted combination of standard loss L_s and knowledge distillation loss L_{KD}. The equation is essential for achieving two key objectives during model training, i.e. minimizing L_s and reducing L_{KD}.

$$L_T = \lambda\, L_s + (1 - \lambda)\, L_{KD}\ . \tag{7}$$

Parameter λ is the weight factor, typically assuming a value between 0 and 1, that acts as a hyperparameter controlling the trade-off between standard loss and knowledge distillation loss.

3.5. Backpropagation

To update the weights of student model P^s using backpropagation, the gradients of total loss L_T with respect to student model weights θ_s are calculated in the following form:

$$\nabla\theta_s = \frac{L_s}{\theta_s}\ , \tag{8}$$

$$\nabla\theta_{KD} = \frac{L_{KD}}{\theta_s}\ , \tag{9}$$

$$\nabla\theta_T = \alpha\, \nabla\theta_s + (1 - \alpha)\, \nabla\theta_{KD}\ , \tag{10}$$

where $\nabla\theta_s$ is the gradient of standard loss L_s with respect to student weights θ_s, $\nabla\theta_{KD}$ is the gradient of L_{KD} with respect to student weights θ_s.

Once $\nabla\theta_T$ gradients are calculated, propagate these gradients backward through student model P^s from the output layer to the input layer, which updates the weights of each layer according to:

$$\theta'_s = \theta_s - \eta\, \nabla\theta_T\ , \tag{11}$$

where θ'_s is the updated weight and η is the learning rate.

3.6. Training Algorithm

The Algorithm 1 outlines the knowledge distillation process during training, where the weight of a student model is updated by backpropagation through gradient calculation.

Algorithm 1 Training process.

Start

Input

1: Training data $D = \{x_i^t, y_i^t\}$, where x_i^t is the input image, y_i^t is the corresponding ground truth label and initialize the hyper parameters such as optimizer, learning rate η, batch size B, number of classes C

2: Train teacher model P^t and obtain corresponding weights θ^t

Backpropagation

3: Initialize student model P^s with the hyper parameters and load teacher model P^t with weight θ^t

Forward pass

4: Set weight factor λ, temperature factor τ and employ student model P^s and teacher model P^t

5: **for** each mini batch B and input $DB = \{x_b^t, y_b^t\}$ **do** forward propagation and compute s weight θ_s, knowledge divergence loss L_{KD}, student loss L_s and total distillation loss L_T using Eqs. (1)–(5)

6: **end for**

Backward pass

7: Compute gradients $\nabla\theta_T$ as defined by Eqs. (6)–(8)

Update

8: Update the weight of student model weight θ_s using the calculated gradients $\nabla\theta_T$ and obtain updated student model weight θ'_s using Eq. (9)

9: Increment the iteration count i

Termination

10: Repeat steps 5 to 8 until the model converges or until a maximum number of iterations is reached

End

4. Experimental Setup and Results

The proposed methodology is evaluated using two benchmark datasets that present diverse and challenging conditions, the CWFID dataset [3] and the Sunflower dataset [20]. The CWFID dataset consists of 60 high-resolution images with detailed pixel-level annotations, collected by the *Bonirob* agricultural robot on an organic carrot farm. The images show carrot plants at the early true leaves, where dense plant clusters with complex double-compound leaves and secondary structures create significant challenges for segmentation due to frequent overlapping and occlusions.

Additionally, the Sunflower dataset, a publicly available resource, is used for crop and weed segmentation experiments. This dataset was acquired by an agricultural robot in sunflower fields in Jesi, Italy. It comprises 500 scene images organized into three subsets representing different stages of crop growth: emergence, intermediate growth, and the final stage before chemical treatment. The images were captured over various days and times to include natural variations in lighting and field conditions.

To facilitate evaluation, the datasets are divided into 80% for training and 20% for testing, with sample images and corresponding ground truth masks shown in Fig. 3.

To measure the effectiveness of the AKDUNet model, performance metrics such as Dice coefficient loss, F1 score, and IoU score are considered in the following manner:

$$Dice = 1 - \frac{2 \times TP}{(TP + FP) + (TP + FN)}\ , \tag{12}$$

$$Precision = \frac{TP}{TP + FP}\ , \tag{13}$$

$$Recall = \frac{TP}{TP + FN}\ , \tag{14}$$

$$\text{F1 score} = \frac{2 \times Precision \times Recall}{Precision + Recall}\ , \tag{15}$$

$$\text{IoU score} = \frac{TP}{TP + FN + FP}\ , \tag{16}$$

where TP, TN, FN, and FP are true positives, true negatives, false negatives, and false positives, respectively.

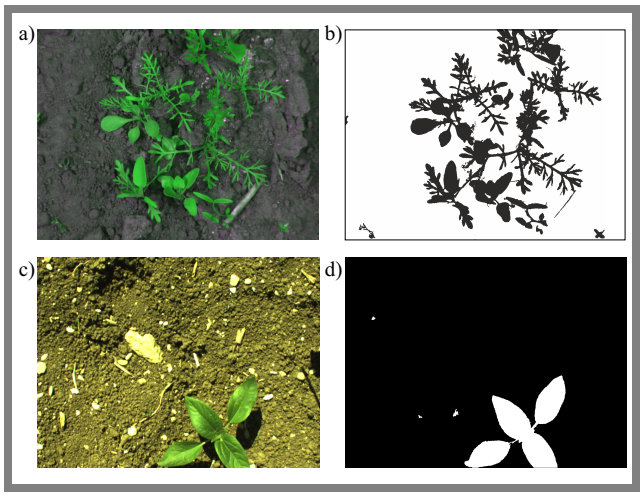

Fig. 3. Sample images and corresponding ground truth masks.

Tab. 2. AKDUNet model parameters.

Parameters	Teacher UNet P^t	Student UNet P^s
Depth	5	2
Filters	16, 32, 64, 128, 256	16, 32
Kernel size	3 × 3	2 × 2
Activation function	ReLU	ReLU
No. of parameters	2.35 M	0.013 M
Inference time [ms]	168 ms	65 ms

4.1. Model Parameters and Implementation Details

The UNet architecture is used as the backbone for both the teacher and student models. To evaluate the AKDUNet model, Google Colab is used equipped with NVIDIA GPUs and 12 GB of RAM to handle the necessary computations. The input images, originally 1296 × 966 pixels in resolution, are resized to 128 × 128 pixels for the evaluation process.

Table 2 presents the key parameters of the two models. The teacher model P^t, with a deeper architecture, employs filters of increasing size and a 3 × 3 kernel size. It has 2.35 m parameters and requires 168 ms for inference of the test image, reflecting its more complex design and computational demand. In contrast, the student model P^s, with a shallower architecture, uses fewer filters, a smaller 2 × 2 kernel size, and has significantly fewer parameters (0.29 M). This model achieves a faster inference time of 65 ms, making it more efficient and computationally powerful compared to the teacher model. Despite the difference in complexity, both models utilize the ReLU activation function, ensuring similar activation behavior across both architectures.

Table 3 describes the hyper-parameters used for the proposed model design. The learning rate η is set to 0.0001, allowing the model to make small, controlled updates to its weights during training. The Adam optimizer is employed to adaptively adjust the learning rate based on gradients, which helps the model to converge efficiently. A temperature factor τ of 4 softens the teacher model output probabilities in the knowledge distillation process. The weight factor λ is 0.5, balancing the contribution of the hard target loss (ground truth) and the soft target loss (teacher's predictions). With a batch size B of 8, the model processes 8 samples per iteration and it is set to perform binary segmentation with two classes C, representing the target leaf area and background in the segmentation task.

4.2. Ablation Study

This ablation study explores the effect of adding an attention module by comparing the models with and without it. The results show that adding the attention layer significantly improves performance, making it a valuable enhancement. Table 4 presents a comparative analysis of three models: student, teacher, and AKDUNet, evaluated on three key metrics such as F1 score, IoU score, and Dice coefficient. The models were evaluated using 5-fold cross-validation, ensuring that the evaluation is robust and not subject to over-fitting. The evaluation is conducted under two conditions, without the attention layer and with the attention layer. The student model performs poorly on all metrics, with an F1 score of 66.79%, an IoU score of 50.16%, and a loss of Dice coefficient of 0.1319.

At baseline (without attention), teacher model P^t outperforms the student model, achieving an F1 score of 90.06% and an IoU score of 81.98%, indicating its strong performance in segmentation tasks. However, the AKDUNet model shows

Tab. 3. Hyperparameters.

Parameters	Value
Learning rate η	0.0001
Optimizer	Adam
Temperature factor τ	4
Weight factor λ	0.5
Batch size B	8
Number of classes C	2

Tab. 4. Comparison of the models' performance with and without the attention layer.

Model	F1 score	IoU score	Dice coefficient loss
Without attention layer			
Student model	66.79%	50.16%	0.1319
Teacher model	90.06%	81.98%	0.0815
AKDUNet model	91.79%	85.82%	0.0414
With attention layer			
Student model	85.86%	75.22%	0.0773
Teacher model	94.69%	89.94%	0.0699
AKDUNet model	96.46%	93.16%	0.0227

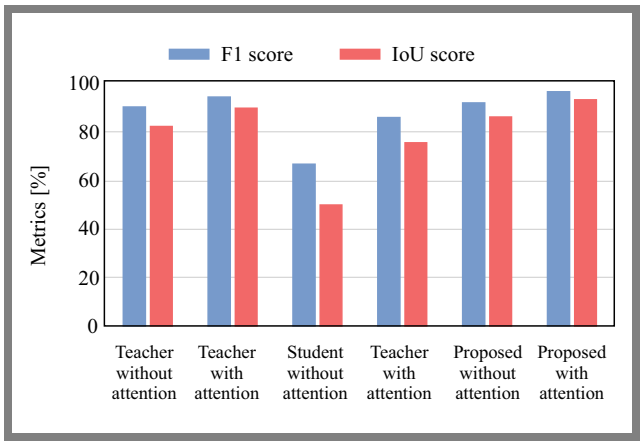

Fig. 4. Comparison of F1 and IoU scores for models with and without the attention layer.

a notable improvement with an F1 score of 91.79% and an IoU score of 85.82%, along with the lowest Dice coefficient loss of 0.0414, suggesting its superior ability to minimize coefficient losses.

Subsequently, an attention module is integrated into all the above-mentioned models and is evaluated for its effectiveness. The inclusion of the attention layer results in significant improvements in all three models. The AKDUNet model demonstrates the most significant performance improvement, achieving an F1 score of 96.46%, an IoU score of 93.16%, and a reduced Dice coefficient loss of 0.0227. The teacher model P^t also benefits from the addition of an attention layer, with an F1 score of 94.69%, an IoU score of 89.94%, and a decrease in the loss of Dice coefficient to 0.0699. Although student P^s model remains the least performing, it still shows improvements, with an F1 score of 85.86%, an IoU score of 75.22% and a Dice coefficient loss of 0.0773. These results show that incorporating an attention layer consistently improves segmentation performance across all models, as illustrated in Fig. 4.

Figure 5, which shows the Dice coefficient loss with and without the attention layer, highlights a significant reduction in prediction error when the attention layer is included. The teacher model P^t shows an improvement, with its loss of the Dice coefficient reducing from 0.0815 to 0.0699, indicating that attention helps minimize errors even in more complex models.

The student model P^s, while showing an improvement by reducing the loss of Dice coefficient from 0.1319 to 0.0773, still has the highest Dice coefficient loss among all models, highlighting its relative difficulty in minimizing prediction errors compared to the two remaining models. The AKDUNet model achieves the lowest loss of Dice coefficients in both conditions, decreasing from 0.0414 (without attention) to 0.0227 (with attention), showcasing its superior efficiency in reducing the error rate. Overall, the graph confirms that adding the attention layer contributes to a significant reduction in Dice coefficient loss, particularly for the proposed model, which achieves the lowest error across all conditions.

Tab. 5. Comparison of the proposed method with other methods using the CWFID data set.

Method	F1 score [%]	IoU score [%]	Loss
UNet++ [21]	77.56	63.37	0.2274
Inception UNet [22]	62.31	45.26	0.6211
SDUNet [15]	83.97	72.39	0.1650
INSCA UNet [16]	94.34	89.30	0.0604
VGG16 UNet [23]	90.70	82.99	0.0954
VGG19 UNet [24]	94.19	85.16	0.0634
ResNet UNet [25]	60.72	75.50	0.1730
SegFormer [18]	95.77	91.85	0.0456
Teacher UNet	94.69	89.94	0.0699
Student UNet	85.86	75.22	0.0773
AKDUNet model	96.46	93.16	0.0227

4.3. Comparison with Other Methods

The effectiveness of the proposed AKDUNet model in segmenting leaf regions from agricultural images was evaluated using two benchmark datasets, namely CWFID and Sunflower. The proposed method was benchmarked against various advanced UNet architectures, including UNet++, Inception UNet, SDUNet, and INSCA UNet, SegFormer, and pre-trained UNet models such as VGG16 UNet, VGG19 UNet, and ResNet UNet. As shown in Tab. 5, which presents results for the CWFID dataset, AKDUNet achieves a remarkable F1 score of 96.46%, IoU score of 94.85%, and the lowest loss value of 0.0208, outperforming all other compared methods.

These results indicate that AKDUNet provides highly accurate and consistent segmentation of leaf regions, which is crucial for downstream tasks such as disease detection, leaf counting, and plant phenotyping. While transformer-based models such as SegFormer also show strong performance with an F1 score of 95.77% and IoU of 91.85%, the superior metrics of AKDUNet highlight the benefits of integrating convolutional feature learning with attention-guided knowledge distillation, enabling the model to capture fine-grained leaf boundaries and spatial details more effectively.

Models such as INSCA UNet with an F1 score of 94.34% and VGG19 UNet with 94.19%, also demonstrate competitive

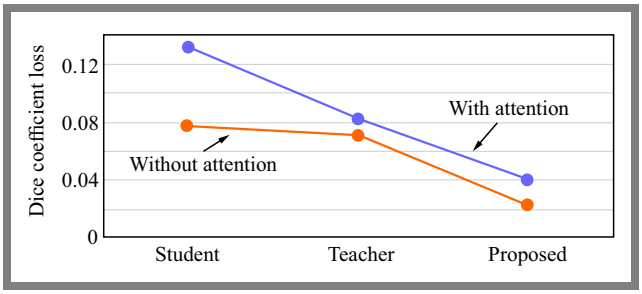

Fig. 5. Dice coefficient loss for models with and without the attention layer.

Tab. 6. Comparison of the proposed method with other methods using the CWFID data set.

Method	F1 score [%]	IoU score [%]	Loss
UNet++ [21]	70.04	69.38	0.0890
Inception UNet [22]	94.31	84.99	0.0040
SDUNet [15]	66.75	50.09	0.0220
INSCA UNet [16]	62.94	46.07	0.0238
VGG16 UNet [23]	90.90	82.35	0.0127
VGG19 UNet [24]	94.78	87.37	0.0064
ResNet UNet [25]	91.00	83.62	0.0152
SegFormer [18]	93.66	82.87	0.0042
Teacher UNet	90.44	77.99	0.0079
Student UNet	85.79	68.70	0.0119
AKDUNet model	95.16	88.17	0.0037

performance, yet were surpassed by AKDUNet in terms of both IoU and loss. This indicates that while attention mechanisms and deep CNN backbones aid in segmentation, the interlaced attention and knowledge transfer mechanism of AKDUNet allows for a more robust and precise extraction of leaf regions. Traditional models like UNet++, Inception UNet, and ResNet UNet show comparatively lower performance, with F1 scores below 78%, suggesting that they may struggle to delineate leaf boundaries in complex backgrounds or overlapping structures.

With the Sunflower dataset, as shown in Tab. 6, a similar pattern emerges. AKDUNet achieves the best performance with an F1 score of 95.16%, an IoU score of 88.17%, and the lowest loss of 0.0037, demonstrating its generalizability across different types of leaf structures and lighting conditions. Competing models such as VGG19 UNet with an F1 score of 94.78% and Inception UNet with 94.31% deliver strong results, but fall short in overall consistency and overlap accuracy compared to AKDUNet.

Interestingly, SegFormer, while highly effective on the CWFID dataset, shows a drop in performance here with an F1 score of 93.66% and IoU of 82.87%, suggesting that AKDUNet's convolutional attention hybrid design is more robust for diverse leaf morphology and field conditions. Moreover, the performance gap between the teacher UNet and the student UNet illustrates the impact of knowledge distillation, whereas AKDUNet significantly exceeds both, validating the advantage of its attention-enhanced distillation strategy.

Analysis performed with the use of both CWFID and Sunflower datasets clearly demonstrates the effectiveness and generalizability of the proposed AKDUNet model for leaf region segmentation in agricultural images. AKDUNet consistently outperforms all other benchmarked methods in terms of F1 score, IoU, and loss, indicating its ability to accurately extract leaf regions with minimal prediction error. Strong performance in both datasets, despite differences in leaf types, image complexity, and lighting conditions, highlights the robustness of the model's architecture.

4.4. Model Complexity Comparison

To assess computational complexity, the AKDUNet model uses floating point operations per second (FLOPs), which consists of number of addition, subtraction, multiplication, and division operations involved during the training process, and the formula to calculate FLOPs is shown in Eq. (17). In addition, the number of parameters for each method is also considered. To ensure fairness in comparison, all methods including the proposed model are evaluated with the same input size of 128 × 128 × 3.

For a Conv2D layer:

$$FLOPs_S = 2 \times H_{out} \times W_{out} \times C_{in} \times K_H \times K_W \times C_{out} \ , \quad (17)$$

where H_{out}, W_{out} – output height and width of the image, C_{in} – input channels of the image, K_H, K_W – kernel height and width of the filter, C_{out} – output channels.

Table 7 presents a comparative analysis of various segmentation models in terms of the number of parameters and computational complexity (FLOPs), while Fig. 6 offers a visual comparison based on parameter count (in millions) and FLOPs (in giga FLOPs).

The proposed AKDUNet model stands out with an exceptionally low parameter count of just 0.013 M and a minimal 0.17 GFLOPs, demonstrating remarkable efficiency in both memory usage and computational cost. Despite its lightweight architecture, AKDUNet achieves a superior segmentation accuracy of 93.16%, surpassing more complex models such as VGG16 UNet with 25.85 M parameters, 32.26 GFLOPs, and VGG19 UNet with 16.24 M parameters, 41.15 GFLOPs, which require significantly more resources. The reduced complexity enables faster execution and lower resource consumption, making it suitable for deployment in resource-constrained environments without compromising segmentation performance.

In comparison, models such as SDUNet with 0.498 M parameters and 5.22 GFLOPs as well as ResNet UNet with 24.29 M parameters and 3.12 GFLOPs offer a trade-off be-

Tab. 7. Comparison of model complexity and computational requirements.

Model	Parameters [M]	FLOPs [G]
UNet++	0.149	4.89
VGG16 UNet	25.85	32.26
VGG19 UNet	16.24	41.15
ResNet UNet	24.29	3.12
Inception UNet	0.177	2.90
SDUNet	0.498	5.22
SegFormer	0.124	5.2
Teacher UNet	2.354	2.67
Proposed AKDUNet	0.013	0.17

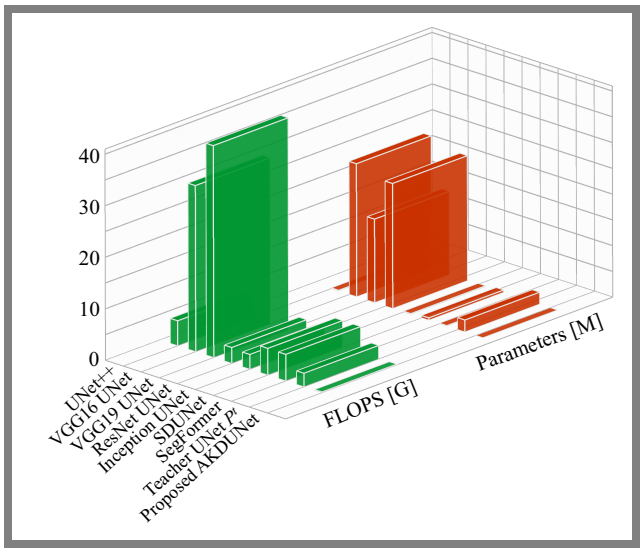

Fig. 6. Comparison of computational complexity with existing models.

tween complexity and accuracy, but fall short in both segmentation precision and computational efficiency. Similarly, the teacher UNet model, with 2.354 M parameters and 2.67 GFLOPs, delivers competitive results, but does not outperform AKDUNet in either metric. The SegFormer model, despite its lightweight nature with only 0.124 M parameters, incurs a relatively high computational cost of 5.2 GFLOPs due to its transformer-based architecture, limiting its usability for real-time deployment on edge devices.

The ability of AKDUNet to achieve superior segmentation accuracy with a significantly lower computational footprint highlights its effectiveness and efficiency, especially when both IoU score and resource demands are critical considerations.

4.5. Qualitative Analysis

Grad CAM visualizations were also generated for the student model P^s, the teacher model P^t, and the AKDUNet model to assess the effectiveness of the distillation process, as shown in Fig. 7.

The results revealed that while both the teacher model P^t and the student model P^s showed attention to relevant features, the AKDUNet model demonstrated superior focus on the critical regions of the input images. This suggests that, while the student model P^s achieves reasonable accuracy, it is not as interpretable or capable of making significant decisions as the teacher model P^t. Specifically, the grad CAM heat maps of the AKDUNet model closely aligned with those of the teacher model P^t, indicating that the student model P^s successfully learned to prioritize the most significant areas through the distillation process.

To evaluate the effectiveness of the proposed architecture, a qualitative analysis was conducted for all models, and the performance was compared for scenarios with and without knowledge distillation. As shown in Fig. 8, segmentation maps from the student model P^s trained from scratch without knowledge distillation lack clear separation between the

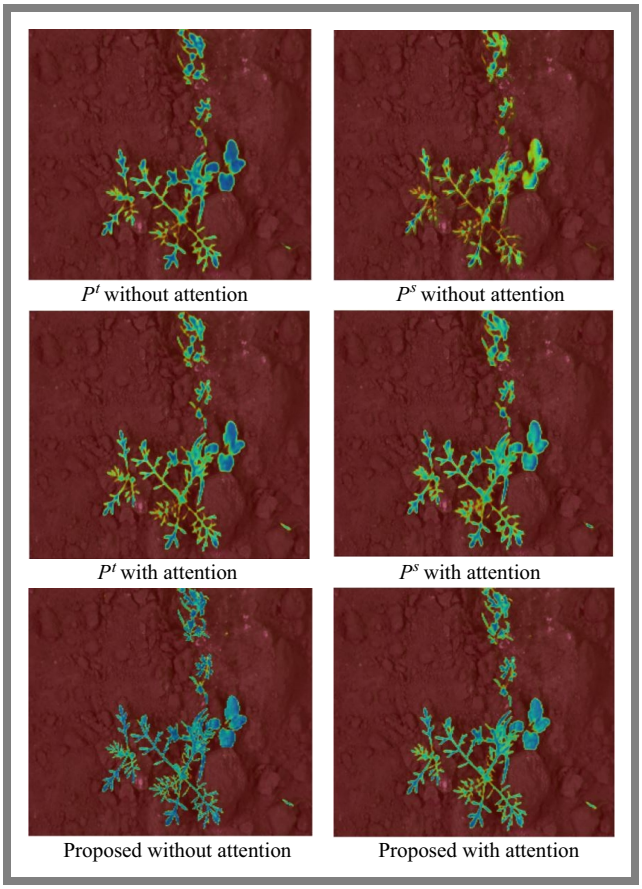

Fig. 7. Grad CAM visualizations of all models.

secondary leaflets and do not present sharp boundaries. This model produces an incomplete segmentation map with many leaf regions incorrectly classified as background.

In contrast, the segmentation map produced by the proposed model with an attention gated module is much clearer and more accurately aligns with the ground truth map, as shown in Fig. 8. This analysis demonstrates that combining knowledge distillation with an attention layer significantly enhances segmentation performance. This improvement is made possible by the transfer of knowledge from the teacher model P^t, whose segmented output outperforms the student model's output image.

Figures 9, 10 show the original images, ground truth, and segmentation outputs for a variety of models, including pretrained UNet architectures such as VGG16 UNet, VGG19 UNet, and ResNet UNet, along with advanced models like UNet++, Inception UNet, SDUNet, and INCSA UNet. By integrating knowledge distillation techniques and attention mechanisms, AKDUNet achieves highly accurate segmentation, even in challenging areas such as secondary leaflets and finer details of leaf structures.

The attention module enhances the model's ability to focus on critical regions, while KD effectively transfers knowledge from the teacher model, improving the student model's generalization and segmentation accuracy. The qualitative results clearly demonstrate that AKDUNet significantly outperforms existing segmentation approaches. It delivers sharper and

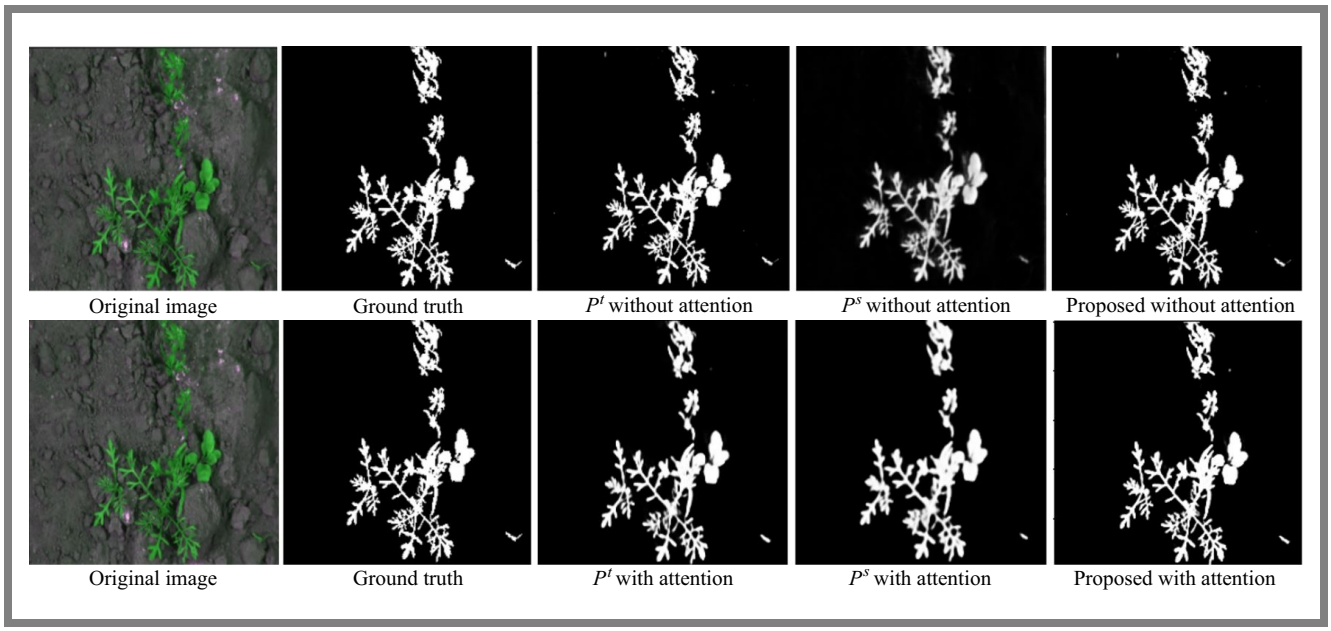

Fig. 8. Leaf area segmentation results with and without the attention module.

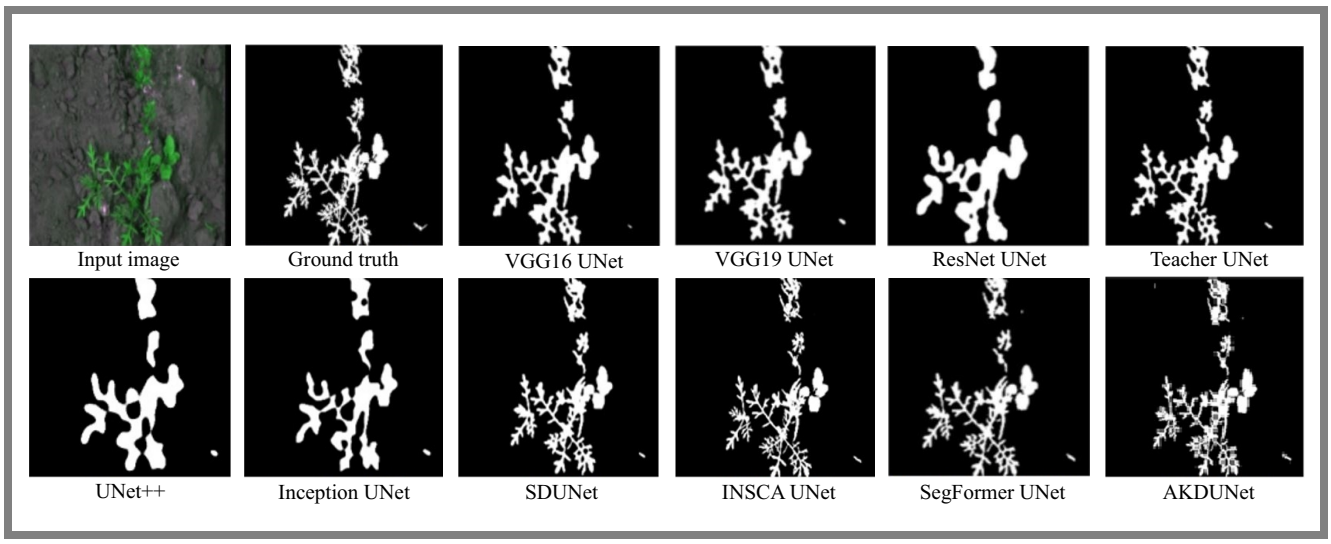

Fig. 9. Results of segmentation – area of the CWFID carrot leaf.

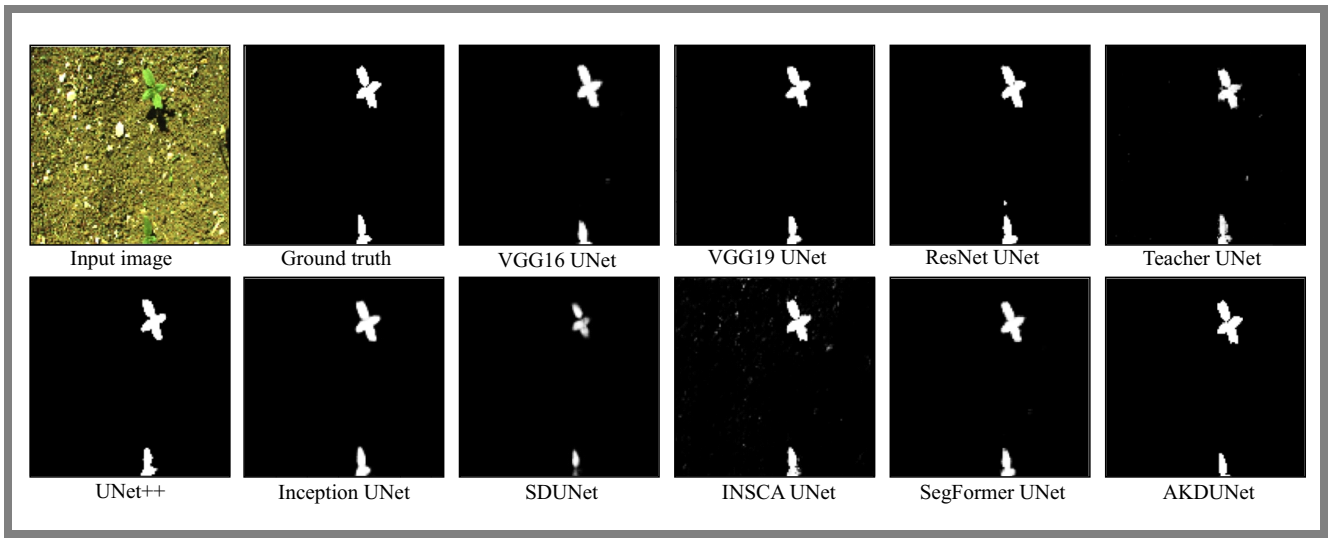

Fig. 10. Results of segmentation – sunflower leaf area.

more precise boundaries, particularly in complex areas, when compared to UNet, VGG16 UNet, VGG19 UNet, ResNet UNet, UNet++, Inception UNet, SDUNet, and INCSA UNet.

5. Conclusions

This study demonstrates the performance of the proposed AKDUNet model while segmenting the leaf region for precision agriculture. Across the CWFID carrot leaf and Sunflower data sets from CWFID, AKDUNet consistently outperformed several state-of-the-art architectures, including SegFormer, UNet++, Inception UNet, SDUNet, INSCA UNet, and pre-trained UNet variants with VGG16, VGG19 and ResNet backbones. The model achieved higher F1 scores and IoU values while maintaining minimal loss, validating its robustness in handling variations in leaf morphology and imaging conditions.

Its significantly lower computational complexity is a key advantage of AKDUNet. Although only 0.013 million parameters and 0.17 GFLOPs were used, the segmentation results were comparable to or better than those of much larger models. Such a high degree of efficiency positions AKDUNet as a strong candidate for real-time deployment on resource-constrained devices such as drones, mobile phones, or edge computing platforms used in agricultural monitoring systems.

The model's architecture, which integrates attention mechanisms and knowledge distillation from a deeper teacher network, enables it to focus on critical regions and segment fine details with a high degree of accuracy. Its ability to generalize across datasets suggests that it can adapt well to diverse plant species and imaging conditions without the need for extensive retraining.

However, there are certain limitations to this work. First, the evaluation was conducted on a limited number of datasets, primarily focused on leaf structures from specific crops. While AKDUNet showed strong generalization across these, its performance on more complex agricultural scenes with overlapping plant parts, occlusions, or mixed crop types has not yet been tested. Second, although the model is lightweight, it still relies on supervised learning and labeled data, which can be costly and time-consuming to obtain at scale. Finally, the knowledge distillation strategy, while effective, may require careful tuning of teacher-student dynamics to ensure stable training across various domains.

Acknowledgments

The datasets used in the study are publicly available in the following repository: https://github.com/cwfid/dataset.

References

[1] A. Walter and U. Schurr, "The Modular Character of Growth in Nicotiana Tabacum Plants under Steady-state Nutrition", *Journal of Experimental Botany*, vol. 50, pp. 1169–1177, 1999 (https://doi.org/10.1093/jxb/50.336.1169).

[2] A.L. Chandra, S.V. Desai, W. Guo, and V.N. Balasubramanian, "Computer Vision with Deep Learning for Plant Phenotyping in Agriculture: A Survey", *ArXiv*, 2020 (https://doi.org/10.48550/arXiv.2006.11391).

[3] S. Haug and J. Ostermann, "A Crop/weed Field Image Dataset for the Evaluation of Computer Vision Based Precision Agriculture Tasks", *Proc. of Computer Vision – ECCV 2014*, pp. 105–116, 2014 (https://doi.org/10.1007/978-3-319-16220-1_8).

[4] R. Liu *et al.*, "TransKD: Transformer Knowledge Distillation for Efficient Semantic Segmentation", *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, 2024 (https://doi.org/10.1109/TITS.2024.3455416).

[5] J. Zhang, Q. Liang, and Y. Shi, "KD-SCFNet: Towards More Accurate and Efficient Salient Object Detection via Knowledge Distillation", *ArXiv*, 2022 (https://doi.org/10.48550/arXiv.2208.02178).

[6] S. An, Q. Liao, Z. Lu, and J.-H. Xue, "Efficient Semantic Segmentation via Self-attention and Self-distillation", *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, pp. 15256–15266, 2022 (https://doi.org/10.1109/TITS.2021.3139001).

[7] T. Zhang *et al.*, "Efficient RGB-T Tracking via Cross-modality Distillation", *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Vancouver, Canada, 2023 (https://doi.org/10.1109/CVPR52729.2023.00523).

[8] X. Hu, K. Yang, L. Fei, and K. Wang, "ACNet: Attention Based Network to Exploit Complementary Features for RGBD Semantic Segmentation", *IEEE International Conference on Image Processing (ICIP)*, Taipei, Taiwan, 2019 (https://doi.org/10.1109/ICIP.2019.8803025).

[9] W. Zhou, J. Yuan, J. Lei, and T. Luo, "TSNet: Three-stream Self-attention Network for RGB-D Indoor Semantic Segmentation", *IEEE Intelligent Systems*, vol. 36, pp. 73–78, 2021 (https://doi.org/10.1109/MIS.2020.2999462).

[10] W. Zhou, E. Yang, J. Lei, and L. Yu, "FRNet: Feature Reconstruction Network for RGB-D Indoor Scene Parsing", *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, pp. 677–687, 2022 (https://doi.org/10.1109/JSTSP.2022.3174338).

[11] J. Zhou *et al.*, "ESA-Net: A Network with Efficient Spatial Attention for Smoky Vehicle Detection", 2021 *IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, Glasgow, UK, 2021 (https://doi.org/10.1109/I2MTC50364.2021.9460078).

[12] P. Chavan, P.P. Chavan, and A. Chavan, "Hybrid Architecture for Crop Detection and Leaf Disease Detection with Improved U-Net Segmentation Model and Image Processing", *Crop Protection*, vol. 190, art. no. 107117, 2025 (https://doi.org/10.1016/j.cropro.2025.107117).

[13] M.K. Surehli, N. Aggarwal, G. Joshi, and H. Nayyar, "Semantic Segmentation of Plant Structures with Deep Learning and Channel-wise Attention Mechanism", *JTIT*, vol. 99, pp. 56–66, 2025 (https://doi.org/10.26636/jtit.2025.1.1853).

[14] A. Das *et al.*, "Deep Learning-based Classification, Detection, and Segmentation of Tomato Leaf Diseases: A State-of-the-art Review", *Artificial Intelligence in Agriculture*, vol. 15, pp. 192–220, 2025 (https://doi.org/10.1016/j.aiia.2025.02.006).

[15] C. Guo *et al.*, "SD-UNet: A Structured Dropout U-Net for Retinal Vessel Segmentation", *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, Athens, Greece, 2019 (https://doi.org/10.1109/BIBE.2019.00085).

[16] I. Delibasoglu, "INCSA-UNET: Spatial Attention Inception UNET for Aerial Images Segmentation", *Computing and Informatics*, vol. 40, pp. 1244–1262, 2022 (https://doi.org/10.31577/cai_2021_6_1244).

[17] M. Agarwal, S.K. Gupta, and K.K. Biswas, "Plant Leaf Disease Segmentation Using Compressed UNet Architecture", *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pp. 9–14, 2021 (https://doi.org/10.1007/978-3-030-75015-2_2).

[18] E. Xie *et al.*, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers", *ArXiv*, 2021 (https://doi.org/10.48550/arXiv.2105.15203).

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, Munich, Germany, 2015 (https://doi.org/10.1007/978-3-319-24574-4_28).

[20] M. Fawakherji *et al.*, "Multi-spectral Image Synthesis for Crop/weed Segmentation in Precision Farming", *Robotics and Autonomous Systems*, vol. 146, art. no. 103861, 2021 (https://doi.org/10.1016/j.robot.2021.103861).

[21] Z. Zhou, M.R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A Nested UNet Architecture for Medical Image Segmentation", in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Lecture Notes in Computer Science*, Springer Cham, pp. 3–11, 2018 (https://doi.org/10.1007/978-3-030-00889-5_1).

[22] N.S. Punn, S. Agarwal, "Inception U-Net Architecture for Semantic Segmentation to Identify Nuclei in Microscopy Cell Images", *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 1, pp. 1–15, 2020 (https://doi.org/10.1145/3376922).

[23] Y. Miao *et al.*, "CT Image Segmentation of Foxtail Millet Seeds Based on Semantic Segmentation Model VGG16-UNet", *Plant Methods*, vol. 20, art. no. 169, 2024 (https://doi.org/10.1186/s13007-024-01288-y).

[24] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition", *ArXiv*, 2014 (https://doi.org/10.48550/arXiv.1409.1556).

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, USA, 2016 (https://doi.org/10.1109/CVPR.2016.90).

A. Shamim Banu, Research Scholar
Department of ECE
https://orcid.org/0009-0002-2329-4235
E-mail: shamiece@gmail.com
National Institute of Technology, Tiruchirappalli, Tamilnadu, India
https://www.nitt.edu
Government Polytechnic College, Tiruchirappalli, Tamilnadu, India
http://gptctrichy.com

S. Deivalakshmi, Ph.D., Associate Professor
Department of ECE
https://orcid.org/0000-0002-7019-9807
E-mail: deiva@nitt.edu
National Institute of Technology, Tiruchirappalli, Tamilnadu, India
https://www.nitt.edu

A Convex Optimization-based Approach for Sidelobe Level Suppression and Null Control in Antenna Arrays by Displacing a Minimum Number of Elements

Magdy A. Abdelhay

Al-Farqadein University College, Basrah, Iraq

https://doi.org/10.26636/jtit.2025.3.2198

Abstract — This paper introduces two methods for peak sidelobe level (PSLL) reduction and null steering in the pattern of linear arrays using position control. While most research on this topic uses stochastic optimization techniques, here convex optimization and the off-grid compressive sensing framework were used to accomplish the required goals. For the first method, the problem of minimizing the PSLL and forming prescribed nulls in the pattern of linear arrays by controlling the elements' positions is cast as a convex optimization problem with the help of first-order Taylor approximation. For the second method, the goals are achieved by perturbing the locations of as few array elements as possible. Towards this end, the problem of forming prescribed nulls in the pattern of non-uniformly spaced linear arrays for a predefined PSLL by elements' position control is formulated as a sparse recovery problem within the off-grid compressive sensing framework. Simulations were performed to evaluate the efficacy of the proposed methods, and the results were compared to results obtained using stochastic optimization techniques.

Keywords — compressive sensing, convex optimization, mechanically adaptive arrays

1. Introduction

In phased array antenna, the radiation pattern can be altered so that the radiation pattern adds up to boost the radiation in the wanted direction while canceling out the radiation in the undesired directions. Numerous algorithms have been studied to create radiation pattern nulls by changing the excitation of elements' amplitude only [1]–[3], phase only [4], [5], amplitude and phase (complex) [6]–[8], or inter-element spacing [9]–[11].

Compressed sensing (CS) is a fairly recent signal processing method to sample and reconstruct signals efficiently by obtaining solutions of underdetermined linear systems [12]. The potential to defy established wisdom in data acquisition based on Shannon's theory [13] and permit the recovery of specific signals from much less observations than standard approaches has received much attention [14], [15].

The cornerstone of CS-based approaches is that a lot of physical variables, both intrinsically or extrinsically sparse, can be portrayed using just a few of nonzero expansion coefficients, given the appropriate expansion bases. The basic objective of CS methods is to determine an approximation of the solution $\mathbf{x}$ to the linear system $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{x}$ must have as few nonzero elements as possible [12].

The evolution of CS has centered on signals having sparse representation in finite discrete dictionaries. The majority of signals we encounter in the actual world, particularly those used in remote sensing, sonar and radar, are characterized by continuous parameters. A discretization process is used to create a finite collection of grid points from the continuous parameter space in order to apply CS theory. When the true signal is not precisely supported on the grid points, performance loss occurs for the traditional CS methods, also known as on-grid CS. This problem is referred to as the basis mismatch problem [16].

Off-grid CS techniques aim to address the basis mismatch issues without trying to solve the problem by using denser discretization, since CS theory suggests that utilizing a finer grid could not improve performance and even might increase the coherence of the dictionary, which contradicts the restricted isometry property necessary for guaranteeing accurate estimation of sparse recovery problems [17], [18].

Off-grid CS framework has been used to synthesize uniformly weighted concentric ring arrays in [19]. A method based on off-grid CS for the synthesis of planar sparse arrays was proposed in [20]. In [21], an alternating algorithm to synthesis planar sparse antenna arrays with complex-excitation and reconfigurable pattern was proposed.

Most research on the topic of optimizing antenna arrays by position-only control employs stochastic optimization techniques [10], [22], [23]. Stochastic optimization techniques suffer from several limitations, including high computational cost, particularly for large array sizes. Additionally, there is no guarantee that the obtained solution is the optimal one, as it may be trapped in a local minimum. Another

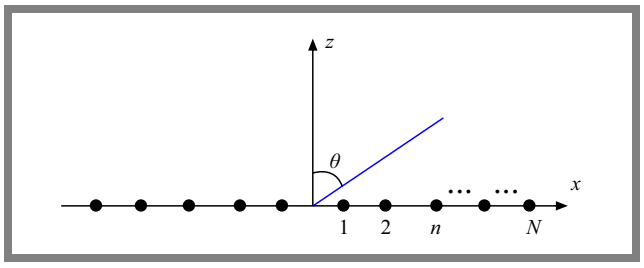

Fig. 1. Geometry of a $2N$ element linear array along the x-axis.

drawback is the inconsistent findings achieved during each run, which necessitates numerous independent runs. Some of the drawbacks of stochastic optimization methods can be overcome by using convex optimization.

In this paper, two methods for peak sidelobe level (PSLL) reduction and null steering in uniformly weighted linear arrays by controlling the position of the array elements are proposed. For the first method, the synthesis problem is cast as a convex optimization problem with the help of first-order Taylor approximation. Here, all the array elements' positions are perturbed to reduce the PSLL and impose prescribed nulls with a predetermined upper bound on the null depth in the radiation pattern. The problem is solved iteratively with a small position perturbations every iteration to minimize the approximation error.

For the second method, the problem is to impose prescribed nulls with a specified upper bound on the null depth while perturbing the positions of as few array elements as possible to achieve a predefined PSLL. The problem is formed as a sparse recovery problem using the off-grid compressive sensing framework. Here, instead of minimizing the ℓ_0 norm of the weight vector, the ℓ_0 norm of the position perturbation vector is minimized. An algorithm based on iterative reweighted ℓ_1 norm minimization of the position perturbation vector is proposed, where the position perturbations are kept small per iteration to control for the approximation error.

The remainder of the paper is structured as follows. Section 2 introduces the problem formulation. Method 1 is presented in Section 3. Method 2 is detailed in Section 4. Section 5 presents the simulation results. Finally, Section 6 draws the conclusions.

2. Problem Formulation

Figure 1 shows a $2N$ element linear array placed symmetrically along the x–axis. In the $x-z$ plane, the array factor is given by:

$$F(\theta) = 2\sum_{n=1}^{N} I_n\, e^{j\left(\frac{2\pi}{\lambda} x_n \sin\theta + \phi_n\right)}\,, \tag{1}$$

where λ is the wavelength, x_n is the position of the n-th element and I_n and ϕ_n are the excitation and the phase of the n-th element, respectively. For a uniformly excited array, i.e., $I_n = 1$ and $\phi_n = 0$, Eq. (1) can be written as:

$$F(\theta) = 2\sum_{n=1}^{N} \cos\left[\frac{2\pi}{\lambda} x_n \sin\theta\right]\,. \tag{2}$$

In this work, the array synthesis problem is modeled as an off-grid CS problem. Suppose that the n-th element position x'_n is not located on the grid points, but is situated at an unknown displacement from the closest grid point x_n. To find the element displacement from the nearest grid point, we present position perturbation to x_n.

Let $a(x_n) = 2\cos\left[\frac{2\pi}{\lambda} x_n \sin\theta\right]$. Using first order Taylor expansion:

$$a(x'_n) \approx a(x_n) + \delta_n \frac{\partial a(x)}{\partial x}\bigg|_{x=x_n}\,, \tag{3}$$

where δ_n is the position perturbation variable for the n-th element and $|\delta_n| \leqslant \Delta d_x/2$. Initially we start with the uniform equally spaced array with interelement spacing Δd_x, then the array elements can have a controlled displacement from their initial locations using the position perturbation variables. The use of $|\delta_n| \leqslant \Delta d_x/2$ ensures that successive elements do not overlap since they can only move by half the interelement spacing in either direction.

The array radiation pattern with the elements' position perturbations may be represented using the first-order Taylor approximation as:

$$F(\theta) \approx \sum_{n=1}^{N} \left[a(x_n) + \delta_n \frac{\partial a(x)}{\partial x}\bigg|_{x=x_n}\right]\,, \tag{4}$$

Equation (4) may be represented in matrix form by sampling the radiation pattern as:

$$\mathbf{F} = (\mathbf{A} + \mathbf{A}_x \Lambda_\delta)\mathbf{1}\,, \tag{5}$$

where $\mathbf{F} = [F(\theta_1), F(\theta_2), \ldots, F(\theta_J)]^T$ is a vector containing the samples of $F(\theta)$ at J directions θ_j, $j = 1, 2, \ldots, J$. $\mathbf{A} = [\boldsymbol{a}(1), \ldots, \boldsymbol{a}(J)]^T$ and $\boldsymbol{a}(j) = [a(x_1), \ldots, a(x_N)]^T$ with $\theta = \theta_j$. $\mathbf{A}_x$ is the partial derivative of $\mathbf{A}$ with respect to x. $\mathbf{\Lambda}_\delta = \text{diag}(\boldsymbol{\delta})$, where $\boldsymbol{\delta}$ is a vector of position perturbations $\boldsymbol{\delta} = [\delta_1, \ldots, \delta_N]^T$. $\mathbf{1} \in \mathbb{R}^N$ is the one vector.

3. Method 1

In this section, we are interested in reducing the PSLL and impose nulls with a predefined upper limit on the null depth in the array's radiation pattern by position control of all the array elements. Towards this end, the array synthesis problem may be expressed as:

$$\min_{\delta} \tau_s \tag{6a}$$

$$\text{subject to } |(\mathbf{A} + \mathbf{A}_x \Lambda_\delta)\mathbf{1}| \leqslant \tau_s, \quad \theta \in \Omega^{sl} \tag{6b}$$

$$|(\mathbf{A} + \mathbf{A}_x \Lambda_\delta)\mathbf{1}| \leqslant \tau_n, \quad \theta \in \Omega^{nl} \tag{6c}$$

$$|\delta_n| \leqslant \Delta d_x/2, \quad n = 1, 2, \ldots, N\,, \tag{6d}$$

where τ_s is a slack variable that represents an upper bound on the array response in the sidelobe region, τ_n is an upper bound on the null depth, Ω^{sl} is the sidelobe region, and Ω^{nl} is the null region (directions).

Since the suggested method relies on the first-order Taylor approximation in Eq. (3), the approximation error needs to

be minimal to ensure the model's accuracy. It is obvious that the inaccuracy will increase as the values of the position perturbations, δ_n, rise. The modeling error can be reduced by reducing Δd_x in Eq. (6)-d to a smaller value $\Delta d'_x < \Delta d_x$, but this will limit the degrees of freedom provided to the algorithm and might reach a solution with a high PSLL.

Here, an iterative algorithm is proposed to mitigate this problem by restricting the value of δ_n in each iteration to $|\delta_n^k| \leqslant \Delta d'_x/2$, where δ_n^k is the value of δ_n in iteration k and $\Delta d'_x < \Delta d_x$. By doing so, we will be able to improve the model's accuracy without facing the aforementioned issues.

Initially, a uniformly spaced array is considered, i.e. the position perturbations δ_n^0, $n = 1, 2, \ldots, N$ are set to zero, $\mathbf{A}^0$ and $\mathbf{A}_x^0$ are calculated accordingly for the sidelobe region and the null directions. The optimization problem at the k-th iteration may be expressed as:

$$\min_{\delta^k} \tau_s \tag{7a}$$

$$\text{subject to } |(\mathbf{A}^{k-1} + \mathbf{A}_x^{k-1}\Lambda_\delta^k)\mathbf{1}| \leqslant \tau_s, \quad \theta \in \Omega^{sl} \tag{7b}$$

$$|(\mathbf{A}^{k-1} + \mathbf{A}_x^{k-1}\Lambda_\delta^k)\mathbf{1}| \leqslant \tau_n, \quad \theta \in \Omega^{nl} \tag{7c}$$

$$|\delta_n| \leqslant \Delta d'_x/2, \quad n = 1, 2, \ldots, N\,. \tag{7d}$$

The optimization problem in Eq. (7) is a convex optimization problem and can be solved using off-the-shelf packages, such as CVX [24]. After solving the optimization problem in Eq. (7), the array elements' positions are adjusted in accordance with their perturbation values:

$$x_n^k = x_n^{k-1} + \delta_n^k, \quad n = 1, \ldots, N\,, \tag{8}$$

where x_n^{k-1} is the position of the n-th element at the past iteration $k-1$. Finally, $\mathbf{A}^k$ and $\mathbf{A}_x^k$ are updated according to the new element positions, and the optimization problem in Eq. (7) is solved again for another iteration. The algorithm continues until the maximum number of iterations is reached. The maximum number of iterations is set experimentally to 10. This algorithm is referred to as method 1 for the remaining of the paper.

4. Method 2

To minimize the amount of elements that needs to be perturbed from their original positions under a predefined PSLL and null depth, the optimization problem may be formulated as:

$$\min_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_0 \tag{9a}$$

$$\text{subject to } |(\mathbf{A} + \mathbf{A}_x\Lambda_\delta)\mathbf{1}| \leqslant \tau_s, \quad \theta \in \Omega^{sl} \tag{9b}$$

$$|(\mathbf{A} + \mathbf{A}_x\Lambda_\delta)\mathbf{1}| \leqslant \tau_n, \quad \theta \in \Omega^{nl} \tag{9c}$$

$$|\delta_n| \leqslant \Delta d_x/2, \quad n = 1, 2, \ldots, N\,, \tag{9d}$$

where $\|\cdot\|_0$ is the ℓ_0 norm, which is the number of the non-zero entries of its argument. The optimization problem in Eq. (9) is an NP-hard optimization problem due to the non-convex objective function. To achieve a convex optimization problem, the convex and sparsity-promoting ℓ_1 norm can be used in place of the ℓ_0 norm:

$$\min_{\boldsymbol{\delta}} \|\boldsymbol{\delta}\|_1 \tag{10a}$$

$$\text{subject to } |(\mathbf{A} + \mathbf{A}_x\Lambda_\delta)\mathbf{1}| \leqslant \tau_s, \quad \theta \in \Omega^{sl} \tag{10b}$$

$$|(\mathbf{A} + \mathbf{A}_x\Lambda_\delta)\mathbf{1}| \leqslant \tau_n, \quad \theta \in \Omega^{nl} \tag{10c}$$

$$|\delta_n| \leqslant \Delta d_x/2, \quad n = 1, 2, \ldots, N\,, \tag{10d}$$

where $\|\cdot\|_1$ is the ℓ_1 norm, which is the sum of the absolute values of its argument. That is $\|\boldsymbol{\delta}\|_1 = \sum_{n=1}^{N} |\delta_n|$.

To lower the approximation error and reduce the number of perturbed element, an algorithm based on the iterative reweighted ℓ_1 norm minimization is proposed [25]. Initially, a uniformly spaced array is considered with zero position perturbations $\delta_n^0 = 0$, $n = 1, 2, \ldots, N$. The matrices $\mathbf{A}^0$ and $\mathbf{A}_x^0$ are calculated accordingly for the sidelobe region and the null directions. The optimization problem at the k-th iteration can be expressed as:

$$\min_{\boldsymbol{\delta}^k} \sum_{n=1}^{N} |\psi_n^k \delta_n^k| \tag{11a}$$

$$\text{subject to } |(\mathbf{A}^{k-1} + \mathbf{A}_x^{k-1}\Lambda_\delta^k)\mathbf{1}| \leqslant \tau_s, \quad \theta \in \Omega^{sl} \tag{11b}$$

$$|(\mathbf{A}^{k-1} + \mathbf{A}_x^{k-1}\Lambda_\delta^k)\mathbf{1}| \leqslant \tau_n, \quad \theta \in \Omega^{nl}\,, \tag{11c}$$

with δ_n^k being the n-th element of $\boldsymbol{\delta}$ at iteration k. $\psi_n^k = 1/(|\delta_n^{k-1}| + \xi)$, where δ_n^{k-1} is the value of δ_n at iteration $k-1$. ξ is a small positive number utilized to retain numerical stability. In this work, ξ is set to 0.0001.

With this relation between ψ_n^k and δ_n^{k-1}, small elements in $\boldsymbol{\delta}$ will be penalized because they are multiplied by a large value ψ_n^k. This will result in even smaller values for the small entries in $\boldsymbol{\delta}$ in the following iteration and boosting the sparsity of the solution [25]. At the first iteration, $\psi_n^1, n = 1, 2, \ldots, N$ are set to one.

After solving the optimization problem in Eq. (11) using CVX, the values of the position perturbations are limited to $|\delta_n| \leqslant \Delta d'_x/2$, i.e., $\delta_n \in \left[-\frac{1}{2}\Delta d'_x,\ \frac{1}{2}\Delta d'_x\right]$. The final values of the position perturbations at iteration k are calculated using:

$$\tilde{\delta}_n^k = \begin{cases} \delta_n^k, & \text{if } \delta_n^k \in \left[-\frac{1}{2}\Delta d'_x,\ \frac{1}{2}\Delta d'_x\right] \\ -\frac{1}{2}\Delta d'_x, & \text{if } \delta_n^k < -\frac{1}{2}\Delta d'_x \\ \frac{1}{2}\Delta d'_x, & \text{otherwise}\,. \end{cases} \tag{12}$$

The array elements' positions are then updated in accordance with their position perturbation values:

$$x_n^k = x_n^{k-1} + \tilde{\delta}_n^k, \quad n = 1, \ldots, N\,, \tag{13}$$

where x_n^{k-1} is the position of the n-th element at the previous iteration $k-1$.

Finally, $\mathbf{A}^k$ and $\mathbf{A}_x^k$ are updated according to the new element positions for the sidelobe region and the null directions. Then, the optimization problem in Eq. (11) is solved again for another iteration. The algorithm continues until it reaches the maximum number of iterations or $\|\boldsymbol{\delta}^k\|_2 \leqslant \epsilon$, where ϵ is a tolerance parameter. Here, ϵ is set to be 0.0001 experimentally.$\|\cdot\|_2$ is the ℓ_2 norm. This indicates that there

Tab. 1. Geometry of the optimized 10-element array using method 1 (normalized with respect to $\lambda/2$).

n	Position	n	Position
1	±0.4880	4	±3.0000
2	±1.0625	5	±4.2082
3	±2.0642		

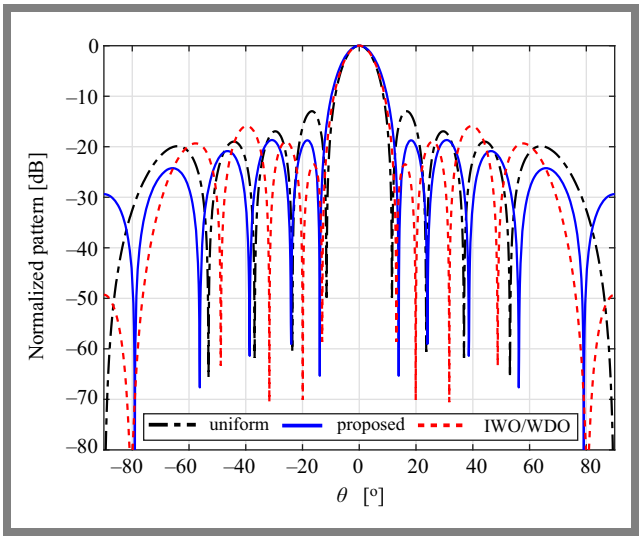

Fig. 2. Patterns of the uniform 10-element array, method 1 and the hybrid IWO/WDO from [10].

is no meaningful change in the positions of the array elements in the current iteration. The maximum number of iterations is set to 10. This algorithm is referred to as method 2 for the remaining of the paper.

5. Simulation Results

For the first example, consider the synthesis of a 10-element linear array (i.e. $N = 5$) with a minimum PSLL. The initial uniformly spaced array has an inter-element spacing of $\lambda/2$. This array was optimized using particle swarm optimization (PSO) in [22], comprehensive learning particle swarm optimizer (CLPSO) in [23], and hybrid invasive weed optimization and wind driven optimization (IWO/WDO) in [10]. The best obtained result was a PSLL of -15.9 dB for the normalized pattern using the hybrid IWO/WDO from [10]. Applying method 1 resulted in obtaining a PSLL of -18.67 dB for the normalized pattern compared to -15.9 dB for the hybrid IWO/WDO. The normalized patterns of the uniform array, method 1 and the hybrid IWO/WDO are shown in Fig. 2. The geometry of the optimized array is given in Tab. 1 with respect to $\lambda/2$.

The second example demonstrates the synthesis of a 28-element linear array ($N = 14$) with a minimum PSLL and three prescribed nulls at 30°, 32.5°, and 35° as in [10]. The initial uniformly spaced array has an inter-element spacing of $\lambda/2$. This array was optimized using particle PSO in [22], CLPSO in [23], and hybrid IWO/WDO in [10]. The hybrid IWO/WDO resulted in the best PSLL of -13.19 dB for the

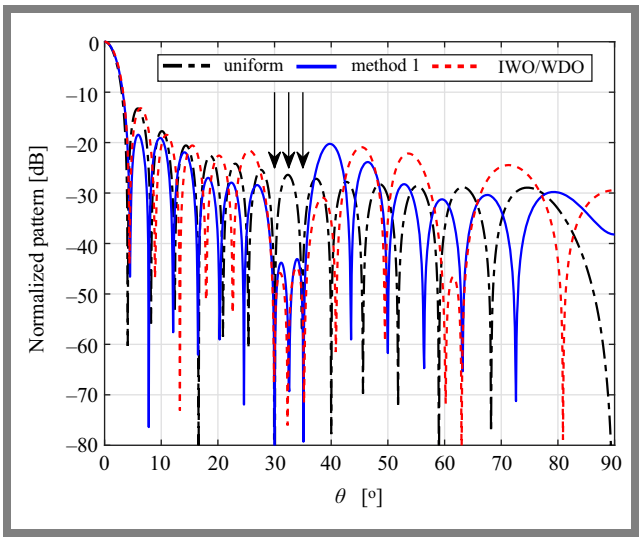

Fig. 3. Patterns of the uniform 28-element array, method 1 and the hybrid IWO/WDO from [10]. The directions of the nulls are indicated by the arrows.

normalized pattern. Applying method 1 resulted in a PSLL of -18.45 dB. The normalized patterns of method 1, the uniform array and the hybrid IWO/WDO are depicted in Fig. 3. Table 2 lists the element positions for the optimized array using method 1 with respect to $\lambda/2$.

Next, for the third example, we apply method 2 for the 28-element linear array ($N = 14$) with initial inter-element spacing of $\lambda/2$. We set the upper bound on the array response in the sidelobe region to that obtained using the hybrid IWO/WDO from [10]. The objective for method 2 is to achieve this PSLL and the three imposed nulls at 30°, 32.5°, and 35° by perturbing the positions of a minimum number of that array elements. Applying method 2 resulted in perturbing the positions of only 6 out of the total 28 array elements while satisfying all the constraints on the radiation pattern. The patterns of the uniform 28-element array, the hybrid IWO/WDO from [10] and the pattern of method 2 are shown in Fig. 4. It

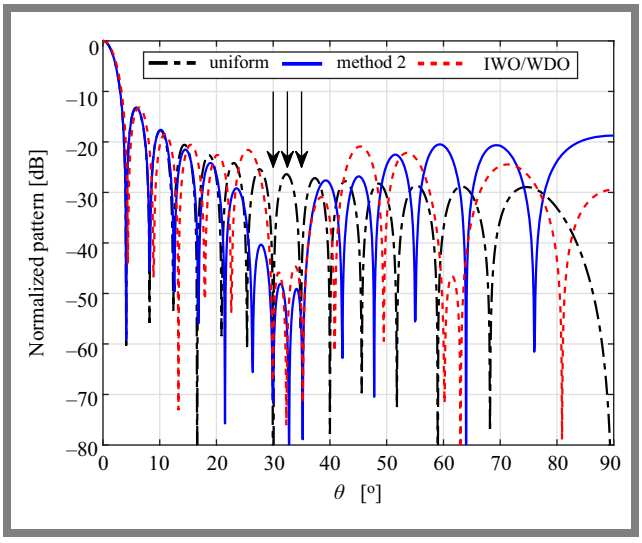

Fig. 4. Patterns of the uniform 28-element array, method 2 and the hybrid IWO/WDO from [10]. The directions of the nulls are indicated by the arrows.

Tab. 2. Geometry of the optimized 28-element array normalized with respect to $\lambda/2$. The perturbed elements using method 2 are marked in bold.

n	Method 1	Method 2
1	±0.6763	±0.5000
2	±1.0383	±1.5000
3	±2.2080	**±2.4339**
4	±2.7874	±3.5000
5	±3.9208	±4.5000
6	±4.6472	±5.5000
7	±5.5773	±6.5000
8	±6.7073	±7.5000
9	±7.8581	±8.5000
10	±8.7781	±9.5000
11	±10.1495	±10.5000
12	±11.2278	±11.5000
13	±12.3734	**±12.1901**
14	±13.8323	**±13.6986**

can be seen from the figure the all the constraints on the radiation pattern are met by perturbing the locations of only 6 array elements. The optimized array geometry using method 2 is given in Tab. 2.

6. Conclusions

In this paper, two methods for the synthesis of aperiodic linear arrays were presented. For the first method, the problem of PSLL reduction and forming prescribed nulls in the radiation pattern of the array by optimizing the position of array elements was formulated as a convex optimization problem and solved iteratively.

For the second method, only a small number of the array elements are perturbed from their original positions to achieve a predefined upper bound on the PSLL and form prescribed nulls in the radiation pattern with an upper bound on the null depth. The two methods were compared to results from the literature using stochastic optimization techniques such as PSO, CLPSO, and hybrid IWO/WDO. The results showed the effectiveness of the proposed methods.

References

[1] W.-P. Liao and F.-L. Chu, "Null Steering in Planar Arrays by Controlling only Current Amplitudes Using Genetic Algorithms", *Microwave and Optical Technology Letters*, vol. 16, pp. 97–103, 1997 (https://doi.org/10.1002/(SICI)1098-2760(19971005)16:2<97::AID-MOP11>3.0.CO;2-5).

[2] H. Ibrahim, "Null Steering by Real-weight Control – a Method of Decoupling the Weights", *IEEE Transactions on Antennas and Propagation*, vol. 39, pp. 1648–1650, 1991 (https://doi.org/10.1109/8.102781).

[3] M.M. Dawoud and M. Nuruzzaman, "Null Steering in Rectangular Planar Arrays by Amplitude Control Using Genetic Algorithms", *International Journal of Electronics*, vol. 87, pp. 1473–1484, 2000 (https://doi.org/10.1080/00207210050192498).

[4] Y. Aslan, J. Puskely, A. Roederer, and A. Yarovoy, "Phase-only Control of Peak Sidelobe Level and Pattern Nulls Using Iterative Phase Perturbations", *IEEE Antennas and Wireless Propagation Letters*, vol. 18, pp. 2081–2085, 2019 (https://doi.org/10.1109/LAWP.2019.2937682).

[5] G. Buttazzoni, M. Comisso, F. Ruzzier, and R. Vescovo, "Phase-only Antenna Array Reconfigurability with Gaussian-shaped Nulls for 5G Applications", *International Journal of Antennas and Propagation*, vol. 2019, pp. 1–8, 2019 (https://doi.org/10.1155/2019/9120530).

[6] S.E. El-Khamy, N.O. Korany, and M.A. Abdelhay, "Minimizing Number of Perturbed Elements in Linear and Planar Adaptive Arrays with Broad Nulls Using Compressed Sensing Approach", *IET Microwaves, Antennas & Propagation*, vol. 13, pp. 1134–1141, 2019 (https://doi.org/10.1049/iet-map.2018.5221).

[7] M.H. Er, "Linear Antenna Array Pattern Synthesis with Prescribed Broad Nulls", *IEEE Transactions on Antennas and Propagation*, vol. 38, pp. 1496–1498, 1990 (https://doi.org/10.1109/8.57004).

[8] M.A. Abdelhay and S.E. El-Khamy, "A Compressed Sensing-based Approach for Null Steering in Partially Adaptive Planar Arrays Using a Reduced Number of Adjustable Array Elements", *Digital Signal Processing*, vol. 145, art. no. 104311, 2024 (https://doi.org/10.1016/j.dsp.2023.104311).

[9] J. Hejres, "Null Steering in Phased Arrays by Controlling the Positions of Selected Elements", *IEEE Transactions on Antennas and Propagation*, vol. 52, pp. 2891–2895, 2004 (https://doi.org/10.1109/TAP.2004.835128).

[10] S.K. Mahto and A. Choubey, "A Novel Hybrid IWO/WDO Algorithm for Interference Minimization of Uniformly Excited Linear Sparse Array by Position-only Control", *IEEE Antennas and Wireless Propagation Letters*, vol. 15, pp. 250–254, 2016 (https://doi.org/10.1109/LAWP.2015.2439959).

[11] M. Pour, T.H. Mitha, and E.C. Brothers, "A Combined Electronic Position – and Partial Amplitude – Control Synthesis Technique for Sidelobe Reductions in Linear Array Antennas", *IEEE Transactions on Microwave Theory and Techniques*, vol. 71, pp. 5074–5081, 2023 (https://doi.org/10.1109/TMTT.2023.3288634).

[12] A. Massa, P. Rocca, and G. Oliveri, "Compressive Sensing in Electromagnetics – A Review", *IEEE Antennas and Propagation Magazine*, vol. 57, pp. 224–238, 2015 (https://doi.org/10.1109/MAP.2015.2397092).

[13] E. Candes and M. Wakin, "An Introduction to Compressive Sampling", *IEEE Signal Processing Magazine*, vol. 25, pp. 21–30, 2008 (https://doi.org/10.1109/MSP.2007.914731).

[14] R.G. Baraniuk, "More is Less: Signal Processing and the Data Deluge", *Science*, vol. 331, pp. 717–719, 2011 (https://doi.org/10.1126/science.1197448).

[15] G. Buttazzoni, F. Babich, F. Vatta, and M. Comisso, "Geometrical Synthesis of Sparse Antenna Arrays Using Compressive Sensing for 5G IoT Applications", *Sensors*, vol. 20, art. no. 350, 2020 (https://doi.org/10.3390/s20020350).

[16] G. Tang, B.N. Bhaskar, P. Shah, and B. Recht, "Compressed Sensing Off the Grid", *IEEE Transactions on Information Theory*, vol. 59, pp. 7465–7490, 2013 (https://doi.org/10.1109/TIT.2013.2277451).

[17] E. Candes and J. Romberg, "Sparsity and Incoherence in Compressive Sampling", *Inverse Problems*, vol. 23, art. no. 969, 2007 (https://doi.org/10.1088/0266-5611/23/3/008).

[18] Z. Tan, P. Yang, and A. Nehorai, "Joint Sparse Recovery Method for Compressed Sensing with Structured Dictionary Mismatches", *IEEE Transactions on Signal Processing*, vol. 62, pp. 4997–5008, 2014 (https://doi.org/10.1109/TSP.2014.2343940).

[19] M.A. Abdelhay, N.O. Korany, and S.E. El-Khamy, "Synthesis of Uniformly Weighted Sparse Concentric Ring Arrays Based on Off-grid Compressive Sensing Framework", *IEEE Antennas and Wireless Propagation Letters*, vol. 20, pp. 448–452, 2021 (https://doi.org/10.1109/LAWP.2021.3052174).

[20] F. Yan, F. Yang, T. Dong, and P. Yang, "Synthesis of Planar Sparse Arrays by Perturbed Compressive Sampling Framework", *IET Microwaves, Antennas & Propagation*, vol. 10, pp. 1146–1153, 2016 (https://doi.org/10.1049/iet-map.2015.0775).

[21] F. Yan *et al.*, "An Alternating Iterative Algorithm for the Synthesis of Complex-excitation and Pattern Reconfigurable Planar Sparse Array", *Signal Processing*, vol. 135, pp. 179–187, 2017 (https://doi.org/10.1016/j.sigpro.2017.01.008).

[22] M. Khodier and C. Christodoulou, "Linear Array Geometry Synthesis with Minimum Sidelobe Level and Null Control Using Particle Swarm Optimization", *IEEE Transactions on Antennas and Propagation*, vol. 53, pp. 2674–2679, 2005 (https://doi.org/10.1109/TAP.2005.851762).

[23] S.K. Goudos *et al.*, "Application of a Comprehensive Learning Particle Swarm Optimizer to Unequally Spaced Linear Array Synthesis with Sidelobe Level Suppression and Null Control", *IEEE Antennas and Wireless Propagation Letters*, vol. 9, pp. 125–129, 2010 (https://doi.org/10.1109/LAWP.2010.2044552).

[24] M. Grant and S. Boyd, "CVX: Matlab Software for Disciplined Convex Programming, version 2.1", 2014.

[25] E.J. Candès, M.B. Wakin, and S.P. Boyd, "Enhancing Sparsity by Reweighted l_1 Minimization", *Journal of Fourier Analysis and Applications*, vol. 14, pp. 877–905, 2008 (https://doi.org/10.1007/s00041-008-9045-x).

Magdy A. Abdelhay, Ph.D., Assistant Professor
Department of Information and Communications Engineering
https://orcid.org/0000-0003-4244-0840
E-mail: abdelhay@ieee.org
Al-Farqadein University College, Basrah, Iraq
https://www.fu.edu.iq/en

Optimal Filter Selection for MIMO F-OFDM Systems in 5G Wireless Communication

Fadila Amel Miloudi[1], Mohammed Sofiane Bendelhoum[1], Fayssal Menezla[2], and Ridha Ilyas Bendjillali[1]

[1] *University Center Nour Bachir, El-Bayadh, Algeria,*
[2] *Université Djillali Liabès of Sidi Bel-Abbès, El-Bayadh, Algeria*

https://doi.org/10.26636/jtit.2025.3.2171

Abstract — Strong demand for mobile broadband cellular systems has boosted the popularity of emerging high-speed modulation technologies such as multiple input multiple output (MIMO) and cyclic prefix orthogonal frequency division multiplexing (CP-OFDM). However, CP-OFDM suffers from some significant drawbacks in 5G networks, including severe out-of-band emissions (OOBE) and poor spectral efficiency. Filtered orthogonal frequency division multiplexing (F-OFDM) has therefore been found to be a good alternative, as it allows to address these shortcomings by relying on digital filtering to eliminate OOBE and improve spectral efficiency. This study focuses on evaluating the performance of MIMO F-OFDM systems and comparing it with the results achieved by MIMO CP-OFDM, with a particular emphasis placed on reducing spectral leakage and improving overall system performance by using various window functions. Six window types, including Hanning, Hamming, Blackman, root raised cosine (RRC), Nuttall, and Blackman-Harris, are investigated. The research aimed to assess the performance of the system in terms of power spectral density (PSD), peak-to-average power ratio (PAPR), and bit error rate (BER), while using different modulation schemes, i.e. QPSK, 16QAM, 64QAM, and 256QAM, over Rayleigh fading and AWGN channels. Simulation results show that the proposed window filter (Nuttall-Blackman-Hanning) significantly reduces OOBE while maintaining efficient spectral performance. The findings demonstrate that MIMO F-OFDM with the proposed filters achieves better spectral efficiency and reliability, making it a promising candidate for 5G applications requiring high data rates, low latency, and robust signal integrity.

Keywords — 5G, BER, CP-OFDM, F-OFDM, MIMO, PAPR, PSD

1. Introduction

Rapid development of 5G wireless communication systems has resulted in a pressing need for higher data rates, enhanced spectral efficiency, and low latency to support a wide range of applications. These applications include enhanced mobile broadband (eMBB), massive machine-type communications (mMTC), and ultra-reliable low-latency communications (URLLC) [1], [2]. As demand for reliable and high-speed communication continues to increase, existing technologies used in previous generation solutions, such as cyclic prefix orthogonal frequency division multiplexing (CP-OFDM), face significant challenges in meeting the stringent requirements of 5G networks.

CP-OFDM, which has been successfully utilized in 4G LTE systems, is well known for its simplicity and effectiveness in managing channel properties [3], [4]. However, its limitations, such as high OOBE and inefficient spectral utilization, hinder its suitability for the demanding environment of 5G. These drawbacks result in interference with adjacent frequency bands, reducing overall system performance and spectral efficiency [5]–[8]. To address these challenges, various alternative waveform techniques have been proposed for 5G, including filter bank multi-carrier (FBMC) and generalized frequency division multiplexing (GFDM), each aiming to optimize spectral efficiency and reduce OOBE [9]– [12]. However, these techniques often suffer from increased system complexity and additional computational overhead.

One promising alternative to CP-OFDM is filtered orthogonal frequency division multiplexing (F-OFDM), which improves spectral efficiency by applying digital filters to reduce OOBE while maintaining the core advantages of CP-OFDM, such as its robustness against multipath fading and ability to manage high data rates. F-OFDM has emerged as a practical solution, particularly for asynchronous transmission scenarios, making it highly compatible with 5G requirements [13]–[15].

Despite their advantages, performance of F-OFDM systems depends heavily on the design of the filter applied. Specifically, the selection of window functions plays a critical role in reducing OOBE and optimizing system efficiency. Previous studies have typically focused on individual window functions such as Hanning, Hamming, and root raised cosine (RRC). However, there is a need for a more systematic evaluation of different window designs to identify the most effective filter for F-OFDM in 5G applications.

1.1. Research Problem and Objective

This study addresses the gap that exists in current research by systematically evaluating the performance of F-OFDM systems in multiple input multiple output (MIMO) setups, using six different window functions: Hanning, Hamming, Blackman, RRC, Nuttall, and Blackman-Harris. The primary

aim of the research is to identify the optimal window function or combination of window functions that minimizes OOBE and enhances overall spectral efficiency, while also analyzing the impact on key performance metrics such as power spectral density (PSD), peak-to-average power ratio (PAPR) and bit error rate (BER). The study further explores how these filters perform under different modulation schemes and channel conditions, specifically in additive white Gaussian noise (AWGN) and Rayleigh fading environments.

By addressing these issues, the research seeks to provide a comprehensive solution to improve spectral efficiency and reliability of F-OFDM systems – properties which are critical for the successful deployment of 5G networks in real-world scenarios such as urban environments, Internet of Things (IoT) applications and high-speed vehicular communications.

Additionally, this study aims to contribute to the ongoing development of filtering techniques that could be essential not only for 5G, but also for future 6G networks.

Compared to traditional MIMO CP-OFDM systems, the proposed MIMO F-OFDM architecture with optimized window filters demonstrates superior performance in terms of OOBE reduction and BER, under various channel conditions. This validates its potential as a robust solution for advanced wireless communication scenarios.

The subsequent sections of this work are organized as follows. Section 2 examines the relevant literature that specifically addresses F-OFDM solutions used in wireless communication systems. Section 3 introduces the F-OFDM-MIMO system model and delineates the process of shaping the symbol spectrum. Within the F-OFDM system, the window function is used to truncate the sinc function for the purpose of constructing the filter. The analysis and discussion of the simulation findings are presented in Section 4. Finally, the document is concluded in Section 5 which summarizes key results and provides insights into future research directions.

2. Related Work

Taking into account previous discussions, the design of the filter needs to be approached in a way that avoids adding unnecessary complexity or delays to the system. Furthermore, the filter-based waveform must meet the requirements of 5G solutions. CP-OFDM utilizes a sinc function-based filter in the frequency domain, which tends to exhibit high side lobes, and thus negatively affects system performance. A feasible filter that can be directly implemented within the system is the truncated-sinc filter, which is created by windowing the sinc function with a suitably selected window [16]. Consequently, F-OFDM based on the windowed-sinc filter has been shown to surpass CP-OFDM in terms of performance.

The authors of [17] performed a benchmark study to compare the performance of windowed orthogonal frequency division multiplexing (W-OFDM) and F-OFDM. Their findings concluded that F-OFDM offers a better reduction of OOBE and is suitable for asynchronous communication, particularly when using higher-order modulation methods. Furthermore, [18] provide a concise overview of the performance of F-OFDM while employing MIMO setups, such as SIMO and MISO, together with several digital modulation methods. Across various settings, the results showed that MIMO F-OFDM improves system performance.

In [15], the effect of neighboring signal interference on MIMO systems was evaluated by employing F-OFDM and CP-OFDM modulation techniques. Using different detection strategies and BER calculations, the simulations conducted in the course of the study revealed that F-OFDM outperforms CP-OFDM as far as the management of interference is concerned.

Paper [19] extended this work by comparing the performance of CP-OFDM, W-OFDM, and F-OFDM under Rayleigh fading conditions. Its results confirmed that F-OFDM offers better spectral efficiency and overall performance than its counterparts.

The authors of [20] proposed a generic function model to design window functions with high energy concentration and rapid attenuation of the side lobes. Their analysis suggested that optimized window functions reduce OOBE while maintaining nearly the same BER as achieved by traditional window filters. The effectiveness of MIMO systems combined with CP-OFDM and F-OFDM was further explored in [21], where performance was analyzed while using various digital modulation techniques and windowed filters (Hanning, Hamming, Blackman, RRC). The simulations showed that the Hanning filter was robust and efficient in signal recovery.

Paper [22] provides an experimental study on F-OFDM using different windowed-sinc filters, both equal and unequal in sub-band sizes. Its results indicated that F-OFDM, compared to traditional CP-OFDM, not only achieves a lower OOBE but also improves spectrum efficiency by 5–6%.

The authors of [23] investigated the performance of MIMO-OFDM systems within 5G networks, incorporating advanced modulation schemes. They evaluated critical performance metrics including throughput, BER, and spectral efficiency. The findings revealed that while higher-order modulation schemes substantially enhance throughput and spectral efficiency, they also lead to increased BER, especially at lower SNR levels.

In essence, while previous studies have focused on optimizing individual window functions or exploring specific modulation techniques, the present study contributes by systematically evaluating six different window designs and analyzing their effects across key performance metrics, offering a comprehensive solution for 5G communication systems.

A closely related study [24] titled “Optimal filter choice for filtered OFDM” compared six window functions for F-OFDM systems and identified the Kaiser window as the most effective in its evaluation settings. While the authors focused primarily on SISO configurations and evaluated filters such as Hamming, Hanning, Blackman, and Kaiser, this study extends the investigation to MIMO scenarios using a broader set of six window functions, including their combinations, such as Nuttall-Blackman-Hanning. These combinations demonstrated an improved performance in minimizing OOBE and reduc-

ing BER. This distinction highlights that the term *optimal* is context-dependent and subject to the system's configuration, channel model, and performance objectives.

In order to enhance filter performance by narrowing the main lobe, lowering side lobe power, and minimizing latency, the main goal of this work is to develop an ideal window function for an F-OFDM system. For filter development, the one that meets the least OOBE requirement among the six suggested window function types has been chosen.

Additionally, we investigate the concept of integrating two or three window functions to further decrease OOBE and improve speed. The MIMO technology is integrated with F-OFDM to improve channel reliability and capacity. Consequently, this combination enhances spectrum utilization and reduces BER.

3. F-OFDM System Model

Despite the advantages offered by CP-OFDM, such as optimizing spectrum usage through the orthogonality of subcarriers and its resilience to inter-symbol interference (ISI) in dispersive channels, it is not ideally suited for 5G networks due to the presence of side lobes that result in increased OOBE. These emissions lead to inefficiencies in spectrum utilization and cause interference with adjacent users in neighboring frequency bands, posing a significant challenge for high-capacity communication systems.

One of the primary objectives of 5G is to improve spectral efficiency by reducing the guard band to less than 10% of the allocation used in 4G systems. F-OFDM addresses the limitations of CP-OFDM by combining its advantages with additional features designed specifically to meet the requirements of 5G. In particular, F-OFDM integrates digital filtering techniques, enabling asynchronous transmission and guaranteeing increased throughput. These enhancements make F-OFDM more efficient in spectrum utilization and more effective in mitigating interference [22].

A key strength of F-OFDM lies in its superior performance in time and frequency localization, ensuring accurate signal processing in both critical domains for applications requiring low latency and high spectral efficiency, such as real-time communications and high-capacity mobile networks. Moreover, F-OFDM requires only a simplified equalizer to manage channel effects, making it flexible and efficient in various communication environments. Its structural design allows it to coexist with other waveforms in different sub-bands, thus offering significant flexibility in filter design across sub-bands and thereby minimizing interference between different frequency ranges [25].

F-OFDM, also known as full-band filtered OFDM, provides a robust solution to the challenges faced by traditional CP-OFDM systems in 5G.

In this study, the focus is on analyzing the branch structure of an F-OFDM system, with its architecture illustrated in Fig. 1.

Following the sampling process of the F-OFDM signal illustrated in Fig. 1 [26], the total length of an F-OFDM symbol,

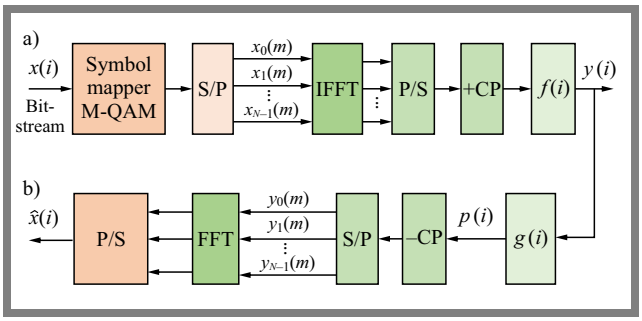

Fig. 1. F-OFDM transceiver scheme: a) upper section of the transmitter and b) lower section of the receiver.

including the cyclic prefix (CP), is denoted as:

$$N_{F-OFDM} = N + N_{cp}\,,$$

where N represents the length of the useful data symbol and N_{cp} denotes the CP length.

The discrete baseband signal $s(i)$ of F-OFDM is defined in the following way:

$$s(i) = \frac{1}{N}\sum_{k=0}^{N-1} x_k(m)\, e^{\frac{j2\pi k(i-N_{cp})}{N}}\,, \tag{1}$$

where $0 \leqslant i \leqslant N_{F-OFDM} - 1$, $x_k(m)$ is the k-th data symbol of the F-OFDM signal of the m-th frame.

Filtering $s(i)$, actually represents a convolution of $s(i)$ and filter $f(i)$, which can be formulated as:

$$y(i) = \frac{1}{N}\sum_{l=-\infty}^{l=\infty} f(i-l)\sum_{k=0}^{N-1} x_k(m)\, e^{\frac{j2\pi k(i-N_{cp})}{N}}\,. \tag{2}$$

Filter design is the primary challenge in F-OFDM systems. Generally, the Hanning window is used to truncate the sinc function and obtain the commonly used filter $f(i)$ [20]. In this work, six types of window functions most commonly used in recent years have been proposed for filter design: Hanning, Hamming, Blackman, RRC, Nuttall, and Blackman-Harris. The window functions are listed bellow [27]:

Hanning window:

$$W_{hn}(i) = 0.5\left[1+\cos\frac{2\,\pi\, i}{N-1}\right],\quad -\frac{N-1}{2} \leqslant i \leqslant \frac{N-1}{2}\,. \tag{3}$$

Hamming window:

$$\begin{gathered} W_{hm}(i) = 0.54 + 0.46\cos\frac{2\,\pi\, i}{N-1}\,, \\ -\frac{N-1}{2} \leqslant i \leqslant \frac{N-1}{2}\,. \end{gathered} \tag{4}$$

Blackman window:

$$\begin{gathered} W_{bl}(i) = 0.42 + 0.5\cos\frac{2\,\pi\, i}{N-1} + 0.08\cos\frac{4\,\pi\, i}{N-1}\,, \\ -\frac{N-1}{2} \leqslant i \leqslant \frac{N-1}{2}\,. \end{gathered} \tag{5}$$

RRC window:

$$\begin{gathered} W_{rrc}(i) = \left[0.5\left(1+\cos\frac{2\,\pi\, i}{N-1}\right)\right]^{0.6}\,, \\ -\frac{N-1}{2} \leqslant i \leqslant \frac{N-1}{2}\,. \end{gathered} \tag{6}$$

Tab. 1. Comparative analysis of this study and previous works.

Ref.	Main focus of previous works	Limitation of previous works	Proposed objectives
[17]	Multirate 5G downlink performance comparison between F-OFDM and W-OFDM schemes using various methods	– Study the performance of F-OFDM only, – Study using RRC and Blackman windows only for F-OFDM, – Uses AWGN only, – Includes effect of QPSK and 256QAM only, – Analyzes PSD and BER vs. SNR	– Investigation of MIMO technology integrated with F-OFDM, – Analysis of different windowing techniques for F-OFDM, – Evaluation using both AWGN and Rayleigh channels, – Examination of the impact of modulation order, – Study includes PSD, PAPR and BER vs. SNR
[18]	Implementing enhanced MIMO with F-OFDM to boost system efficiency in future 5G cellular networks	– Analyzes the performance of MIMO F-OFDM system-based RRC window filter, – Study includes PSD and BER vs. SNR	– Study the performance of MIMO combined with F-OFDM, – Six different window functions have been suggested for the filter design, – Includes PSD, PAPR and BER vs. SNR
[15]	Performance comparison of F-OFDM and OFDM for MIMO systems in a 5G scenario	– Analyzes the performance of a MIMO F-OFDM system using an RRC window filter, – Study uses QPSK modulation only	– Performance analysis of MIMO combined with F-OFDM, – Six different window functions have been suggested for the filter design. – Work includes effect of modulation order
[19]	Performance evaluation of OFDM, W-OFDM, and F-OFDM over Rayleigh fading channels for 5G systems	– Study the performance of the MIMO F-OFDM system-based RRC window filter, – Work includes PSD and BER vs. SNR	– Performance evaluation of MIMO combined with F-OFDM system-based six types of window filters for filter design, – Work includes PSD, PAPR and BER vs. SNR
[20]	Filter design based on a generic function model for reducing OOBE in F-OFDM	– Evaluation of the performance of F-OFDM only with different window filters using the effect of single, combination, and squared window filters, – Study uses AWGN and 64QAM only	– Analysis of the performance of MIMO combined with F-OFDM, – Six different window functions have been proposed for the filter design, – Using the effect of single, combination, and squared window filters, – Study using the AWGN and Rayleigh channel and includes the effect of modulation order
[21]	Design and performance evaluation of MIMO F-OFDM systems for 5G and beyond	– Four different types of window functions have been proposed for the filter design, – Study includes PSD and BER vs. SNR	– Six different window functions have been introduced for the filter design, – Work includes PSD, PAPR and BER vs. SNR
[22]	An experimental investigation of F-OFDM spectrum efficiency for 5G applications	– Analysis of F-OFDM only, – Study using AWGN only, – Five types of window functions have been proposed for the filter design. – Work includes PSD and BER vs. SNR	– Study the performance of MIMO combined with F-OFDM, – Two cases: AWGN and Rayleigh channel, – Six types of window function have been proposed for the filter design, – Includes PSD, PAPR and BER vs. SNR comparison
[23]	Performance analysis of MIMO-OFDM systems	– Performance analysis of MIMO-OFDM, – Study includes PSD and BER vs. SNR	– Study the performance of MIMO F-OFDM, – Research includes PSD, PAPR and BER vs. SNR

Nuttall window:

$$W_{nutt}(i) = a_0 + a_1 \cos\frac{2\pi i}{N-1} + a_2 \cos\frac{4\pi i}{N-1} + a_3 \cos\frac{6\pi i}{N-1}, \quad -\frac{N-1}{2} \leqslant i \leqslant \frac{N-1}{2}. \tag{7}$$

with: $a_0 = 0.355768$, $a_1 = 0.487396$, $a_2 = 0.144232$, $a_3 = 0.012604$.

Blackman–Harris window:

$$W_{bl_{harris}}(i) = a_0 + a_1 \cos\frac{2\pi i}{N-1} + a_2 \cos\frac{4\pi i}{N-1} + a_3 \cos\frac{6\pi i}{N-1}, \quad -\frac{N-1}{2} \leqslant i \leqslant \frac{N-1}{2}. \tag{8}$$

with: $a_0 = 0.35875$, $a_1 = 0.48829$, $a_2 = 0.14128$, $a_3 = 0.01168$.

In Eq. (8), N is the signal length after sampling and N is an odd number.

The definition of the discredited sinc function is as follows:

$$\mathrm{sinc}(i) = \frac{\sin(w_c\, i)}{w_c\, i}, \quad -\frac{N-1}{2} \leqslant i \leqslant \frac{N-1}{2}\,. \tag{9}$$

w_c is the cutoff frequency of an ideal low pass filter (LPF).

Therefore, a windowed-sinc filter can be formulated as follows:

$$f(i) = \mathrm{sinc}(i).W(i)\,, \tag{10}$$

where $\mathrm{sinc}(i)$ represents the time domain sinc response of LPF and i is the time domain window function.

Based on Fig. 1 and the matched filtering method from signal detection theory [28], the transmitter applies a filter $f(i)$ to filter signal symbols in the time domain, while the receiver uses a matched filter $g(i)$. In accordance with the signal detection theory, at the receiver side, the matched filter satisfies $g(i) = f(-i)$; if $f(i)$ is set as a real-valued even function, then $f(i) = f(-i)$, implying $g(i) = f(i)$.

Consequently, after passing through the matched filter, signal $p(i)$ is obtained as $p(i) = y(i) * g(i)$, where $*$ denotes the convolution operation. After eliminating the cyclic prefix (CP), performing serial-to-parallel conversions, fast Fourier transform (FFT), and parallel-to-serial conversion, $\hat{x}(i)$ can be derived as follows:

$$\hat{x}(i) = \frac{1}{N} \sum_{l=-\infty}^{l=\infty} p(i-l) \sum_{k=0}^{N-1} y_k(i)\, e^{\frac{j2\pi k(i-N_{cp})}{N}}\,, \tag{11}$$

While the types of window functions and simulation parameters have been outlined, this section aims to provide a more comprehensive explanation of the specific procedures followed throughout the simulations by applying them to the time-domain signal by truncating the sinc function, in line with the filter design methodology. For each window type, the corresponding filter coefficients were derived and integrated into the F-OFDM.

Regarding modulation schemes, we employed quadrature phase shift keying (QPSK) as well as m-ary quadrature amplitude modulation (M-QAM) with M= 16, 64 and 256. The system's performance was evaluated in terms of key metrics, including PSD, PAPR, and BER, using varying modulation schemes. The simulations were conducted under both AWGN and Rayleigh fading channel conditions to ensure a thorough and representative assessment of filter performance across different channel environments.

This approach ensures the accuracy and reliability of the results, establishing a foundation for comparing the performance of MIMO F-OFDM and MIMO CP-OFDM systems.

4. Results and Discussion

As noted in [29], the filter length in an F-OFDM system typically follows the inequality:

$$N \leqslant \frac{N_{F-OFDM}}{2}\,,$$

where N represents the filter length and N_{F-OFDM} refers to the length of the F-OFDM symbol. In this case, the symbol length is set to 1024, which makes the FFT/IFFT length 1024 as well. Since the filter length N is usually chosen as an odd number, this study uses a filter length of 513. The modulation techniques used are QPSK and M-QAM, with M =16, 64, 256 representing 16QAM, 64QAM and 256QAM, respectively. A total of 600 subcarriers have been utilized.

In this study, the MIMO technology, known for improving capacity and enhancing channel reliability, is combined with F-OFDM. Specifically, the paper explores the integration of MIMO 4 × 4 with F-OFDM using an optimal window function to maximize system performance.

The filters exhibited varying performance, and a comparative analysis is necessary for clarity. For example, the Nuttall window provided the best overall performance in reducing OOBE. Although the two proposed window filters 1 and 2 performed well in terms of spectral efficiency and signal recovery, the proposed window filter 2 was effective in minimizing BER but less successful in reducing OOBE. This comparison highlights the need to select the appropriate window filter according to the specific performance criteria, such as spectral efficiency, BER, or PAPR.

To gain further insight into the OOBE characteristics, PSD of the F-OFDM system is calculated. This analysis offers a deeper understanding of how OOBE is affected by different filter designs.

Let us rewrite Eq. (1) in a continuous form:

$$s(t) = \frac{1}{N} \sum_{k=0}^{N-1} x_k(m)\, e^{j2\pi k \Delta f t}, \quad -\frac{T_s}{2} \leqslant t \leqslant \frac{T_s}{2}\,, \tag{12}$$

where: $T_s = \frac{1}{\Delta f} + T_{cp}$ and T_{cp} is the time length of the CP.

Upon completion of the filter $f(t)$:

$$y(t) = s(t) * f(t)\,. \tag{13}$$

The signal at the transmitter end of the F-OFDM system is known as $y(t)$, $f(t) = w(t).\mathrm{sinc}(t)$.

The Fourier transformation is used to transform the transmitted signal:

$$Y(f) = S(f).F(f)\,. \tag{14}$$

where $Y(f)$ is the Fourier transform of the transmitted signal $y(t)$, $S(f)$ is the Fourier transform of $s(t)$, $F(f)$ is the Fourier transform of the filter $f(t)$.

$$S(f) = \frac{1}{N} \sum_{k=0}^{N-1} x_k(m)\, T_s\, \mathrm{sinc}\big[\pi(f - k\,\Delta f)T_s\big]\,, \tag{15}$$

$$\begin{aligned} F(f) = \mathcal{F}\big[w(t).\mathrm{sinc}(t)\big] &= \mathcal{F}\big[w(t)\big] * \mathcal{F}\big[\mathrm{sinc}(t)\big] \\ &= W(f) * \mathcal{F}\big[\mathrm{sinc}(t)\big]\,, \end{aligned} \tag{16}$$

where $\mathcal{F}[.]$ is the Fourier transform symbol and $W(f)$ represents the Fourier transform of the window function. The degree of OOBE of $S(f)$ is determined by the degree of OOBE of the system and $W(f)$ is responsible for the PSD of $S(f)$.

Tab. 2. Overview of the parameters of the simulated model.

Parameters	Considerations for simulation
Message type	Binary bits
Channel	AWGN and Rayleigh
SNR	0 to 22 dB
Filters used for F-OFDM	Hanning, Hamming, Blackman, RRC, Nuttall, and Blackman-Harris
Number of subcarriers	600
CP length	74 for CP-OFDM and 72 for F-OFDM
Filter order	512
Tone offset	5 subcarriers
IFFT / FFT size	1024
Digital modulation	QPSK, 16QAM, 64QAM, and 256QAM
Antenna configuration	T4 × R4

PSD for the F-OFDM system is:

$$10\log_{10}\left|Y(f)\right|^2 = 10\log_{10}\left|S(f).\left\{W(f) * \mathcal{F}[\mathrm{sinc}(t)]\right\}\right|^2. \tag{17}$$

PAPR of the F-OFDM system is as follows:

$$PAPR_{F\_OFDM} = \frac{\max(|y(t)|^2)}{E(|y(t)|^2)}, \tag{18}$$

in which $\max(.)$ is the F-OFDM signal's maximum value and $E(.)$ is its average value.

Table 2 provides a summary of the performance characteristics of the simulated model, namely PSD, PAPR, and BER.

4.1. Power Spectral Density

This subsection compares the performance of several window-based filters, including RRC, Hanning, Blackman, Hamming, Blackman-Harris, and Nuttall, in terms of their impact on the power spectrum.

As illustrated in Fig. 2, the filters significantly reduce OOBE when compared to traditional OFDM. Specifically, greater attenuation is achieved in the stopband region with the application of these filters, leading to a marked improvement in spectral containment.

Table 3 highlights the difference in OOBE between the CP-OFDM and F-OFDM systems using various window filters. In CP-OFDM, an attenuation of 66 dB is achieved, with the highest spectral leakage (out-of-band emission). Windows such as RRC, Hanning, Blackman, Hamming, Blackman-Harris, and Nuttall show a clear improvement in reducing spectral leakage compared to CP-OFDM. Among the window functions tested, the proposed windows 1 and 2 offer a sharp and rapid reduction in out-of-beam PSD, which means that they are more effective in reducing spectral interference.

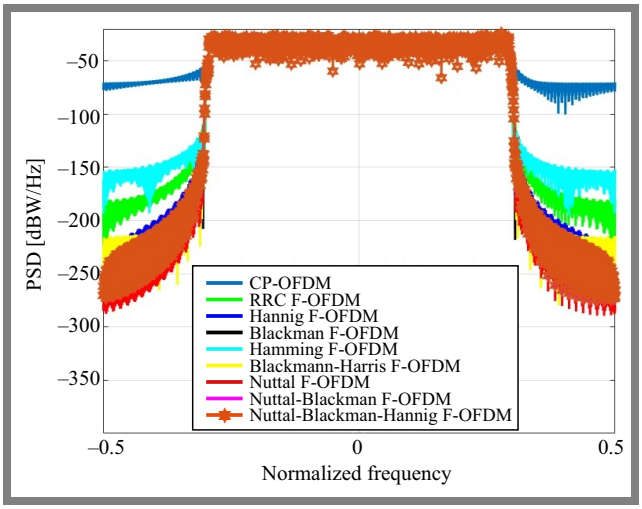

Fig. 2. PSDs of different windows-based F-OFDM systems (50 resource blocks, 12 subcarriers each) compared to CP-OFDM.

Tab. 3. Comparison of F-OFDM OOBE with CP-OFDM.

Window filter	OOBE difference [dB]
Hamming	–152
RRC	–175
Blackman-Harris	–207
Hanning	–233
Blackman	–250
Nuttall	–256
Proposed window 1 (Nuttall-Blackman)	–253
Proposed window 2 (Nuttall-Blackman-Hanning)	–252

The comparison reveals that windows-based filtering, particularly with the Nuttall window and the two proposed windows, offers superior OOBE suppression. We also notice that the PSD values are very close, which makes it a highly effective solution for enhancing spectral efficiency of F-OFDM systems in 5G networks. Although the performance curves (e.g., PAPR and BER) exhibit very close numerical values, even slight improvements can lead to noticeable enhancements in real-time 5G applications, particularly in ultra-reliable low-latency communications (URLLC), where every decibel counts.

4.2. Peak to Average Power Ratio

Next, we assess PAPR for a CP-OFDM system, both with and without the application of a windowed sinc filter, across various modulation schemes. The evaluation is based on the complementary cumulative distribution function (CCDF), which illustrates the probability that the instantaneous power of a signal exceeds a specified threshold relative to its average power level. Figures 3–6 show CCDF comparisons for

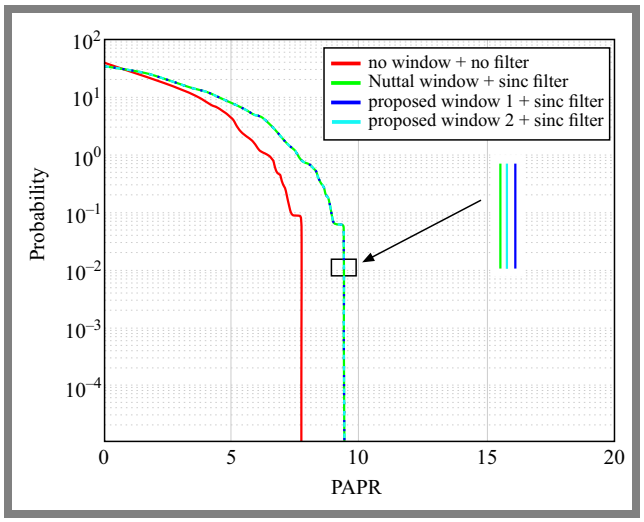

Fig. 3. Comparison of CCDF for CP-OFDM and F-OFDM systems with QPSK modulation.

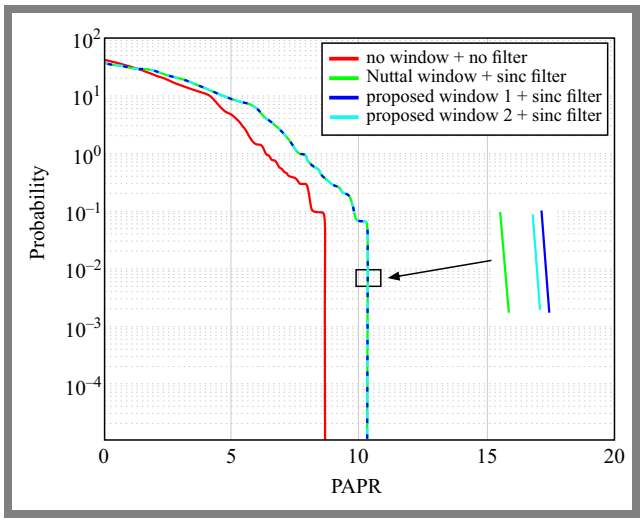

Fig. 4. Comparison of CCDF for CP-OFDM and F-OFDM systems with 16QAM modulation.

different modulation schemes: QPSK, 16QAM, 64QAM and 256QAM.

The results indicate that CP-OFDM systems across all modulation orders exhibit lower PAPR values compared to their F-OFDM counterpart. This suggests that CP-OFDM systems are less prone to high instantaneous power peaks. However, it is important to note that F-OFDM systems, particularly when using single and combination window filters, follow similar PAPR trends. These trends reflect that while F-OFDM achieves superior spectral efficiency, it may lead to slightly higher PAPR values, which could affect energy efficiency in selected applications.

Table 4 provides a detailed comparison of PAPR values between CP-OFDM and F-OFDM systems using the Nuttall window filter and the proposed window filters 1 and 2, for different modulation orders. The two proposed window filters 1 and 2 demonstrate a match in PAPR values with the Nuttall filter, indicating that filter design does not have a significant impact on power efficiency in F-OFDM systems.

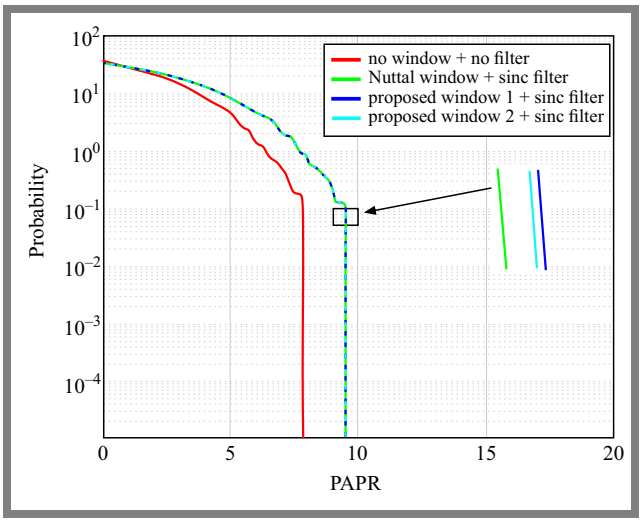

Fig. 5. Comparison of CCDF for CP-OFDM and F-OFDM systems with 64QAM modulation.

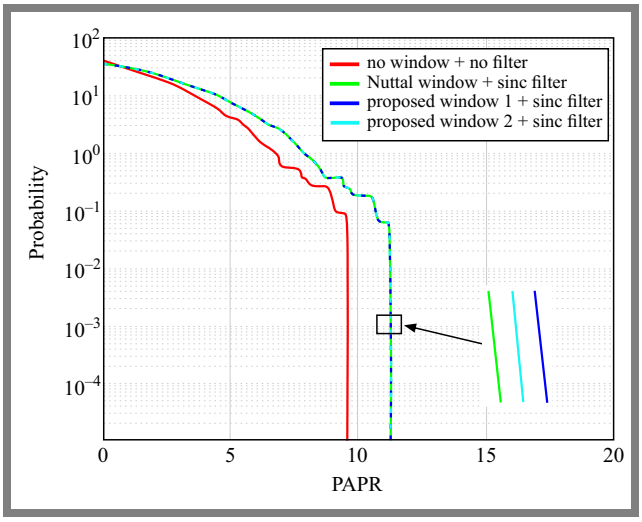

Fig. 6. Comparison of CCDF for CP-OFDM and F-OFDM systems with 256QAM modulation.

The data show that as the modulation order increases, PAPR values tend to rise, which is a common characteristic of higher-order modulation schemes. When QPSK modulation is employed, we see an improvement in PAPR performance compared to all modulation orders, but F-OFDM systems generally present higher PAPR values compared to CP-OFDM, suggesting a trade-off between spectral efficiency and power efficiency.

4.3. Bit Error Rate

The primary performance indicators of a communication system are its efficiency and reliability. While efficiency is improved by maximizing frequency band, reliability remains a critical factor, particularly in systems, where maintaining a low error rate is essential. This study evaluates the reliability of MIMO systems combined with F-OFDM, specifically focusing on the Nuttall window filter and the two proposed window filters 1 and 2. For comparison, BER performance of MIMO systems using CP-OFDM filters is also assessed.

Tab. 4. PAPR with different modulation order of CP-OFDM and F-OFDM.

Modulation order	PAPR for CP-OFDM [dB]	PAPR for F-OFDM-based Nuttall window filter [dB]	PAPR for F-OFDM-based prop. window filter 1 [dB]	PAPR for F-OFDM-based prop. window filter 2 [dB]
4	7.8011	9.4640	9.4662	9.4663
16	7.9388	9.6076	9.6070	9.6070
64	8.6504	10.3150	10.3156	10.3159
256	9.6376	11.3052	11.3052	11.3049

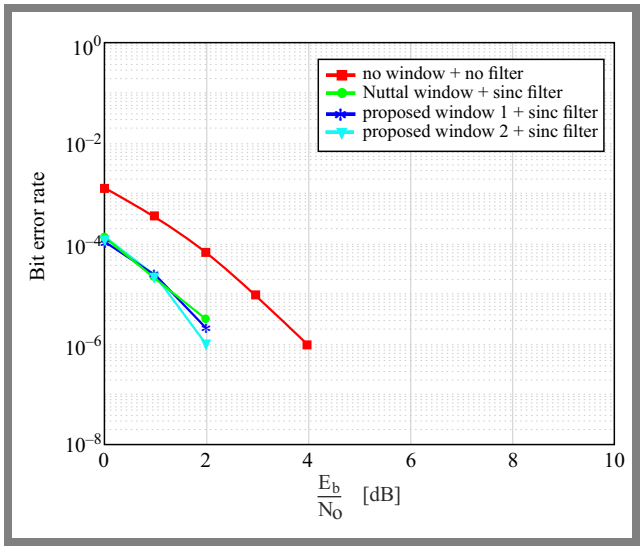

Fig. 7. BER for MIMO CP-OFDM- and F-OFDM-based single window filter and a combination thereof, using QPSK modulation.

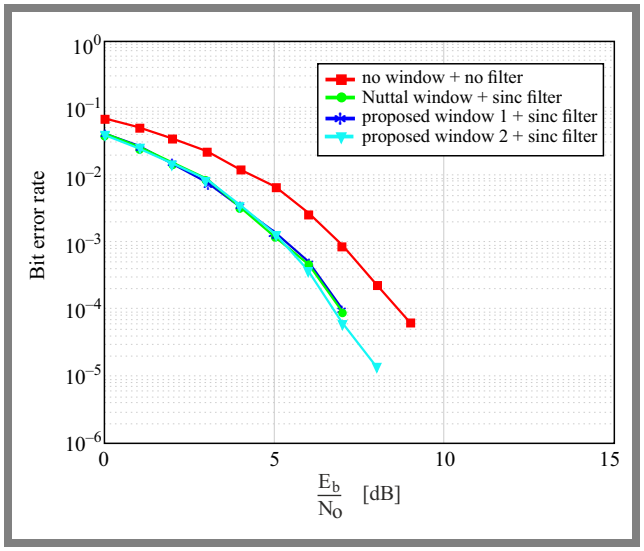

Fig. 8. BER for MIMO CP-OFDM- and F-OFDM-based single window filter and a combination thereof, using 16QAM modulation.

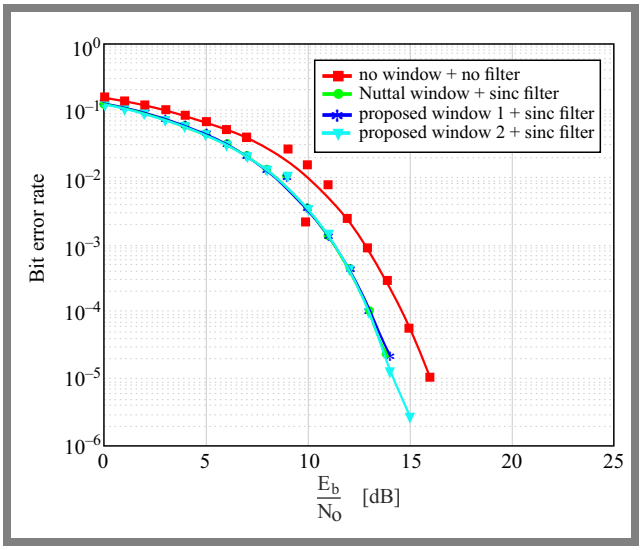

Fig. 9. BER for MIMO CP-OFDM- and F-OFDM-based single window filter and a combination thereof, using 64QAM modulation.

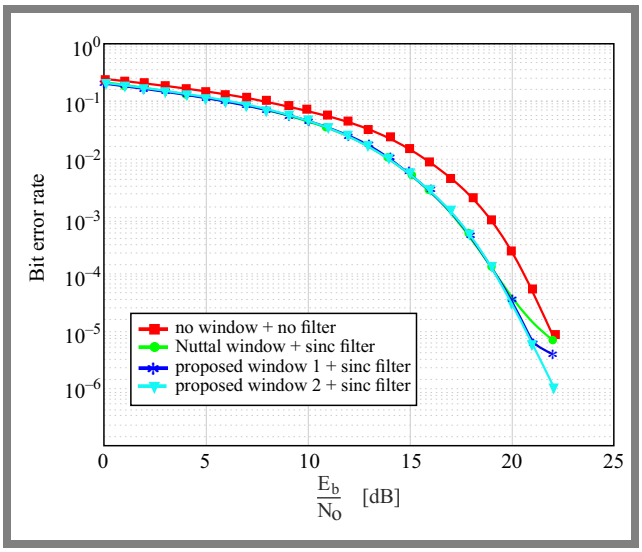

Fig. 10. BER for MIMO CP-OFDM- and F-OFDM-based single window filter and a combination thereof, using 256QAM modulation.

The evaluation is conducted by measuring BER across several modulation schemes, including QPSK, 16QAM, 64QAM and 256QAM, under different conditions. Figures 7–10 display the performance of MIMO CP-OFDM and MIMO F-OFDM systems with single and combined window filters, offering a comprehensive view of BER for each modulation scheme.

The results indicate that MIMO combined with F-OFDM achieves better BER performance compared to MIMO CP-OFDM with all varieties of digital modulation schemes employed. Specifically, when QPSK modulation is used, nearly identical BER values were obtained by MIMO F-OFDM-based Nuttall window filter, proposed window filter 1 (Nuttall-

Blackman) and proposed window filter 2 (Nuttall-Blackman-Hanning). But when SNR values are greater than 1 dB, we notice that proposal no. 2 offers better performance. After that, when employing higher-order modulation schemes, MIMO F-OFDM offers lower BER performance than CP-OFDM. When transmitted over the same Rayleigh channel, nearly identical BER values were obtained by the OFDM filtered with a single and with a combination of window filters when SNR was lower than 5 dB, 13 dB, and 20 dB, respectively.

Although at SNR values greater than those mentioned previously, the BER values achieved for the proposed window filter 2 are the best compared to those observed with the Nuttall window filter and the proposed window filter 1.

Despite the convergent spectral properties of all 3 filters, superior BER performance achieved by the proposed window filter 2 remains a key advantage, particularly in scenarios where spectral efficiency and interference minimization are critical. Therefore, it is recommended that the proposed window filter 2 (Nuttall-Blackman-Hanning) be utilized when designing F-OFDM systems, as it offers an optimal balance between OOBE reduction and system performance in 5G communication environments.

It should be noted that the term *optimal*, as used in this work, refers to the best-performing filter configuration under the defined simulation parameters, including modulation schemes, channel models, and window combinations. It does not imply a universally best solution, but rather one that achieves the most favorable trade-offs in terms of OOBE, BER, and PAPR for the studied MIMO F-OFDM system.

5. Conclusions and Future Outlook

In this paper, we analyze the performance of MIMO combined with F-OFDM for 5G mobile communications and compare it to MIMO using CP-OFDM. The results demonstrated that the proposed window filter 2 (Nuttall-Blackman-Hanning) outperforms other window functions, particularly in reducing OOBE and enhancing overall system performance. Consequently, it is recommended for use in F-OFDM systems to ensure optimal spectral efficiency and system reliability.

Although this study has provided a comprehensive analysis of window functions and their impact on MIMO F-OFDM performance, several opportunities for future research remain open. First, future studies could focus on developing adaptive windowing techniques that adjust dynamically in response to varying network conditions. Such techniques could significantly enhance system performance in rapidly changing environments, including dynamic 5G scenarios.

Additionally, testing the proposed system in more complex propagation environments, such as urban or high-density networks, would provide valuable insight into its robustness and scalability. These tests could help further optimize the system for real-world deployments where signal interference and attenuation are more pronounced.

Future research should also focus on optimizing filter design for energy efficiency, a critical factor for energy-sensitive applications such as IoT networks and mobile devices.

References

[1] I.F. Akyildiz, S. Nie, S.-C. Lin, and M. Chandrasekaran, "5G Roadmap: 10 Key Enabling Technologies", *Computer Networks*, vol. 106, pp. 17–48, 2016 (https://doi.org/10.1016/j.comnet.2016.06.010).

[2] ITU-R, "IMT Vision – Framework and Overall Objectives of the Future Development of IMT for 2020 and Beyond", *Recommendation ITU-R M.2083-0*, 2015 (https://www.itu.int/rec/r-rec-m.2083).

[3] S. Tiwari, S. Chatterjee, and S.S. Das, "Comparative Analysis of Waveforms for Fifth Generation Mobile Networks", *IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)*, Bangalore, India, 2016 (https://doi.org/10.1109/ANTS.2016.7947770).

[4] F.-L. Luo and C. Zhang, *Signal Processing for 5G: Algorithms and Implementations*, Hoboken, USA: Wiley, 581 p., 2016 (https://doi.org/10.1002/9781119116493).

[5] S.B. Weinstein, "The History of Orthogonal Frequency-division Multiplexing", *IEEE Communications Magazine*, vol. 47, pp. 26–35, 2009 (https://doi.org/10.1109/MCOM.2009.5307460).

[6] R.Y. Mesleh *et al.*, "Spatial Modulation", *IEEE Transactions on Vehicular Technology*, vol. 57, pp. 2228–2241, 2008 (https://doi.org/10.1109/TVT.2007.912136).

[7] B. Farhang-Boroujeny, "OFDM Versus Filter Bank Multicarrier", *IEEE Signal Processing Magazine*, vol. 28, pp. 92–112, 2011 (https://doi.org/10.1109/MSP.2011.940267).

[8] M.K. Gupta and S. Tiwari, "Performance Evaluation of Conventional and Wavelet Based OFDM System", *AEU - International Journal of Electronics and Communications*, vol. 67, pp. 348–354, 2013 (https://doi.org/10.1016/j.aeue.2012.10.005).

[9] A. Sahin, I. Guvenc, and H. Arslan, "A Survey on Multicarrier Communications: Prototype Filters, Lattice Structures, and Implementation Aspects", *IEEE Communications Surveys & Tutorials*, vol. 16, pp. 1312–1338, 2013 (https://doi.org/10.1109/SURV.2013.121213.00263).

[10] R. Nissel, S. Schwarz, and M. Rupp, "Filterbank Multicarrier Modulation Schemes for Future Mobile Communications", *IEEE Journal on Selected Areas in Communications*, vol. 35, pp. 1768–1782, 2017 (https://doi.org/10.1109/JSAC.2017.2710022).

[11] N. Michailow *et al.*, "Generalized Frequency Division Multiplexing for 5th Generation Cellular Networks", *IEEE Transactions on Communications,* vol. 62, pp. 3045–3061, 2014 (https://doi.org/10.1109/TCOMM.2014.2345566).

[12] A. Farhang, N. Marchetti, and L.E. Doyle, "Low-complexity Modem Design for GFDM", *IEEE Transactions on Signal Processing*, vol. 64, pp. 1507–1518, 2016 (https://doi.org/10.1109/TSP.2015.2502546).

[13] J. Abdoli, M. Jia, and J. Ma, "Filtered-OFDM: A New Waveform For Future Wireless Systems", *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Stockholm, Sweden, 2015 (https://doi.org/10.1109/SPAWC.2015.7227001).

[14] X. Zhang *et al.*, "Filtered-OFDM-enabler for Flexible Waveform in the 5th Generation Cellular Networks", *IEEE Global Communications Conference (GLOBECOM)*, San Diego, USA, 2015 (https://doi.org/10.1109/GLOCOM.2015.7417854).

[15] F.A.P. de Figueiredo *et al.*, "Comparing F-OFDM and OFDM Performance for MIMO Systems Considering a 5G Scenario", *Preprints*, 2019 (https://doi.org/10.20944/preprints201905.0307.v2).

[16] A. Idris *et al.*, "PAPR Reduction Using Huffman and Arithmetic Coding Techniques in F-OFDM System", *Bulletin of Electrical*

Engineering and Informatics, vol. 7, pp. 257–263, 2018 (https://doi.org/10.11591/eei.v7i2.1169).

[17] F. Di Stasio, M. Mondin, and F. Daneshgaran, "Multirate 5G Downlink Performance Comparison for F-OFDM and W-OFDM Schemes with Different Numerologies", *International Symposium on Networks, Computers and Communications (ISNCC)*, Rome, Italy, 2018 (https://doi.org/10.1109/ISNCC.2018.8530905).

[18] S.S.U. Shah, A.H. Sodhro, H.A. Baber, and M. Imran, "Implementing Enhanced MIMO with F-OFDM to Increase System Efficiency for Future 5G Cellular Networks", *International Journal of Communication Networks and Information Security*, vol. 10, pp. 403–409, 2018 (https://doi.org/10.17762/ijcnis.v10i2.3286).

[19] M.H. Mahmud *et al.*, "Performance Analysis of OFDM, W-OFDM and F-OFDM under Rayleigh Fading Channel for 5G Wireless Communication", *3rd International Conference on Intelligent Sustainable Systems (ICISS)*, Thoothukudi, India, 2020 (https://doi.org/10.1109/ICISS49785.2020.9316134).

[20] M. Liu, W. Xue, Y. Xu, and S.B. Makarov, "Design of Filters Based on Generic Function Model for Reducing Out-of-band Emissions of the F-OFDM Systems", *AEU – International Journal of Electronics and Communications*, vol. 139, art. no. 153908, 2021 (https://doi.org/10.1016/j.aeue.2021.153908).

[21] S. Sarker, L.A. Ara, T. Alam, and T. Debnath, "Design and Analysis of MIMO F-OFDM Systems for 5G and Beyond Wireless Communications", *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 10, pp. 203–210, 2021 (https://doi.org/10.35940/ijrte.B6274.0710221).

[22] D. Ali and Z.Z. Yahya, "An Experimental Study of F-OFDM Spectrum Efficiency for 5G Applications", *International Journal of Microwave and Optical Technology*, vol. 17, pp. 1–9, 2022.

[23] A.H. Babalola, O.A. Abdulkarim, S.A. Salihu, and T.O. Adebakin, "Performance Analysis of MIMO-OFDM Systems in 5G Wireless Networks", *Proc. of International Conference on Applied Informatics*, pp. 278–291, 2024 (https://doi.org/10.1007/978-3-031-75147-9_19).

[24] S.R. Thakre, "Optimal Filter Choice for Filtered OFDM", *3rd International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2019 (https://doi.org/10.1109/ICECA.2019.8821847).

[25] A.A. Sahrab and A.D. Yaseen, "Filtered Orthogonal Frequency Division Multiplexing for Improved 5G Systems", *Bulletin of Electrical Engineering and Informatics*, vol. 10, pp. 2079–2087, 2021 (https://doi.org/10.11591/eei.v10i4.3119).

[26] A. Sahin, I. Guvenc, and H. Arslan, "A Survey on Multicarrier Communications: Prototype Filters, Lattice Structures, and Implementation Aspects", *IEEE Communications Surveys & Tutorials*, vol. 16, pp. 1312–1338, 2013 (https://doi.org/10.1109/SURV.2013.121213.00263).

[27] F.J. Harris, "On the Use of Windows for Harmonic Analysis with The Discrete Fourier Transform", *Proceedings of the IEEE*, vol. 66, pp. 51–83, 1978 (https://doi.org/10.1109/PROC.1978.10837).

[28] S.M. Kay, *Fundamentals of Statistical Signal Processing: Detection Theory*, vol. 2., Englewood Cliffs, USA: Prentice Hall PTR, 576 p., 1998 (ISBN: 9780135041352).

[29] M.A. Taher, H.S. Radhi, and A.K. Jameil, "Enhanced F-OFDM Candidate for 5G Applications", *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, pp. 635–652, 2020 (https://doi.org/10.1007/s12652-020-02046-3).

Fadila Amel Miloudi, M.Sc.
Department of Electrical Engineering, Institute of Technology, Instrumentation Laboratory and Advanced Materials
https://orcid.org/0009-0002-5224-7833
E-mail: f.miloudi@cu-elbayadh.dz
University Center Nour Bachir, El-Bayadh, Algeria
https://www.cu-elbayadh.dz

Mohammed Sofiane Bendelhoum, Ph.D., Associate Professor
Department of Electrical Engineering, Laboratory of Electronic Systems, Telecommunications and Renewable Energies
https://orcid.org/0000-0002-9789-8712
E-mail: m.bendelhoum@cu-elbayadh.dz
University Center Nour Bachir, El-Bayadh, Algeria
https://www.cu-elbayadh.dz

Fayssal Menezla, Ph.D.
Department of Electrical Engineering, Laboratory LEPO
https://orcid.org/0009-0000-2909-7203
E-mail: f.menezla@cu-elbayadh.dz
Université Djillali Liabès of Sidi Bel-Abbès, El-Bayadh, Algeria
https://www.univ-sba.dz

Ridha Ilyas Bendjillali, Ph.D.
Department of Electrical Engineering, Laboratory of Electronic Systems, Telecommunications and Renewable Energies
https://orcid.org/0000-0003-2465-8192
E-mail: r.bendjillali@cu-elbayadh.dz
University Center Nour Bachir, El-Bayadh, Algeria
https://www.cu-elbayadh.dz

Performance Optimization of M/M/1 Queues with Working Vacations and Server Breakdowns in Wireless Communication Systems

S. Muthukumar[1,2], J. Ebenesar Anna Bagyam[1], and K. Basarikodi[2]

[1]*Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India,*
[2]*Ramco Institute of Technology, Rajapalayam, Tamil Nadu, India*

https://doi.org/10.26636/jtit.2025.3.2195

Abstract — This paper presents a unified analytical and simulation framework for optimizing the performance of M/M/1 queueing systems that incorporate differentiated working vacations, server breakdowns, and customer balking behavior. Other features of the solution include dynamical transitions between full-service mode, two levels of working vacation (with reduced service rates) phases, and random breakdown-repair cycles. Customers arrive via a Poisson process and decide to join or balk based on the server's current state. Embedded Markov chains, probability generating functions, and Matlab based discrete event simulation are applied to analyze key performance metrics, including average waiting time, queue length, and server utilization. A particle swarm optimization (PSO) algorithm is used to identify parameter configurations that minimize congestion and delay. Application scenarios in 5G/6G networks and service platforms demonstrate how adaptive vacation scheduling and resilience strategies improve energy efficiency and throughput. The results offer valuable information for performance tuning in resource-constrained telecommunication systems.

Keywords — 5G/6G networks, queueing performance analysis, server reliability, single-server queues, working vacations

1. Introduction

Modern wireless communication systems, including 5G, 6G and Internet of Things (IoT) networks, face increasing levels of demand for high-quality services, yet have limited computational and energy resources at their disposal. The dynamic nature of network traffic, combined with energy constraints and unpredictable system behavior, calls for deploying adaptive queueing mechanisms that are capable of managing service delays, optimizing resource usage, and ensuring operational resilience. The queue theory provides a robust analytical framework for addressing such challenges and has become a key tool in the process of modeling and optimizing telecommunication networks [1], [2].

Among the various queueing models, the working vacation queue, originally proposed in [3], has gained attention for its practical applicability in energy-aware systems. Unlike traditional vacation models where the server becomes completely idle, the working vacation model assumes that the server continues to operate at a reduced rate, closely resembling energy-saving or degraded service modes known from wireless infrastructures. This framework has been extended to include re-trials, server unreliability, customer impatience, and heterogeneous vacation behaviors [4], [5], improving its relevance to complex environments such as sensor networks and cloud-edge systems [6], [7].

Recent advances in network management include intelligent queue control in 5G, hybrid optimization models for traffic handling, and delay-tolerant service designs [8]–[10]. However, few studies jointly address such phenomena as server breakdowns, multiple working vacation phases, and balking behavior, especially under the constraints of wireless systems, where these conditions often coexist.

In our previous work [5], we analyzed the steady state behavior of an M/M/1 queueing system with differentiated working vacations and customer balking. This paper extends that model by introducing the following improvements:

- Two distinct working vacation phases, each with its own reduced service rate.
- Random server breakdowns and repair dynamics, modeling real-world hardware unreliability.
- State-dependent customer balking, where join/balk decisions are influenced by the server's operating mode.
- Integration of discrete-event simulation in Matlab to validate the analytical findings.
- Application of particle swarm optimization (PSO) to identify optimal system parameters while minimizing delays and improving performance metrics.

The inclusion of PSO in this study is a significant methodological enhancement. PSO is a robust, population-based metaheuristic that efficiently explores complex, non-linear search spaces and converges quickly. It is particularly well-suited for optimizing queueing systems with stochastic behavior, where traditional gradient-based methods may struggle. By applying PSO to tune service rates, vacation parameters, breakdown rates, and balking thresholds, this paper offers

Tab. 1. Model parameters and their descriptions.

Parameter	Description	Typical values
λ	Customer arrival rate (Poisson process)	Example: 1.0
μ	Service rate during busy period	Must satisfy $\mu > \mu_1 > \mu_2$
μ_1	Service rate during type I working vacation	Reduced service rate
μ_2	Service rate during type II working vacation	Further reduced service rate
γ_1	Probability of entering type I vacation	Example: 0.1
γ_2	Probability of entering type II vacation	Example: 0.05
α	Probability of server breakdown	Example: 0.03
β	Server repair rate after breakdown	Exponential repair rate
b_1	Balking probability during type I vacation	Set to 0.3 (heuristic)
b_2	Balking probability during type II vacation	Set to 0.5 (heuristic)
b	Balking probability during server breakdown	Set to 0.7 (heuristic)

an actionable optimization framework for improving system responsiveness and energy efficiency in dynamic wireless communication environments.

Using embedded Markov chains, probability-generating functions, and discrete event simulation, we analyze key performance indicators such as average waiting time, server utilization, and system throughput. Our results provide practical design insights for customer service platforms, edge computing nodes, and base stations in energy- and reliability-constrained systems.

2. Model Description

This study analyses a single-server M/M/1 queueing system that incorporates differentiated working vacations, server breakdowns, and customer balking behavior. The system is structured under the following assumptions:

- **Arrival process**. Customers arrive following a Poisson process with rate λ.
- **Service mechanism**. During regular busy periods, the server operates at a service rate of μ. In type I vacation mode, the server continues to operate at a reduced rate μ_1, and in type II vacation mode, at an even slower rate μ_2, where $\mu > \mu_1 > \mu_2$.

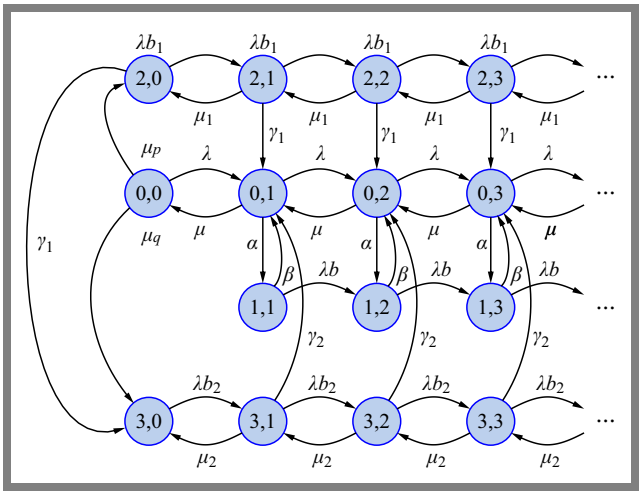

Fig. 1. State transition diagram.

- **Server behavior**. After completing a busy period, the server may enter a type I vacation with probability γ_1 or a type II vacation with probability γ_2.
- **Server breakdowns**. Random breakdowns occur with probability α, and the server undergoes repair at an exponential rate β.
- **Customer balking**. Customers may choose to balk depending on the server's state, with a 30% balking probability during type I vacation, 50% during type II vacation, and 70% during server breakdowns.

This model effectively captures real-world operational constraints such as fatigue, partial service availability, failure events, and customer impatience, making it highly relevant for the analysis and optimization of both traditional service systems and next-generation wireless networks.

2.1. Model Framework

This study builds on the queueing model presented in [5], which explored a steady-state M/M/1 system with differentiated working vacations, breakdowns, and balking. The current model retains the structural foundation of that system but introduces refined interpretations and supports simulations required for validating performance.

Let $N(t)$ denote the number of customers in the system at time t, and let $S(t) \in \{0, 1, 2, 3\}$ represent the server's state, where:

- $S(t) = 0$ – server is busy,
- $S(t) = 1$ – server is under breakdown,
- $S(t) = 2$ – server is on a type I working vacation,
- $S(t) = 3$ – server is on a type II working vacation.

The system is modeled as a continuous-time Markov process $S(t), N(t), t \geqslant 0$ with state space $\Lambda = \{(i, j) : i = 0, 1, 2, 3; j \geqslant 0\}$. Transition probabilities and the governing balance equations are retained from the earlier model, with minor notational refinements. For completeness, the main steady-state equations and performance metrics, including the expected number of customers in system and average waiting time, are summarized in Appendix A.

3. Practical Applications in Wireless Networks

This section demonstrates how the proposed M/M/1 queueing model with differentiated working vacations, server breakdowns, and customer balking can be effectively applied to modern service systems. Two representative domains are considered: customer service centers and next-generation wireless networks (NGWN).

3.1. Customer Service Centers and Call Centers

In service-oriented platforms like call centers, help desks or support chat systems, human agents act as servers processing customer requests. The proposed model offers several relevant analogies:

- **Server breakdowns** correspond to sudden unavailability of agents due to technical issues, fatigue, or shift changes.
- **Working vacations** represent scheduled breaks or periods of reduced service effort, e.g. multitasking, handling low-priority tasks.
- **Customer balking** models real-world impatience such as callers hanging up or users exiting queues when facing perceived delays.

By implementing the proposed model, organizations may design dynamic staffing policies to:

- Reduce call abandonment rates and improve response times.
- Optimize agent workload while avoiding burnout.
- Adapt service capacity based on real-time traffic.

Strategically timed low-effort periods (type I vacations) can preserve service quality while allowing recovery time, as long as breakdown probability and balking are carefully managed.

3.2. Next Generation Wireless Networks

In wireless systems, particularly 5G/6G networks, IoT gateways, and edge computing nodes, the model maps directly to network elements and protocols in the following manner:

- Servers represent base stations or edge nodes that process data packets or user requests.
- Working vacations correspond to energy savings or degraded operating modes during off-peak hours or congestion periods.
- Breakdowns simulate hardware failures, link failures, or shutdowns caused by overheating.
- Balking models packet drops or user session termination due to degraded quality of service (QoS).

The model enables adaptive resource allocation and energy-aware operations. It supports the following:

- Sleep/wake cycles in small cells or relay nodes for energy efficiency,
- Fault tolerance mechanisms through predictive repair and redundancy,

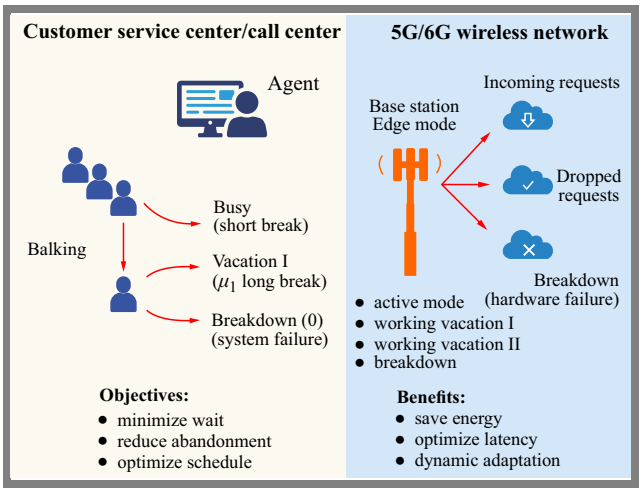

Fig. 2. Analogies of the proposed queueing model used in call centers and wireless network systems.

- Dynamic load balancing to reduce service delays and user drop rates.

Benefits for network operators include improved resilience and responsiveness, lower energy consumption without sacrificing throughput, scalable performance modeling for smart city infrastructure, vehicular networks, and cloud-edge orchestration.

4. Simulation Setup and Dataset Description

A simulation was performed to evaluate the performance of the system under the following parameters:

- Arrival rate $\lambda = 1.0$,
- Service rates: $\mu = 1.5$ (normal), $\mu_1 = 1.0$ (type I vacation), $\mu_2 = 0.5$ (type II vacation),
- Vacation probabilities: $\gamma_1 = 0.1$ (type I), $\gamma_2 = 0.05$ (type II),
- Breakdown probability $\alpha = 0.03$,
- Repair rate $\beta = 0.2$,
- Balking probabilities: $b_1 = 0.3$ (type I), $b_2 = 0.5$ (type II), and $b = 0.7$ (breakdowns).

The simulation ran for a total of 10 000 customer events, ensuring statistically significant results that capture typical system behavior under varying server states.

This representative subset highlights how customer arrivals, service commencement, and balking behavior are influenced by the server's state, providing practical insights into the system's dynamics.

5. Numerical Results and Validation

To evaluate the performance of the proposed M/M/1 queueing system with differentiated working vacations, server breakdowns, and customer balking, we developed a Matlab-based simulation. The simulation model mimics system behavior

Tab. 2. Sample simulation output (first five events).

Arrival time	Service start	Service end	Server state	Decision
0.47	0.47	1.08	Busy	Joined
0.64			Vacation I	Balked
2.65	2.65	2.66	Busy	Joined
6.15	6.15	6.29	Busy	Joined
6.36	6.36	6.73	Busy	Joined

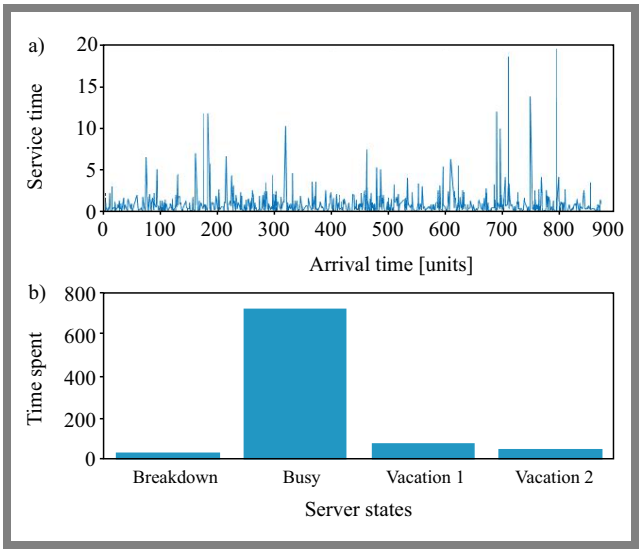

Fig. 3. Simulation outputs: a) service times for customers and b) cumulative time spent in different server states.

over 10 000 customer arrivals, ensuring statistical robustness and replicability. The key outputs are visualized in Fig. 3.

In Fig. 3a, the x-axis represents individual customer arrival times, while the y-axis shows their corresponding service durations. The figure illustrates how service times vary depending on server state: they are the shortest during busy periods, longer during type I vacations, and the longest during type II vacations or breakdown periods. Clusters of elevated service times reflect transitions into low-efficiency or failure states.

Figure 3b illustrates the time spent in different states. This bar chart quantifies the cumulative time spent in each server state (busy, vacation I, vacation II, breakdown). It clearly shows how working vacations and breakdowns reduce effective service capacity, helping identify bottlenecks and guide parameter tuning (e.g. reducing γ_2 or improving β).

These visualizations offer actionable insights into system dynamics and underscore the impact that server fatigue and failures exert on customer experience. The simulation provides a practical validation layer to the analytical model, confirming its relevance for real-world service systems and telecommunication networks, where non-ideal behaviors like balking and degradation are common.

5.1. Simulation Environment

All simulations, including queue dynamics and particle swarm optimization (PSO), were implemented in Matlab R2023b. Matlab's built-in functions and custom scripts were used to model queue states, implement PSO algorithms, and generate figures. Random number seeds were set to ensure consistency across repeated runs.

6. Optimization Framework Using PSO

Particle swarm optimization is employed to minimize cost functions and optimize key performance metrics, expected queue length, waiting time, and server utilization, in an M/M/1 queue with working vacations, breakdowns, and customer balking. PSO is particularly suitable for this problem due to its fast convergence, simplicity, and robustness in non-linear, high-dimensional search spaces. Compared to other meta-heuristics (e.g. GA, ACO), PSO requires fewer parameters and is more computationally efficient.

The objective of the PSO algorithm is to minimize cost function C that represents the trade-off between queueing performance metrics, i.e. the average number of customers in the system, and the average waiting time.

The cost function is defined as:

$$C = w_1.E(L) + w_2.E(W)\,, \tag{1}$$

where:

- $E(L)$ – expected number of customers in the system,
- $E(W)$ – expected waiting time from Little's law: $E(W) = \frac{E(L)}{\lambda}$,
- $w_1, w_2 \in [\,0, 1\,]$ are user-defined weights such that: $w_1 + w_2 = 1$, reflecting the relative importance of queue length and delay.

For example, setting $w_1 = w_2 = 0.5$ gives equal weight to both performance criteria.

The PSO algorithm searches for the parameter set $\theta = [\,\mu_1, \mu_2, \gamma_1, \gamma_2, \alpha, \beta, b_1, b_2, \text{and}\, b\,]$ that minimizes this cost function:

$$\theta^* = \arg\min_{\theta} c(\theta)\,. \tag{2}$$

It is defined based on parameters such as arrival/service rates $\lambda, \mu, \mu_1, \mu_2$, breakdown and repair rates α, β, vacation probabilities γ_1, γ_2, and balking probabilities b_1, b_2, and b.

Initialization of the algorithm includes swarm size, inertia weight w, and coefficients c_1, c_2. Parameter bounds are defined, e.g., $\mu_1 \in [\,0.1, \mu\,]$. Each particle encodes queue parameters:

$$Particle_i = [\mu_1, \mu_2, \gamma_1, \gamma_2, \alpha, \beta, balking\ par, \ldots]\,. \tag{3}$$

At iteration $t+1$, particle velocities and positions are updated as:

$$v_i^{t+1} = w\, v_i^t + c_1\, r_1 (pBest_i - x_i^t) + c_2\, r_2 (gBest - x_i^t)\,, \tag{4}$$

$$x_i^{t+1} = x_i^t + v_i^{t+1}\,, \tag{5}$$

Tab. 3. Baseline vs. optimized performance.

Metric	Baseline	PSO optimized	Improvement
Average queue length	6.12	3.48	↓~43.1%
Average waiting time [units]	5.01	2.23	↓~55.5%
Server utilization	78.4%	85.7%	↑ +7.3%

where $r_1, r_2 \sim U(0,1)$ are random factors, $pBest_i$ is the best position of particle i, and $gBest$ is the global best found so far. The algorithm terminates when a maximum number of iterations is reached or the improvement falls below a threshold.

6.1. PSO Results

To assess the impact of the PSO algorithm, we compared performance metrics before and after optimization (Tab. 3). The PSO was applied to tune key parameters, e.g. $\mu_1, \mu_2, \gamma_1, \gamma_2, \alpha, \beta$, and balking probabilities, with the objective of minimizing average queue lengths and waiting times.

These results confirm that PSO effectively identifies superior parameter configurations reducing congestion and improving response times.

The plot shown in Fig. 4 shows a smoothed trend of service durations aligned with arrival times. It illustrates how optimized parameters help maintain service times within a tighter range, thus avoiding spikes observed under baseline settings. This uniformity leads to greater predictability and reduced customer waiting time.

6.2. PSO Parameter Settings and Cost Convergence

To validate the performance of the PSO algorithm in optimizing the M/M/1 queue system with working vacations and server breakdowns, Tab. 4 summarizes the parameter settings used in the simulation.

The cost function used in optimization reflects a weighted combination of average queue length and average waiting time. These metrics capture the overall efficiency of the system and customer satisfaction under varying conditions, for example, vacations, breakdowns, and balking.

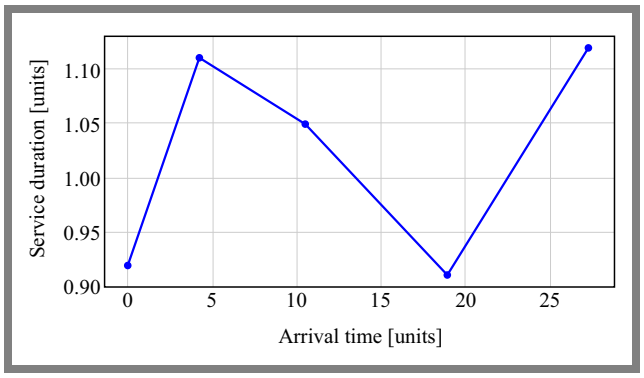

Fig. 4. Optimized service times across arrival times.

Tab. 4. PSO parameter settings.

Parameter	Value	Description
Swarm size	30	Number of particles in the swarm
Number of iterations	100	Maximum iterations of the algorithm
Inertia weight w	0.7	Controls influence of previous velocity
Cognitive coefficient c_1	1.5	Weight toward personal best position
Social coefficient c_2	1.5	Weight toward global best position

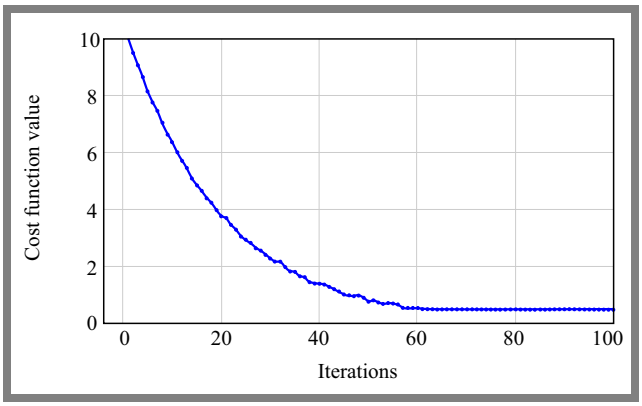

Fig. 5. Convergence of the PSO cost function.

A cost convergence plot was generated to visualize the performance of the PSO algorithm in the individual iterations. As shown in Fig. 5, the cost steadily decreases and stabilizes as the algorithm converges toward an optimal parameter set.

The convergence curve demonstrates that the algorithm reliably reduces the cost within 100 iterations, indicating successful optimization of the queue parameters under the defined constraints.

7. Conclusions

This article presents an enhanced M/M/1 queueing model that integrates differentiated working vacations, server breakdowns, and state-dependent customer balking, offering a realistic framework for analyzing and optimizing performance in modern telecommunication systems. The model captures essential operational dynamics often encountered in environments such as call centers, wireless base stations, and edge computing nodes, where service degradation, unreliability, and impatient user behavior are common.

The model demonstrates significant improvements in performance, offers a reduction in average waiting time (up to 55%) and a 43% decrease in queue length. These results underscore the potential of intelligent queue management and metaheuristic optimization approaches, as these are capa-

ble of improving responsiveness and energy efficiency in resource-constrained networks. The proposed framework offers actionable insights for designing adaptive scheduling policies, optimizing energy usage, and enhancing user satisfaction in 5G/6G, IoT gateways, and customer-facing service platforms.

Although PSO has proven effective in optimizing the proposed queueing model due to its fast convergence and simplicity, it is beneficial to briefly consider alternative metaheuristic approaches. Genetic algorithms (GAs), for instance, offer robustness and flexibility, particularly for discrete optimization problems, but often require more computational effort and parameter tuning.

Reinforcement learning (RL), on the other hand, enables adaptive learning in dynamic environments and is well-suited for online decision making. However, its applicability is limited in settings where training data is sparse or where system states evolve slowly.

PSO was selected for this study because of its ease of implementation, compatibility with simulation-based optimization, and lower computational overhead in static system scenarios. Future comparative studies may further explore the trade-offs among these techniques to guide optimization choices across various application domains.

Acknowledgments

Data supporting the findings of this study were reported in our previous publication [5]. Additionally derived data supporting this study are available from the authors upon reasonable request.

Appendix A.
Local Balance Equations

Let $p_{i,j}$ denote the steady-state probability of being in server state i with j customers in the system.

The balance equations for the continuous-time Markov process are as follows:

$$(\lambda+\mu)\,p_{0,0} = \mu\,p_{0,1}\ , \tag{6}$$

$$\begin{aligned}(\lambda+\mu+\alpha)\,p_{0,n} = {}&\gamma_1\,p_{2,n}+\gamma_2\,p_{3,n}+\beta\,p_{1,n}\\ &+\mu\,p_{0,n+1}+\lambda\,p_{0,n-1},\ n\geqslant 2\ ,\end{aligned} \tag{7}$$

$$(\lambda b+\beta)\,p_{1,1} = \alpha\,p_{0,1}\ , \tag{8}$$

$$(\lambda b+\beta)\,p_{1,n} = \alpha\,p_{0,n}+\lambda b\,p_{1,n-1},\ n\geqslant 1\ , \tag{9}$$

$$(\lambda b_1+\gamma_1)\,p_{2,0} = \mu_1\,p_{2,1}+\mu p\ p_{0,0}\ , \tag{10}$$

$$(\lambda b_1+\gamma_1+\mu_1)\,p_{2,n} = \mu_1\,p_{2,n+1}+\lambda b_1\,p_{2,n-1},\ n\geqslant 1\ , \tag{11}$$

$$\lambda b_2\,p_{3,0} = \gamma_1\,p_{2,0}+\mu q\,p_{0,0}+\mu_2\,p_{3,1}\ , \tag{12}$$

$$(\lambda b_2+\gamma_2+\mu_2)\,p_{3,n} = \lambda b_2\,p_{3,n-1}+\mu_2\,p_{3,n+1},\ n\geqslant 1\ , \tag{13}$$

Generating functions

Using corresponding probability generating functions

$$P_i(z) = \sum_{n=0}^{\infty} p_{i,n}z^n$$

the expressions are:

$$\begin{aligned}P_0(z) = {}&\frac{(\mu-\alpha z)\,p_{0,0}-\big[P_1(z)]\beta z+[p_{2,0}-P_2(z)\big]\gamma_1 z}{\lambda z^2-(\lambda+\mu+\alpha)z+\mu}\\ &+\frac{\big[p_{3,0}-P_3(z)\big]\gamma_2 z}{\lambda z^2-(\lambda+\mu+\alpha)z+\mu}\ ,\end{aligned} \tag{14}$$

$$P_1(z) = \frac{\alpha\big[p_{0,0}-P_0(z)\big]}{\lambda b z-(\lambda b+\beta)}\ , \tag{15}$$

$$P_2(z) = \frac{\mu_1\,p_{2,0}-\mu_1 z\,p_{2,0}-\mu p\,z\,p_{0,0}}{\lambda b_1 z^2-(\lambda b_1+\mu_1+\gamma_1)z+\mu_1}\ , \tag{16}$$

$$P_3(z) = \frac{\mu_2\,p_{3,0}-z\,p_{3,0}(\mu_2+\gamma_2)-\gamma_1 z\,p_{2,0}-\mu q z\,p_{0,0}}{\lambda b_2 z^2-(\lambda b_2+\mu_2+\gamma_2)z+\mu_2}\ , \tag{17}$$

Steady-state probability at idle state

$$p_{0,0} = \frac{\gamma_1\gamma_2}{\gamma_1\gamma_2+(\gamma_2\,\mu\,p+\gamma_1\gamma_2\,s+\gamma_1^2\,t+\gamma_1\,\mu\,q)}\ , \tag{18}$$

where:

$$s = \frac{\mu p\,z_1}{\mu_1(z_1-1)}\ ,$$

$$t = -\frac{z_2\,(s\gamma_1+\mu q)}{z_2\,(\gamma_2+\mu_2)-\mu_2}\ ,$$

and:

z_1 is the positive root of

$$\lambda b_1 z^2-(\lambda b_1+\mu_1+\gamma_1)z+\mu_1\ ,$$

z_2 is the positive root of

$$\lambda b_2 z^2-(\lambda b_2+\mu_2+\gamma_2)z+\mu_2\ ,$$

Expected system lengths by state

Busy state:

$$\begin{aligned}E(L_B) = {}&\frac{\gamma_1\gamma_2\,\lambda b_2\,p_{3,0}+[\gamma_1\gamma_2\,\mu_1+\gamma_1^2(\lambda b_2-\mu_2)]\,p_{2,0}}{\gamma_1\gamma_2\,\alpha}\\ &+\frac{[\gamma_2\,\mu p(\lambda b_1-\mu_1)+\gamma_1\,\mu q(\lambda b_2-\mu_2)+\gamma_1\gamma_2\lambda]\,p_{0,0}}{\gamma_1\gamma_2\,\alpha}\ ,\end{aligned} \tag{19}$$

Breakdown state:

$$E(L_1) = \frac{\alpha\,\big[E(L_B)-p_{0,0}\big]}{\beta}\ , \tag{20}$$

Type-I vacation:

$$E(L_2) = \frac{\big[\mu p\,(\lambda b_1-\mu_1)\big]\,p_{0,0}+\gamma_1\mu_1 p_{2,0}}{\gamma_1^2}\ , \tag{21}$$

Type-II vacation:

$$E(L_3) = \frac{\gamma_2 \lambda b_2\, p_{3,0} + \gamma_1(\lambda b_2 - \mu_2)\, p_{2,0} + \mu q(\lambda b_2 - \mu_2)\, p_{0,0}}{\gamma_2^2}, \tag{22}$$

Expected waiting time

Using Little's law:

$$E(W) = \frac{E(L_B) + E(L_1) + E(L_2) + E(L_3)}{\lambda}. \tag{23}$$

References

[1] S.A. Afolalu *et al.*, "A Short Review on Queuing Theory as a Deterministic Tool in Sustainable Telecommunication System", *Materials Today: Proceedings*, vol. 44, pp. 2884–2888, 2021 (https://doi.org/10.1016/j.matpr.2021.01.092).

[2] A. Roy, J.L. Pachuau, and A.K. Saha, "An Overview of Queuing Delay and Various Delay Based Algorithms in Networks", *Computing*, vol. 103, pp. 2361–2399, 2021 (https://doi.org/10.1007/s00607-021-00973-3).

[3] L.D. Servi and S.G. Finn, "M/M/1 Queues with Working Vacations (M/M/1/WV)", *Performance Evaluation*, vol. 50, pp. 41–52, 2002 (https://doi.org/10.1016/S0166-5316(02)00057-3).

[4] J.E.A. Bagyam, P. Suganthi, and P. Visali, "Unreliable Single Server Retrial Queueing Model with Repeated Vacation", *International Journal of Mathematics in Operational Research*, vol. 26, pp. 357–372, 2023 (https://doi.org/10.1504/IJMOR.2023.134838).

[5] S. Muthukumar and J.E.A. Bagyam, "A Steady-state Behavior of an M/M/1 Queue with Optional Differentiated Working Vacations, Server Breakdown, and Customer Balking", *Advances and Applications in Statistics*, vol. 92, pp. 603–631, 2025 (https://doi.org/10.17654/0972361725025).

[6] G. Bouloukakis, I. Moscholios, N. Georgantas, and V. Issarny, "Performance Analysis of Internet of Things Interactions via Simulation-based Queueing Models", *Future Internet*, vol. 13, art. no. 87, 2021 (https://doi.org/10.3390/fi13040087).

[7] X. Li, "Performance Evaluation of M/M/1 Queueing Models in Cloud Computing Environments", *Highlights in Science, Engineering and Technology*, vol. 85, pp. 841–848, 2024 (https://doi.org/10.54097/1xcw8y74).

[8] S. Jung, J. Kim, and J.H. Kim, "Intelligent Active Queue Management for Stabilized QoS Guarantees in 5G Mobile Networks", *IEEE Systems Journal*, vol. 15, pp. 4293–4302, 2020 (https://doi.org/10.1109/JSYST.2020.3014231).

[9] A.A. Rezaee and F. Pasandideh, "A Fuzzy Congestion Control Protocol Based on Active Queue Management in Wireless Sensor Networks with Medical Applications", *Wireless Personal Communications*, vol. 98, pp. 815–842, 2018 (https://doi.org/10.1007/s11277-017-4896-6).

[10] Z. Zhang *et al.*, "A Particle Swarm Optimization-based Queue Scheduling and Optimization Mechanism for Large-scale Low-Earth-orbit Satellite Communication Networks", *Sensors*, vol. 25, art. no. 1069, 2025 (https://doi.org/10.3390/s25041069).

S. Muthukumar, M.Phil., Assistant Professor
Department of Mathematics
https://orcid.org/0009-0007-1320-6082
E-mail: muthukumarmsc99@gmail.com
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
https://kahedu.edu.in
Ramco Institute of Technology, Rajapalayam, Tamil Nadu, India
https://www.ritrjpm.ac.in

J. Ebenesar Anna Bagyam, Ph.D., Assistant Professor
Department of Mathematics
https://orcid.org/0000-0002-9943-5893
E-mail: ebenesar.j@gmail.com
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, India
https://kahedu.edu.in

K. Basarikodi, Ph.D., Professor
Department of Mathematics
https://orcid.org/0000-0002-2668-7096
E-mail: basarikodi@ritrjpm.ac.in
Ramco Institute of Technology, Rajapalayam, Tamil Nadu, India
https://www.ritrjpm.ac.in

A Comprehensive Study on Path Loss Estimation Using Deep Hybrid Learning in 5G Networks

Kazi Md Abrar Yeaser[1] and Kazi Md Abir Hassan[2]

[1]*Premier University, Chittagong, Bangladesh,*
[2]*Islamic University of Technology, Gazipur, Bangladesh*

https://doi.org/10.26636/jtit.2025.3.2100

Abstract — One of the most important factors in radio network design is path loss – a phenomenon that may be measured using a variety of techniques, including deterministic, empirical, machine learning, and deep learning models. Each approach has its own limitations, such as inability to capture non-linear interactions, high computational resource demand, and inability to reflect changes in environmental conditions, among many others. The deep learning model has the capacity to recognize intricate patterns and has been essential in removing those obstacles; therefore, in this study it is used for path loss prediction in 5G communications in the South Asian region. The model makes use of long- and short-term memory (LSTM), gated recurrent unit (GRU), convolutional neural network (CNN), and dense neural network (DNN) approaches to take advantage of all the benefits that each algorithm provides. The performance of the proposed strategy was validated by testing it against multiple state-of-the-art approaches, while relying on the same dataset. An examination of the relevance of characteristics has also been carried out to gain a better understanding of the influence of path loss. A variety of characteristics that are directly related to path loss were evaluated, followed by an examination of how they affect the decision-making process. The results show a possible solution that can help handle this path loss estimation for mmWave communication, especially for 5G networks and beyond.

Keywords — 5G, deep learning, machine learning, mmWave, path loss

1. Introduction

5G networks use higher frequency and smaller cell sizes, rendering the issues of signal degradation, fading, and interference more important [1], [2]. Signal strength is the most important parameter for maintaining a reliable communication link and determines throughput. Path loss refers to the degradation of the electromagnetic signal as it propagates through a channel [3]. Mathematically, it is the difference between the transmitting power and the receiving power of a signal. Knowledge of path loss in a given environment makes it easier to efficiently plan radio networks [4]. Path loss measurement can help optimize power usage according to channel conditions. Additionally, knowledge about the path loss of a channel may greatly improve the quality of service and resource allocation.

There are several methods to measure path loss. The early methods involve empirical models developed based on data observed under real-world scenarios [5]. Several parameters such as distance, frequency, and attenuation factor are taken into account when developing specific formulas. The models capture the average path loss value of a channel in a certain setting. There are several empirical models such as the free space path loss model [6], the Hata model [7], the Okumura model [8], the 3GPP TR 38.901 model [9], and the log distance model [10]. Simplicity is one of the main reasons behind their widespread use.

They are pretty basic models that require some simple parameters, like distance, frequency, and some environmental data that change based on a given setting. Consequently, it is very easy to modify the model according to environmental needs. For example, the Hata model can be adopted in urban, suburban, and rural settings. Its simplicity eliminates the need for computational resources and makes it an economical option. Its ease of use and thorough understanding have led to its adoption as the main foundation of radio network planning [10].

3GPP has developed one such empirical model, known as 3GPP TR 38.901, keeping in mind the nature of 5G communication and its requirements. However, there are several major limitations that led to the adoption of other techniques. One of the major limitations is their rigidity. Although the models capture some of the environmental parameters, they cannot reflect a sudden change of a certain parameter. Also, the models are much more generalized. Consequently, they cannot truly capture the difference in settings that vary from one country to another.

The urban setting prevailing in Europe does not necessarily reflect the conditions that exist in Asia. Therefore, path loss measurements may not be accurate. Additionally, changes in some parameters may cause significant changes in path loss readouts.

In some communication modes, such as vehicular communication, the parameters change rapidly, making the path loss models inconsistent in such scenarios. Also, some of the empirical models require the transmitter and receiver to be in the line-of-sight setting, which reduces their usability in non-line-of-sight environments.

Statistical analysis is another approach that is similar to the empirical method. Instead of relying on curve fitting based on real-world data in the empirical method, the statistical method uses probabilistic and statistical analysis to model the propagation of electromagnetic waves. Many statistical methods such as log distance path loss model and log normal shadowing [11] exist.

One of the major limitations of the method is its inability to model optical phenomena such as diffraction, reflection, and scattering. These phenomena are very common in high-frequency environments, which makes the statistical method ineffective in high-frequency cases.

Another method of measuring path loss relies on deterministic models [12]. These models make use of electromagnetic principles to predict the propagation of signals through the environment based on their interaction with several environmental factors [13]. Instead of the observed values, these models are developed by simulating the real world environment. As a result, they are environment-specific and yield measurements with a higher degree of accuracy.

There are several deterministic models in use. Ray tracing [14] is the most popular deterministic model. It tracks the rays from the transmitter to the receiver and finds out how all the ray components interact with the environment. Several parameters such as diffraction, scattering, and reflections are taken into account based on the shape and materials of a given object. The model is used mainly in urban and indoor environments.

Ray launching [15] follows a similar principle of simulating the multipropagation of rays from the transmitter to the receiver. However, it does not trace each ray, which makes it faster. It is mainly used where there is a trade-off between computational resources and accuracy.

It needs to be borne in mind there several other models, such as the uniform theory of diffraction (UTD) [16] and the finite difference time domain (FDTD) [17] exist. These, however, suffer from some disadvantages. The high computational resource requirement makes them a costly option. Due to the high computational volume, they require more time for processing and do not perform accurately in a complex environment where the parameters change rapidly. The high dependence on environmental factors makes them very sensitive to small-scale variations.

Another deterministic method that solves the problem is the parabolic equation method. Unlike the UTD and FDTD, it allows for wave modeling in one main direction only. The method greatly reduces the computation load, as it ignores the backward waves. However, it finds limited use in the near-field region and, like other methods, it also lacks accuracy when the parameters change abruptly.

The geometry method is based on the same foundation as the deterministic model, i.e. it measures path loss by estimating the interaction of the wave with various environmental factors. However, the geometry-based method introduces a statistical method to simulate multipath effects. It is a hybrid method that utilizes the statistical method while also relying on physical accuracy. However, the need for detailed environment data makes its use complex in larger areas. Also, the computational complexity is very high in this case.

Today, artificial intelligence has gained greater traction in every aspect of engineering [18]. Due to its ability to understand complex patterns and make decisions, it is also relied upon in path loss measurements. Machine learning [19] is one of the subsets of artificial intelligence. Machine learning (ML) makes estimations based on previously observed data [20]. It uses several algorithms to find a common pattern among the various features that may influence path loss and makes decisions based on those determinations.

Several algorithms, such as linear regression [21], support vector regression [22], and decision tree [23] are used. However, machine learning lacks the ability to capture non-linear relationships. Deep learning (DL) networks have become very useful in this regard. It is a subset of machine learning that has been constructed to mimic the operation of a human brain [24]. The inclusion of neurons, layers, and activation functions enables deep learning algorithms to capture non-linear relationship as well [25].

In this paper, a deep hybrid model is proposed to estimate path loss. The model consists of long short-term memory (LSTM), a gated recurrent unit (GRU), a convolution neural network (CNN), and a dense neural network. The model was trained using data tailored for the South Asia region. The model was fed with several parameters, such as distance between transmitter and receiver, time delay, received power, phase, azimuth angle of departure, azimuth angle of arrival, elevation angle of departure, elevation angle of arrival, frequency, season, phase, and RMS delay spread. The model explores the three distinct algorithms to take advantage of all of their functionalities. Along with the estimation, the importance of the features and their influence on estimating path loss have been explored.

Table 1 shows the several methods and their limitations in estimating path loss. It is evident from the table that deep learning algorithms can estimate path loss more accurately compared to the deterministic model, and at a lower cost. But the high data requirement is hurdle affecting its adoption. The proposed hybrid model may offer a potential solution to the problem.

Our contributions are as follows.

1) Combining several deep learning algorithms to develop hybrid models for the estimation of path loss in the South Asia region. Instead of focusing only on one kind of algorithm, we have combined several algorithms like LSTM, GRU, CNN, and DNN to capture both temporal and spatial dependencies while predicting path loss.
2) Conducting a comparative study benchmarking the solution against other commonly used algorithms to validate the performance of the proposed hybrid model.
3) Interpreting the model's decision-making process by studying the impact of each feature utilized in the model.
4) Investigating the model's ability to detect path loss to boost it in real world scenarios.

Tab. 1. Comparison of various path-loss models.

Method	Advantage	Limitation
Empirical	Simple and fast	Poor generalization
Deterministic	Very accurate capture of several physical effects	Computationally demanding
Statistical	Scalable	Fixed distribution
Geometry based	Balance between physical realism and efficiency	Need of detailed environment info
Machine learning	Ability to capture non-linear relationship	Requires large amounts of labeled data
Deep learning	Very high accuracy	Computationally demanding

2. Literature Review

Due to the superior performance of machine learning and deep learning algorithms in recognizing the relationship between path loss and various factors, several researchers have explored different approaches based on these models.

2.1. Machine Learning-based Approaches

Several researchers have used machine learning algorithms to estimate path loss. While evaluating the best models, almost all commonly used algorithms have been tested, but the best-performing solutions varied depending on a specific environment. AdaBoost was found to show superior performance in tropical regions [26], random forests showed better performance in the region of uneven terrain attributes [27], and gradient tree boosting for millimeter wave communication (mmWave) communication was best suited for indoor environments [28].

Several researchers adopted numerous performance enhancement steps during the data preparation and training stages, instead of relying on the algorithm alone. In [29], before moving on to support vector machine-based model, dimensionality reduction techniques – such as principal component analysis – were employed to lower the use of computational resources. In another work, support vector regression was relied upon to reduce complexity and training time with different kernels to find the optimal model [30].

Various machine learning algorithms such as AdaBoost and random forest were employed to find the best model for predicting path loss in aircraft cabins [31]. The data expansion method generating partial data samples using the empirical approach has also been adopted to achieve further prediction accuracy improvements. Instead of relying on one specific algorithm, an ensemble model named voting regression was proposed. It consisted of k-nearest neighbors (KNN), support vector regression (SVR), random forest (RF), AdaBoost, and gradient tree boosting (GTB) algorithms to improve its performance [32].

2.2. Neural Network-based Approaches

Due to their ability to capture complex patterns, deep learning approaches have gained momentum, replacing machine learning algorithms where the availability of data is not a problem. Several researchers have also used deep learning-based approaches to estimate path loss. Artificial neural networks (ANN) are one of the most commonly used solutions. An ANN has been used to build path loss prediction models in corridor environments with varying frequencies [33].

Two types of ANN (multilayer perception (MLP) and radial basis function (RBF)) were used to model path loss of an ultra wide-band channel in a mine environment [34]. The model was designed to focus on the balance between generalization and precision. In [35], an ANN-based model was adopted to predict path loss in a multi-wall, multi-frequency indoor environment. The model is based on MLP and the training of data follows the backpropagation algorithm.

Instead of simply using ANN, some researchers conducting data preprocessing and training stages to increase the level of accuracy. In [36], an ANN-based model was deployed to predict path loss in urban environments. To optimize the ANN model and adapt it to a specific problem, an adaptive differential evolution algorithm named CoDe was used. The authors of [37] used ANN to predict path loss for very high-frequency wireless communication. In the study, extensive analysis has been performed to find the optimal numbers of input parameters, neurons, activation functions, and learning algorithms. MLP combined with ADALINE was used to predict the loss of signal propagation in microcellular urban environments [38].

Just like it was the case with machine learning approaches, rather than depending on one type of algorithm, researchers utilized various algorithms with ANN to build more robust models. The authors of [39] proposed a two-layer RBF neural network-based model. It predicted path loss using hybrid rival penalized competitive learning (RPCL) and recursive least squares (RLS) algorithms. The model offered better performance compared to empirical approaches such as the data model. In [40], field strength was predicted using a combination of an empirical model and an artificial neural network. The research was based on a dense urban environment.

A hybrid model was developed using the Hata model and low complexity ANN to predict path loss in [41]. The model outperformed a high-complexity ANN model by accurately predicting path loss.

Another approach based on neural networks is the backpropagation neural network. In [42], a backpropagation neural network was used to predict the received power in a suburban scenario, while in paper [43], backpropagation neural networks were used to build a model that can be useful in multiple environmental settings (rural, urban, and suburban). In addition to the ANN and backpropagation neural network, other types of networks such as the 3-layer wavelet neural

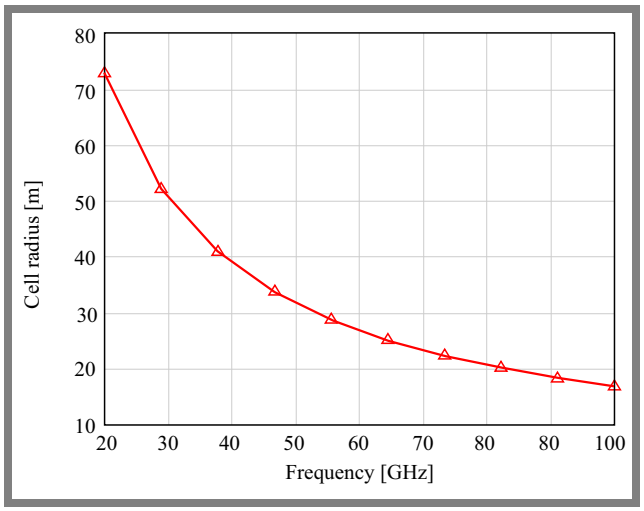

Fig. 1. Impact of using higher frequencies on cell radius.

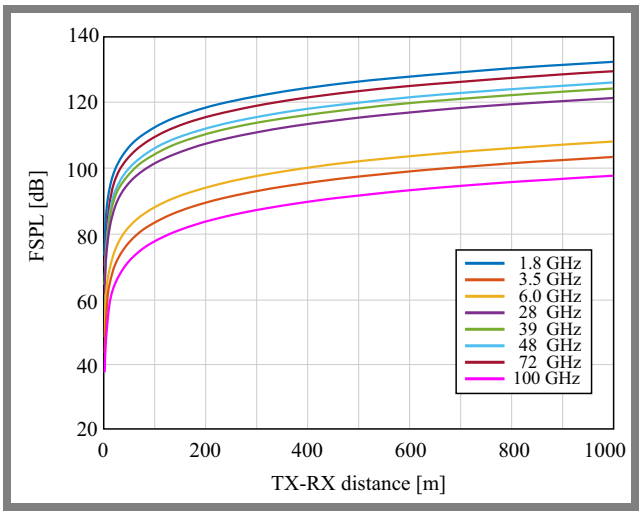

Fig. 2. Impact of using higher frequencies on path loss.

network have been developed to predict field strength for different frequencies [44].

CNN is another prominent deep learning-based approach. It has been used extensively for model building. In [45], two convolutional neural networks based on group-16 visual geometry (VGG) and residual network (ResNet-50) were tested. ResNet-50 was found to be the best performing solution. Multitask learning was introduced to further enhance the accuracy of the predictions made. The introduction of multitask learning increased the accuracy rate to 2–4%. In [46], a CNN-based model was proposed to predict the path loss exponent of outdoor millimeter wave band channels. In [47], the authors used CNN to build a path loss prediction model for high-traffic scenarios. The environment has various obstacles that greatly impact the communication using high-frequency bands, and it makes it very difficult for conventional methods to accurately estimate path loss.

3. Problem Analysis

5G networks use millimeter wave (mmWave) spectrum (24 to 100 GHz) to facilitate higher capacity and ultra-low latency. The use of higher frequencies allows to support massive device connectivity. However, using higher frequencies comes with its disadvantages too. One of the key challenges is the reduction of cell size. Figure 1 shows the impact of frequency increment on the radius of the cell. As one may notice, at 20 GHz the cell radius is marginally higher than 70 m. As we continue to increase the frequency even further, the cell radius declines sharply. At 100 GHz, the cell radius reduces to less than 20 m. A lower cell radius will result in frequent cell switching events that greatly impact path loss and throughput. When employing the free space path loss model, it can be seen that higher frequency has a significant impact on path loss. The free space path loss (in decibels) can be expressed as:

$$FSLP = 20\log_{10} d + 20\log_{10} f + 20\log_{10}\frac{4\pi}{c}, \qquad (1)$$

where d is the separation between the transmitter (Tx) and the receiver (Rx), f is the operating frequency, and speed of light is denoted as c. Figure 2 shows the impact of higher frequencies on path loss.

As the number of steps increases, the path loss also increases. At 1000 m separation, using the 1.8 GHz frequency results in a path loss that is closer to 90 dB. Using higher frequencies results in greater path loss. As one may see from the figure, an operating frequency of 100 GHz causes a path loss exceeding 120 dB at 100 m separation. Consequently, path loss estimation is of greater importance in 5G networks, as it allows for efficient radio network planning.

4. Methodology

The proposed hybrid method consists primarily of LSTM, GRU, and CNN layers. While LSTM layers are used to capture temporal dependencies, convolution layers are used for capturing spatial dependencies. GRU layers are used here instead of a stack of LSTM layers to reduce computation requirements. In Fig. 3, the working process is shown in the form of a block diagram. The process involves collection of the data set, preprocessing, designing, training and evaluating the model, and then comparing it with baseline ML-DL algorithms.

The data set has been based on a 5G communication environment in the South Asia region, as described in [48]. The dataset contains multiple data which were obtained through a simulation relying on NYUSIM, but only those variables that are closely related to and can be used to predict path loss are considered in this research. These include the following: transmitter-receiver (T-R) separation distance, time delay, received power, RMS delay spread, and frequency. As far as frequency is concerned, the dataset is mainly focused on the high band and the frequencies used here are 7.125, 24.25, 52.6, and 71 GHz.

The process of preparing raw data trainable for the deep learning model is important, since it is closely related to finding the best outcome from the prediction model. The

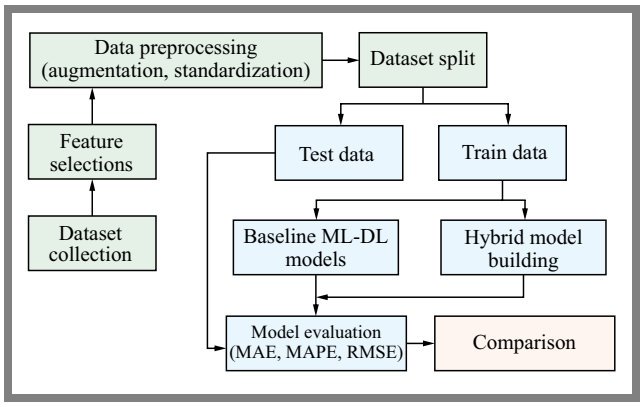

Fig. 3. Block diagram depicting the methodology used.

cleaner the data set, the better outcome can be obtained from the prediction model; thus, it is an important phase before training the model. Two stages are developed:

- Data augmentation is a method through which newer artificial data can be created. Deep learning algorithms require a robust larger dataset to build a general model. Data augmentation helps in that regard to create more samples and increase the size of the dataset. Several methods are used, such as adding noise and transformation. One of such methods is bootstrapping. It creates new data using a method known as "resample with replacement". The primary reasons for choosing bootstrapping for data augmentation are its non-parametric character and flexibility. Bootstrapping not only allows to increase the size of the dataset, but also helps generalize the model, making it more robust to noise.
- Standardization. As several values have an outlier effect and fail to follow normal distribution, we have applied min-max scaler to all variables. This will augment convergence and prevent any bias caused by the outliers.

$$x_{transformed} = \frac{x - x_{min}}{x_{max} - x_{min}}\,, \tag{2}$$

where x is the value of a feature, x_{min} and x_{max} are the minimum and maximum values of the feature, respectively.

4.1. Proposed Deep Hybrid Model

The deep learning model has to be built in such a way that it is able to take into account every aspect of the dataset and can predict accordingly. To build a solution that fulfills this requirement, a hybrid model turns out to be the best approach, as it combines different deep learning algorithms. In this research, a hybrid model is built by combining LSTM, GRU, CNN, and a dense layer, as shown in Tab. 2. It is well known that LSTM is a resource-intensive model, as it can be expanded and may utilize large datasets to predict better outcomes.

After this layer, GRU is utilized. It is less resource-intensive and more efficient in producing a better outcome, as GRU is used for capturing the temporal dependencies.

However, it is much simpler than LSTM, which results in faster training. After GRU, a dilated convolution layer is used. It offers a unique feature, as it is able to drop the value after a specific range based on the dilation rate. The dilation rate has been essential for increasing the receptive field of the layers. The dilation rate has been varied here to capture both local and global dependencies. Thus, LSTM is used to expand the values. It is then the task of GRU to concise them, with dilated convolution taking over to make the prediction precise.

Lastly, a dense layer is used that will be activated based on the ReLU activation function to predict the final result and for fast convergence. The Adam optimizer has been employed for the model with a learning rate of 0.001.

Tab. 2. Deep hybrid model parameters.

Layers	Units	Parameters	Activation function
LSTM	128	68 608	tanh
LSTM	64	49 408	tanh
GRU	64	24 960	tanh
Conv1D	32	4 128	tanh
Conv1D	32	2 080	tanh
Conv1D	32	2 080	tanh
Conv1D	32	2 080	tanh
Conv1D	32	2 080	tanh
Dense	1	33	ReLU
Total parameters 472 613 (1.80 MB)			
Trainable parameters 157 537 (615.38 KB)			
Non-trainable parameters 0 (0.00 B)			
Optimizer parameters 315 076 (1.20 MB)			

4.2. Baseline ML/DL Models for Comparison

To validate the performance of the proposed deep hybrid model, its outcomes are compared with those achieved by several commonly used baseline ML and DL models. In the following section, a brief analysis of the baseline models is presented.

- Linear regression (LR). The model estimates path loss by assuming a linear relationship between path loss and the input features. The model was chosen for its simplicity and interpretability. The parameters used include fit intercept (set to true) and no regularization.
- Polynomial regression (PR). It is an extension of linear regression. It models non-linear relationships by including polynomial terms of the input features up to degree 2. The model is used to capture the non-linear relationship while maintaining computational efficiency.
- Random forest regression (RFR). It is an ensemble model. It includes 100 decision trees, with a maximum depth of 10 and a minimum of 2 samples per split. RFR is implemented because of its ability to capture non-linear relationships and robustness against overfitting.
- Support vector regression (SVR). The SVR model uses a kernel of radial basis function (RBF) with $C = 10$ and

$\varepsilon = 0.1$. SVR is efficient in handling higher-dimensional data and can model non-linear relationships through kernel transformations.

- Artificial neural network (ANN). The model consists of two hidden dense layers with 64 and 32 neurons, respectively, both using the ReLU activation function. The model also incorporates a dropout layer to reduce overfitting. ANN by far outperforms machine learning-based models in capturing non-linear relationships.
- Long short-term memory (LSTM). LSTM is used to capture temporal dependencies of the features. The model consists of two stacked LSTM layers with 64 and 32 units, respectively, both using the hyperbolic tangent (tanh) activation function. A dropout layer is also added to reduce overfitting. The model helps capture the sequential dependencies that may be overlooked by other approaches.
- Gated recurrent unit (GRU). Like LSTM, GRU also captures temporal dependencies. The model consists of two stacked GRU layers containing 64 and 32 units, respectively, each using the tanh activation function. The dropout layer is also used here to reduce overfitting. Although LSTM shows superior performance in capturing long-term dependencies, GRU can perform better in scenarios with limited data or where the temporal dependencies are moderately long.
- Convolution neural network (CNN). CNN is used to capture spatial dependencies. The model has a stack of six 1D convolutional layers, each with 32 filters, with a kernel size of 2. The dilated convolutional architecture is effective in capturing the spatial characteristics.

5. Result Analysis

5.1. Error Matrix Analysis

The evaluation metrics include mean absolute error (MAE), mean absolute percentage error (MAPE), and root mean square error (RMSE). MAE measures the average magnitude of errors between the predicted and actual values. It can be expressed as follows:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \hat{y}_i\right|, \tag{3}$$

where y_i is the actual value and $\hat{y}_i$ is the predicted value.

MAPE measures the average percentage error. It is sensitive to small actual values. It can be expressed as follows:

$$MAPE = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|. \tag{4}$$

RMSE gives more weight to larger errors by squaring them before averaging. It can be formulated in the following way:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}. \tag{5}$$

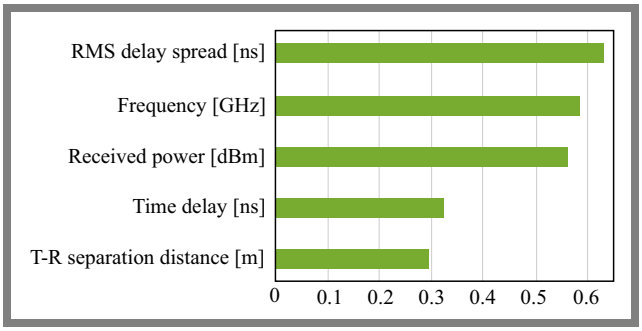

Fig. 4. Comparisons of feature sensitivity.

Tab. 3. Deep hybrid model parameters.

Model	MAE	MAPE	RMSE
LR	17.0053	10.8424	21.9414
RF	16.6703	10.6766	21.6529
SVM	15.0382	9.6501	19.3329
Polynomial	13.8306	8.9102	17.7603
LSTM	5.3798	3.4097	6.9062
ANN	4.8112	2.9737	6.1571
CNN	4.2579	2.6533	5.4366
GRU	4.0791	2.5039	5.1116
Proposed hybrid model	**3.9742**	**2.4512**	**5.0747**

Table 3 presents the performance metrics for all models. From the table, it is evident that the neural network-based model outperforms machine learning-based models, as it is capable of capturing non-linear relationships more effectively by incorporating spatial and temporal dependencies. The proposed deep hybrid model achieves the lowest RMSE and MAE, outperforming all baseline models. The results indicate that the model is able to successfully integrate several algorithms to extract their individual qualities in order to produce the best result.

5.2. Feature Sensitivity Analysis

While building the model, a total of five important variables were considered to estimate path loss. These variables have a direct influence on path loss. Studying sensitivity of the features is important to analyze whether a given model is biased towards one parameter only, which can severely impact its accuracy. The five variables, including RMS delay spread, operating frequency, received power, time delay, and distance between the transmitter and the receiver were taken into account to predict path loss.

After analyzing sensitivity of the features, it is evident from Fig. 4 that the prediction model is more sensitive to RMS delay spread, operating frequency, and power received by user equipment (UE). A minor alteration in these variables will significantly affect prediction values. RMS delay spread is a crucial parameter for determining path loss, as it indicates time dispersion of the signal arrival phase, present due to

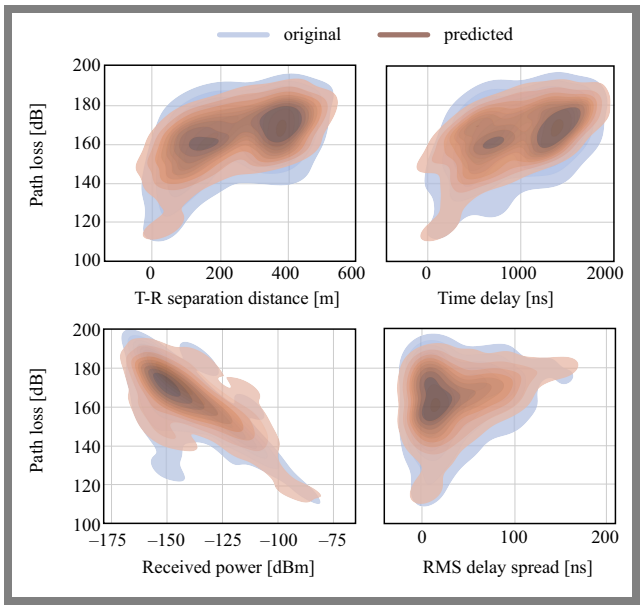

Fig. 5. Distribution of predicted values.

the multipath nature of the system. It identifies how spread out these arrival times are, thus helping manage the signal's integrity, especially in high-speed communication systems.

On the contrary, the prediction model proposed in this research is less sensitive to the T-R separation value – a characteristic that is desired in the practical world, because this separation distance can vary significantly, especially when UE is mobile. If the prediction model relies mostly on this parameter, then path loss prediction will be significantly impacted for high-speed users, and the prediction model will predict an arbitrary value which will eventually lead to a complete failure to operate efficiently in a real-world scenario.

The sensitivity is quantified using partial derivatives, which can be written as:

$$S_{(y,x_i)} = \frac{\partial y}{\partial x_i}\,, \tag{6}$$

where:
$S_{(y,x_i)}$ is the sensitivity of output y with respect to x. $\frac{\partial y}{\partial x_i}$ represents how y varies in response to small changes in x.

5.3. Analyzing the Distribution of the Prediction Values

It is important to know what value is produced from the deep learning model to make it suitable for real-world scenarios. It will help to better understand the model and tune its attributes to produce the best results. The five important parameters are plotted against predicted and actual path loss values in such a way that the probability density function (PDF) of the actual and predicted path loss is explained. The reason behind this is to see the range of predicted and original values and to detect any outliers or wrongly predicted values.

From Fig. 5 it is clear that the prediction model has captured the scenario clearly, as there is no outlier present in the plots. Moreover, the prediction values are more confined than the actual values, which not only represents the accuracy of the prediction model but also shows that it operates consistently in all the scenarios. As RMS delay spread plays the most crucial

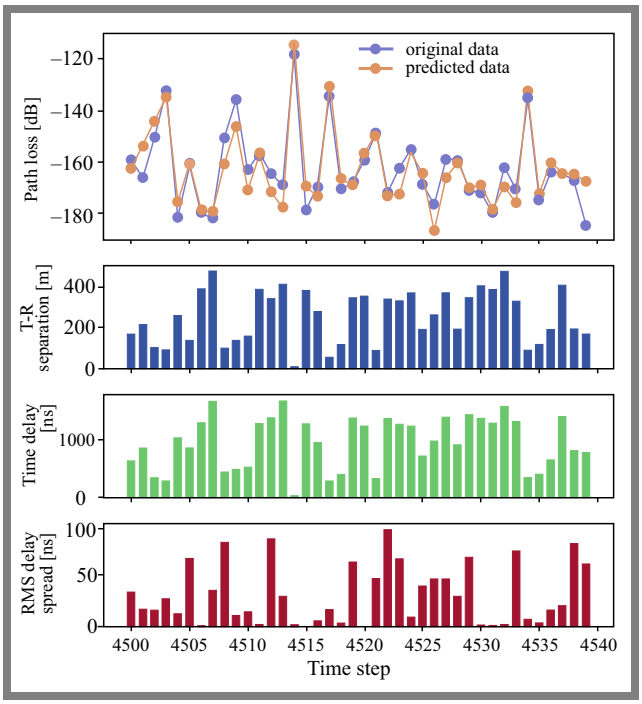

Fig. 6. Comparison of original and predicted data.

role in terms of impact, the distribution of predicted values against a specific parameter is more confined than in the case of actual values. This indicates that deviation from actual data is not significant. Moreover, the model could analyze every aspect of the related parameter and make proper justification before making the prediction.

5.4. Original and Predicted Data Patterns

Figure 6 illustrates that estimations of the prediction model not only closely align with multiple data points of actual values but also effectively capture the underlying patterns. For example, at time step 4514, both T-R separation and RMS delay spread decrease over time, while the actual path loss increases, and the predicted path loss also reflects the increase. A similar pattern is observed in steps 4504, 4517, and 4534. On the contrary, the model also accurately predicts the opposite scenarios, as seen in time steps 4507, 4526, and 4539. These prediction values indicate that the model successfully recognizes various scenarios and patterns, demonstrating its prediction accuracy.

6. Conclusions

Since path loss is a critical component of high frequency wireless communication, in this research a deep hybrid model was developed to predict path loss for high frequency communication, specifically for 5G and B5G. By combining LSTM, GRU, convolutional layers, and dense layers in the model development phase and utilizing the distinct characteristics of each algorithm, optimal results were achieved.

The approach becomes more robust and versatile when all types of dependencies are combined into one model. The convolutional layer offers spatial domain information, while

LSTM and GRU provide the temporal viewpoint of the features. Numerous simulations have demonstrated how well the suggested model predicts path loss and identifies its variance pattern.

By examining the dependency of the model on various parameters, the study further investigated the significance of the individual characteristics in the decision-making process of the suggested model.

Lastly, this study examines the ability of the hybrid model to predict actual outcomes by examining each pattern that might potentially emerge in a real-world environment. The results clearly demonstrated the model's potential for use in real-world scenarios.

References

[1] A.H. Kelechi *et al.*, "The Four-C Framework for High-capacity Ultra-low Latency in 5G Networks: A Review", *Energies*, vol. 12, art. no. 3449, 2019 (https://doi.org/10.3390/en12183449).

[2] M. Pons *et al.*, "Utilization of 5G Technologies in IoT Applications: Current Limitations by Interference and Network Optimization Difficulties – A Review", *Sensors*, vol. 23, art. no. 3876, 2023 (https://doi.org/10.3390/s23083876).

[3] C. Phillips, D. Sicker, and D. Grunwald, "A Survey of Wireless Path Loss Prediction and Coverage Mapping Methods", *IEEE Communications Surveys & Tutorials*, vol. 15, pp. 255–270, 2013 (https://doi.org/10.1109/surv.2012.022412.00172).

[4] S. Kurt and B. Tavli, "Path-loss Modeling for Wireless Sensor Networks: A Review of Models and Comparative Evaluations", *IEEE Antennas and Propagation Magazine*, vol. 59, pp. 18–37, 2017 (https://doi.org/10.1109/map.2016.2630035).

[5] V.S. Abhayawardhana *et al.*, "Comparison of Empirical Propagation Path Loss Models for Fixed Wireless Access Systems", *2005 IEEE 61st Vehicular Technology Conference*, Stockholm, Sweden, 2005 (https://doi.org/10.1109/VETECS.2005.1543252).

[6] J. Caffery, "A New Approach to the Geometry of TOA Location", *52nd Vehicular Technology Conference*, Boston, USA, 2000 (https://doi.org/10.1109/VETECF.2000.886153).

[7] M. Hata, "Empirical Formula for Propagation Loss in Land Mobile Radio Services", *IEEE Transaction on Vehicular Technology*, vol. 29, pp. 317–325, 1980 (https://doi.org/10.1109/t-vt.1980.23859).

[8] Y. Okumura, "Field Strength and Its Variability in VHF and UHF Land-mobile Radio Service", Review of the Electrical Communication Laboratory, vol. 16, pp. 825–873, 1968 (https://ci.nii.ac.jp/naid/10010001461).

[9] Q. Zhu *et al.*, "3GPP TR 38.901 Channel Model", *Wiley 5G Ref*, pp. 1–35, 2021 (https://doi.org/10.1002/9781119471509.w5gref048).

[10] S.I. Popoola, A.A. Atayero, O.D. Arausi, and V.O. Matthews, "Path Loss Dataset for Modeling Radio Wave Propagation in Smart Campus Environment", *Data in Brief*, vol. 17, pp. 1062–1073, 2018 (https://doi.org/10.1016/j.dib.2018.02.026).

[11] P.K. Sharma and R.K. Singh, "Comparative Analysis of Propagation Path Loss Models with Field Measured Data", *International Journal of Engineering Science and Technology*, vol. 2, 2010.

[12] R. Luebbers, "Propagation Prediction for Hilly Terrain Using GTD Wedge Diffraction", *IEEE Transaction on Antennas and Propagation*, vol. 32, pp. 951–955, 1984 (https://doi.org/10.1109/tap.1984.1143449).

[13] K. Yang, T. Ekman, T. Røste, and F. Bekkadal, "A Quasi-deterministic Path Loss Propagation Model for the Open Sea Environment", *14th International Symposium on Wireless Personal Multimedia Communications (WPMC)*, Brest, France, 2011.

[14] V. Mohtashami and A.A. Shishegar, "Modified Wavefront Decomposition Method for Fast and Accurate Ray-tracing Simulation", *IET Microwaves Antennas & Propagation*, vol. 6, pp. 295–295, 2012 (https://doi.org/10.1049/iet-map.2011.0264).

[15] S.Y. Seidel and T.S. Rappaport, "Site-specific Propagation Prediction for Wireless in-building Personal Communication System Design", *IEEE Transactions on Vehicular Technology*, vol. 43, pp. 879–891, 1994 (https://doi.org/10.1109/25.330150).

[16] R.G. Kouyoumjian and P.H. Pathak, "A Uniform Geometrical Theory of Diffraction for an Edge in a Perfectly Conducting Surface", *Proceedings of the IEEE*, vol. 62., pp. 1448–1461, 1974 (https://doi.org/10.1109/proc.1974.9651).

[17] K.S. Kunz and R.J. Luebbers, *The Finite Difference Time Domain Method for Electromagnetics*, Routledge: Boca Raton, USA, 464 p., 2018 (ISBN 9780367402372).

[18] Y. Wang *et al.*, "Machine Learning-enhanced Flexible Mechanical Sensing", *Nano-Micro Letters*, vol. 15, art. no. 55, 2023 (https://doi.org/10.1007/s40820-023-01013-9).

[19] R. Gupta *et al.*, "Artificial Intelligence to Deep Learning: Machine Intelligence Approach for Drug Discovery", *Molecular Diversity*, vol. 25, pp. 1–46, 2021 (https://doi.org/10.1007/s11030-021-10217-3).

[20] X. Wu and A. Che, "A Memetic Differential Evolution Algorithm for Energy-efficient Parallel Machine Scheduling", *Omega*, vol. 82, pp. 155–165, 2019 (https://doi.org/10.1016/j.omega.2018.01.001).

[21] X. Su, X. Yan, and C.L. Tsai, "Linear Regression", *Wires Computational Statistics*, vol. 4, pp. 275–294, 2012 (https://doi.org/10.1002/wics.1198).

[22] M. Awad and R. Khanna, *Efficient Learning Machines*, Apress Berkeley, Canada, 268 p., 2015 (https://doi.org/10.1007/978-1-4302-5990-9).

[23] B. de Ville, "Decision Trees", *Wires Computational Statistics*, vol. 5, pp. 448–455, 2013 (https://doi.org/10.1002/wics.1278).

[24] S. Dong, P. Wang, and K. Abbas, "A Survey on Deep Learning and Its Applications", *Computer Science Review*, vol. 40, art. no. 100379, 2021 (https://doi.org/10.1016/j.cosrev.2021.100379).

[25] O.A Montesinos Lopez, A. Montesinos Lopez, and J. Crossa, *Multivariate Statistical Machine Learning Methods for Genomic Prediction*, Springer Cham, Switzerland, 691 p., 2022 (https://doi.org/10.1007/978-3-030-89010-0).

[26] O.J. Famoriji and T. Shongwe, "Path Loss Prediction in Tropical Regions Using Machine Learning Techniques: A Case Study", *Electronics*, vol. 11, art. no. 2711, 2022 (https://doi.org/10.3390/electronics11172711).

[27] C.A. Oroza, Z. Zhang, T. Watteyne, and S.D. Glaser, "A Machine-learning-based Connectivity Model for Complex Terrain Large-scale Low-power Wireless Deployments", *IEEE Transactions on Cognitive Communications and Networking*, vol. 3, pp. 576–584, 2017 (https://doi.org/10.1109/tccn.2017.2741468).

[28] Y. Nunez, L. Lovisolo, L.D.S. Mello, and C. Orihuela, "Path-loss Prediction of Millimeter-wave using Machine Learning Techniques", *2022 IEEE Latin-American Conference on Communications (LATINCOM)*, Rio de Janeiro, Brasil, 2022 (https://doi.org/10.1109/LATINCOM56090.2022.10000523).

[29] M. Piacentini and F. Rinaldi, "Path Loss Prediction in Urban Environment Using Learning Machines and Dimensionality Reduction Techniques", *Computational Management Science*, vol. 8, pp. 371–385, 2010 (https://doi.org/10.1007/s10287-010-0121-8).

[30] R.D. Timoteo, D.C. Cunha, and G.D. Cavalcanti, "A Proposal for Path Loss Prediction in Urban Environments Using Support Vector Regression", *The Tenth Advanced International Conference on Telecommunications*, Paris, France, 2014.

[31] J. Wen *et al.*, "Path Loss Prediction Based on Machine Learning Methods for Aircraft Cabin Environments", *IEEE Access*, vol. 7, pp. 159251–159261, 2019 (https://doi.org/10.1109/ACCESS.2019.2950634).

[32] D. Karra, S.K. Goudos, G.V. Tsoulos, and G. Athanasiadou, "Prediction of Received Signal Power in Mobile Communications Using Different Machine Learning Algorithms: A Comparative Study", *2019 Panhellenic Conference on Electronics & Telecommunications (PACET)*, Volos, Greece, 2019 (https://doi.org/10.1109/PACET48583.2019.8956271).

[33] M.K. Elmezughi, O. Salih, T.J. Afullo, and K.J. Duffy, "Comparative Analysis of Major Machine-learning-based Path Loss Models for Enclosed Indoor Channels", *Sensors*, vol. 22, art. no. 4967, 2022 (https://doi.org/10.3390/s22134967).

[34] N. Zaarour, N. Kandil, N. Hakem, C. Despins, "Comparative Experimental Study on Modeling the Path Loss of an UWB Channel in a Mine Environment Using MLP and RBF Neural Networks", *2012 International Conference on Wireless Communications in Underground and Confined Areas*, Clermont-Ferrand, France, 2012 pp. 1–6. (https://doi.org/10.1109/ICWCUCA.2012.6402503).

[35] A.B. Zineb and M. Ayadi, "A Multi-wall and Multi-frequency Indoor Path Loss Prediction Model Using Artificial Neural Networks", *Arabian Journal for Science and Engineering*, vol. 41, pp. 987–996, 2015 (https://doi.org/10.1007/s13369-015-1949-6).

[36] S.P. Sotiroudis *et al.*, "Application of a Composite Differential Evolution Algorithm in Optimal Neural Network Design for Propagation Path-loss Prediction in Mobile Communication Systems", *IEEE Antennas and Wireless Propagation Letters*, vol. 12, pp. 364–367, 2013 (https://doi.org/10.1109/lawp.2013.2251994).

[37] S.I. Popoola *et al.*, "Determination of Neural Network Parameters for Path Loss Prediction in Very High Frequency Wireless Channel", *IEEE Access*, vol 7, pp. 150462–150483, 2019 (https://doi.org/10.1109/ACCESS.2019.2947009).

[38] V.C. Ebhota, J. Isabona, and V.M. Srivastava, "Environment-adaptation Based Hybrid Neural Network Predictor for Signal Propagation Loss Prediction in Cluttered and Open Urban Microcells", *Wireless Personal Communications*, vol. 104, pp. 935–948, 2018 (https://doi.org/10.1007/s11277-018-6061-2).

[39] P.-R. Chang and W.-H. Yang, "Environment-adaptation Mobile Radio Propagation Prediction Using Radial Basis Function Neural Networks", *IEEE Transactions on Vehicular Technology*, vol. 46, pp. 155–160, 1997 (https://doi.org/10.1109/25.554747).

[40] T. Balandier, A. Caminada, V. Lemoine, and F. Alexandre, "170 MHz Field Strength Prediction in Urban Environment Using Neural Nets", *6th International Symposium on Personal, Indoor and Mobile Radio Communications*, Toronto, Canada, 1995 (https://doi.org/10.1109/PIMRC.1995.476416).

[41] G. Panda, R.K. Mishra, and S.S. Palai, "A Novel Site Adaptive Propagation Model", *IEEE Antennas and Wireless Propagation Letters*, vol. 4, pp. 447–448, 2005 (https://doi.org/10.1109/lawp.2005.860213).

[42] L. Wu *et al.*, "Received Power Prediction for Suburban Environment based on Neural Network", *2020 International Conference on Information Networking (ICOIN)*, Barcelona, Spain, 2020 (https://doi.org/10.1109/ICOIN48656.2020.9016532).

[43] M. Ayadi, A.B. Zineb, and S. Tabbane, "A UHF Path Loss Model Using Learning Machine for Heterogeneous Networks", *IEEE Transactions on Antennas and Propagation*, vol. 65, pp. 3675–3683, 2017 (https://doi.org/10.1109/tap.2017.2705112).

[44] F. Cheng and H. Shen, "Field Strength Prediction Based on Wavelet Neural Network", *2010 2nd International Conference on Education Technology and Computer*, Shanghai, China, 2010 (https://doi.org/10.1109/ICETC.2010.5529392).

[45] H.F. Ates, S.M. Hashir, T. Baykas, and B.K. Gunturk, "Path Loss Exponent and Shadowing Factor Prediction from Satellite Images Using Deep Learning", *IEEE Access*, vol. 7, pp. 101366–101375, 2019 (https://doi.org/10.1109/access.2019.2931072).

[46] J.Y. Lee, M.Y. Kang, and S.C. Kim, "Path Loss Exponent Prediction for Outdoor Millimeter Wave Channels through Deep Learning", *2019 IEEE Wireless Communications and Networking Conference (WCNC)*, Marrakesh, Morocco, 2019 (https://doi.org/10.1109/WCNC.2019.8885668).

[47] N. Kuno and Y. Takatori, "Prediction Method by Deep-learning for Path Loss Characteristics in an Open-square Environment", *2018 International Symposium on Antennas and Propagation (ISAP)*, Busan, South Korea, 2018.

[48] R.R. Ratul *et al.*, "Atmospheric Influence on the Path Loss at High Frequencies for Deployment of 5G Cellular Communication Networks", *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Delhi, India, 2023 (https://doi.org/10.1109/ICCCNT56998.2023.10307972).

Kazi Md Abrar Yeaser, B.Sc., Lecturer
Faculty of Engineering
https://orcid.org/0009-0001-9922-1342
E-mail: abrar.yeaser@puc.ac.bd
Premier University, Chittagong, Bangladesh
https://puc.ac.bd

Kazi Md Abir Hassan, Graduate Student
Faculty of Engineering and Technology
https://orcid.org/0009-0003-3839-974X
E-mail: abirhassan@iut-dhaka.edu
Islamic University of Technology, Gazipur, Bangladesh
https://www.iutoic-dhaka.edu

Intelligent Secure Data Aggregation in WSNs

Olena Semenova[1], Natalia Kryvinska[2], Serhii Baraban[3], Maksym Prytula[1], and Volodymyr Martyniuk[1]

[1] *Vinnytsia National Technical University, Vinnytsia, Ukraine,*
[2] *Comenius University in Bratislava, Bratislava, Slovakia,*
[3] *Poznan University of Technology, Poznan, Poland*

https://doi.org/10.26636/jtit.2025.3.2220

Abstract — The paper discusses the problem of secure data aggregation in wireless sensor networks (WSNs) – a procedure that is of critical importance for reducing energy consumption, minimizing transmission overhead, and thus prolonging network lifetime. Due to the limited computational and energy resources of WSN nodes, traditional aggregation methods often fail to perform effectively in dynamic heterogeneous environments. With such a context taken into consideration, this study emphasizes the potential of artificial intelligence techniques, such as neural networks, genetic algorithms, and fuzzy logic, to enable adaptive aggregation approaches tailored to environmental and network-specific parameters. Furthermore, the integration of fuzzy logic, genetic algorithms, and artificial neural networks into a hybrid system leverages the strengths of each approach, resulting in enhanced adaptability and accuracy of the aggregation process. As part of the investigation, a fuzzy inference system (FIS) model was developed that incorporates attributes such as energy, current load, distance to the base station, and trust level. The model was implemented in Matlab using the Fuzzy Logic Designer toolbox. To further improve system performance, a genetic algorithm was applied to optimize membership functions. In the final phase, the model was transformed into an adaptive neurofuzzy inference system (ANFIS) which was trained using simulated data within Matlab. The simulation results demonstrate that the proposed hybrid approach ensures flexible, robust and energy-efficient control of the data aggregation process under dynamically changing conditions in which WSNs operate.

Keywords — artificial intelligence, data aggregation, fuzzy logic, security, wireless sensor networks

1. Introduction

Wireless sensor networks (WSNs) play an important role in a wide range of applications, including industrial manufacturing, smart cities, automotive, healthcare and environmental monitoring [1]. In these environments, autonomous sensor nodes are distributed to monitor various conditions. The sensors gather data and transmit it to a central node for processing and further analysis.

Scalability, self-organization, and adaptability are among the characteristics of WSNs that make them effective tools for processing data in real-time, particularly in potentially hazardous or hard-to-reach situations. Data aggregation and routing are two essential processes in WSNs that contribute to improving energy efficiency, extending network lifetime, and minimizing communication overhead [2].

Data aggregation refers to the process of collecting useful data. In WSNs, appropriate data aggregation procedures are required to preserve limited resources. The primary objective of aggregation algorithms is to collect data in a manner that optimizes energy efficiency, thereby extending the network's lifespan.

As WSNs are characterized by restricted computational capabilities, limited memory, and finite battery capacity, all this complicates the development process. Moreover, in some cases, this may result in applications that are tightly integrated with network protocols [3]. Furthermore, data aggregation is employed to address overlapping in data routing. When data from various sensor nodes converge at the same node on their return to the sink, they are aggregated as if they pertain to the same data.

In general, data aggregation methods relied upon in WSNs can be classified into four categories: cluster-based, tree-based, in-network, and centralized data aggregation [4].

In the cluster-based approach, the network is segmented into clusters. Each cluster is made up of a group of sensor nodes, with one node designated as the cluster head. The cluster head is responsible for data aggregation, where the collected data is combined and subsequently sent to the sink.

These clusters operate in two distinct phases: during the initial (setup) phase the cluster selection process occurs and clusters are established. The steady phase follows, in which the cluster is functioning. Throughout the steady phase, all nodes within the cluster, including the cluster head, continuously sense their environment for specific data in a regular way.

All member nodes transmit the detected data to the cluster head which aggregates the information and forwards it to the sink. This strategy minimizes bandwidth usage by decreasing the number of packets that need to be transmitted. Furthermore, the data aggregation process relied upon in this method not only reduces the number of packets sent directly to the sink, but also lowers energy consumption due to the shorter transmission distances. However, it suffers from a drawback, namely increased latency. Cluster-based data

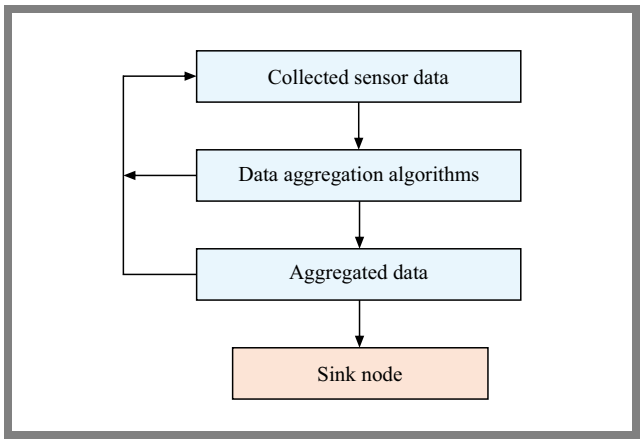

Fig. 1. Diagram presenting the general data aggregation algorithm.

aggregation techniques include LEACH, HEED, SEP, and PAgloT.

In the tree-based approach, aggregation trees are established in such a way that every data transmission requires the formation of spanning trees. In this framework, the base station functions as the root of the tree, whereas sensor nodes act as leaves. Data are collected by the leaves and transmitted towards the root with the parent nodes consolidating the data throughout the networks.

In-network aggregation represents a comprehensive method for collecting and processing data at intermediary nodes, in addition to facilitating the routing of information through multi-hop networks. Its primary objective is to minimize the consumption of energy required to perform the process. This method may either decrease the size of the data, leading to a reduction in the amount of data that needs to be transmitted subsequently, or maintain the width by combining all received packets.

In the centralized approach, all sensors transmit the collected data as data packets to a central node or base station via the shortest available route. The function of the aggregator or header node is to compile the data received from the other nodes, after which the consolidated data are sent as a single packet.

Figure 1 shows the general data aggregation algorithm through different aggregation techniques [5]. The algorithm utilizes sensor data from the sensor nodes and aggregates them using several aggregation algorithms, including the centralized approach, low energy adaptive clustering hierarchy (LEACH), and tiny aggregation (TAG), among others. These aggregated data are then transmitted to the sink node by selecting the most efficient path.

However, various efficient aggregation protocols do not meet the resource constraints. Moreover, numerous security threats exist, including but not limited to snooping attacks, wormhole attacks, black hole attacks, packet replication attacks, denial-of-service (DoS) attacks, and distributed denial-of-service (DDoS) attacks. In many cases, the process of optimizing performance of WSNs concerns more than one of the metrics, thus necessitating the application of multi-objective optimization [6].

2. Problem Definition

During the data aggregation process, several challenges must be addressed. It is evident that it is difficult to overcome all these challenges simultaneously. The most significant challenges include the following [7]:

- **Data redundancy**. Sensor nodes often detect similar types of data and even the same events, leading the sink node to collect redundant information. This results in a waste of time, energy, and other resources.
- **Delay**. In some cases, data from more distant nodes arrive late at the sink or root node, causing the aggregation process to commence later than intended. Additionally, aggregations at intermediate levels can further increase the delay.
- **Accuracy**. There are two primary types of accuracy-related issues. Firstly, the aggregator function serves as an approximation mechanism. Therefore, some precision is inevitably lost during the data forwarding process. Secondly, there may be a compromised node that transmits false or inappropriate data to the aggregator node. The aggregator node does not guarantee the correctness of these data and proceeds to process them.
- **Traffic load**. In specific situations, the aggregator node may become overloaded. This occurs when load balancing is not effectively implemented or when clusters are of unequal sizes.
- **Aggregation freshness**. Data from similar frames should be aggregated, while the use of outdated stored data or the aggregation of data from multiple frames across different time periods should be avoided, as such an approach compromises freshness.
- **Security**. As wireless sensor networks are often implemented in hostile environments, security issues, particularly involving data confidentiality and integrity, become crucial.

Therefore, when a malicious node infiltrates the network, ensuring the delivery of packets to the base station becomes a challenge. Given the resource limitations inherent in WSNs, it is essential to guarantee packet delivery while minimizing energy consumption. Transmission of redundant data within the network accelerates the depletion of energy in a node. This leads to network partitioning which, in turn, results in increased energy consumption and a consequent reduction in the overall lifetime of the network.

Therefore, in conjunction with data aggregation, it is essential to ensure the security of successful data transmission to the base station through the efficient use of available resource parameters.

To address a variety of challenges related to energy efficiency, coverage maximization, and security provision in WSNs, artificial intelligence (AI) can be applied effectively in wireless sensor networks by enabling smart decision making [8].

AI denotes the ability of a system to perform tasks that require human-like intelligence, emulating human thought processes or concepts. It is regarded a significant domain within com-

puter science that aims to enhance machine intelligence. The predominant techniques employed in AI include search algorithms, learning methodologies, fuzzy systems, knowledge representation, and reasoning processes [9].

AI finds application in addressing numerous intricate issues across diverse fields such as security, finance, healthcare, and transportation, leveraging its proficiency in managing incomplete and noisy data, tackling nonlinear problems, and demonstrating suitability for prediction and accelerated post-training generalization. The different AI techniques used to tackle WSN-related challenges comprise fuzzy logic, artificial neural networks, evolutionary computation, nature-inspired approaches, swarm intelligence, deep learning, reinforcement learning, and hybrid models [10].

3. Literature Review

This section discusses various publications associated with the application of AI used for data aggregation in WSNs.

Study [11] proposed a fuzzy-based secure data aggregation protocol which improves the lifetime of the network, maximizes the packet delivery ratio, and minimizes the end-to-end delay. Paper [12] introduces a fuzzy-based data aggregation technique to ensure energy efficiency in wireless sensor networks. A similarity-aware data aggregation process using a fuzzy c-means approach is discussed in [13].

Current secure data aggregation protocols represent a trade-off between security and the shortest path. To address this problem, the protocols mentioned in [15] are developed by integrating the distributed k-means algorithm and the fuzzy c-means algorithm. In work [15], a data aggregation algorithm was constructed based on a self-organizing feature mapping neural network.

In [16], a machine learning-based approach for an energy efficient data aggregation model in WSNs was described. The security feature can also be incorporated into the proposed model. Study [17] presents four algorithms for data aggregation. Three of them are based on backpropagation neural networks.

To mitigate energy consumption and ensure data aggregation in WSNs, study [18] presents a cluster-based data aggregation routing approach with a genetic search algorithm. This method aims to reduce energy use. In [19], a novel hybrid LEACH algorithm was introduced for data aggregation based on a genetic algorithm to optimize WSN parameters, including energy consumption.

As different AI-based techniques for data aggregation in WSNs have their own pros and cons [20], a combination of these approaches may be beneficial, as it leverages the complementary strengths of the different models, thus overcoming the limitations of individual methods.

Here, the authors propose using a hybrid approach for data aggregation in wireless sensor networks that combines three AI technologies: fuzzy logic, artificial neural networks, and genetic algorithms. This integration allows one to account for the uncertainty and incompleteness of the sensor data as well as dynamic changes in the network and the surrounding environment. The proposed approach ensures a high level of adaptability and reliability while making data aggregation-related decisions.

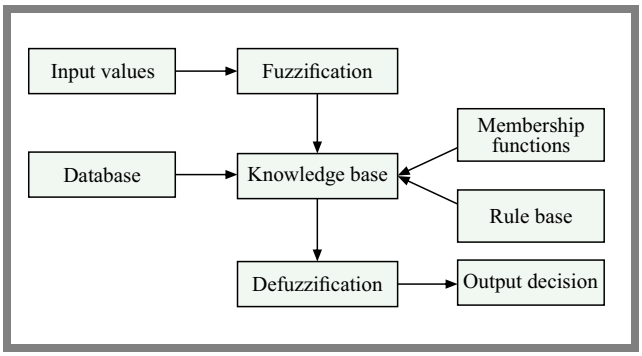

Fig. 2. Diagram presenting the proposed FIS architecture.

4. Methodology

In this section, a fuzzy inference system (FIS) intended for data aggregation in WSNs will be developed. A common FIS consists of four functional units: fuzzification, rule base, decision making, and defuzzification (Fig. 2).

The fuzzification unit transforms crisp inputs into linguistic variables. A rule base consists of a collection of fuzzy if-then rules. The decision-making unit conducts inference based on the fuzzy if-then rules. The defuzzification unit converts the fuzzy results generated by the inference system into precise outputs [21].

To develop an FIS, input in the form of linguistic variables along with their corresponding terms must be established. Then, the membership functions associated with these inputs are to be defined. The inputs include all conceivable states of the process being controlled, whereas the output represents all potential control actions. Subsequently, a rule base must be specified, consisting of a set of if-then fuzzy rules to characterize the controlled states. Finally, to evaluate the performance of the FIS, a simulation is performed.

The inputs of the (FIS) proposed in this study include the following: energy level of the node, distance from the base station or the sink node, load, and node trust level. Its output variable is the aggregation priority of the node.

The energy level of a sensor node is the remaining battery capacity which directly influences its suitability for performing energy-intensive tasks, such as data aggregation. Nodes with higher energy levels are preferred to act as aggregators to prolong the overall network lifetime. From a security perspective, maintaining a minimum energy threshold is critical, as low-energy nodes are more vulnerable to exhaustion attacks and may be less reliable in executing secure aggregation protocols.

For the energy input variable E such linguistic terms as “low” and “high” are applied. These may be regarded intuitive categories offering a general assessment and facilitating decision

making. Thus, we obtain the following:

$$E \to \left\{\mu_E^{low}(E),\, \mu_E^{high}(E)\right\}. \tag{1}$$

The distance between a sensor node and the base station or the sink node impacts energy consumption and transmission latency. Therefore, nodes located closer to the sink are generally better suited for aggregation due to reduced communication costs and lower risk of packet loss. From a security standpoint, longer distances may increase the exposure of transmitted data to interception or manipulation, thereby necessitating stronger encryption or trust mechanisms.

For the distance input variable D such linguistic terms as "near" and "far" are applied as:

$$D \to \left\{\mu_D^{near}(D),\, \mu_D^{far}(D)\right\}. \tag{2}$$

The traffic load of a sensor node corresponds to the volume of data packets it processes or forwards over a given time interval, which significantly impacts its efficiency in data aggregation operations. High traffic load can lead to increased latency, buffer overflows, and reduced aggregation accuracy due to packet collisions or losses. From a security point of view, excessive traffic may signal potential threats, such as flooding or spoofing attacks.

For the load input variable L such linguistic terms as "small" and "large" are applied in the following way:

$$L \to \left\{\mu_L^{small}(L),\, \mu_L^{lar}(L)\right\}. \tag{3}$$

Node trust is a quantified reliability of a sensor node. High-trust nodes are prioritized to serve as aggregators in order to ensure that the fused data is accurate, timely, and free from manipulation. From a security perspective, incorporating trust evaluation helps mitigate the risks posed by compromised or malicious nodes, leading to enhanced robustness of the aggregation process.

For the trust input variable T such linguistic terms as "low" and "high" are applied as follows:

$$T \to \left\{\mu_T^{low}(T),\, \mu_T^{high}(T)\right\}. \tag{4}$$

Aggregation priority can be regarded as a level of preference assigned to a given sensor node when it comes to performing data aggregation tasks within the wireless network. From a security point of view, this priority should be determined by evaluating factors such as node trustworthiness, energy availability, and exposure to potential threats, ensuring that only reliable and resilient nodes are selected.

Prioritizing secure nodes for aggregation reduces the likelihood of data tampering, spoofing, or compromised fusion, thus enhancing the overall integrity and confidentiality of aggregated information. For the aggregation priority input variable P, we propose to apply 8 linguistic terms: "no priority", "very weak", "weak", "medium weak", "medium", "medium strong", "strong", "very strong". Thus, we obtain the following:

$$\begin{aligned} P \to \big\{&\mu_p^{no}(P),\, \mu_p^{wv}(P),\, \mu_p^{v}(P),\, \mu_p^{mv}(P),\\ &\mu_p^{m}(P),\, \mu_p^{ms}(P),\, \mu_p^{s}(P),\, \mu_p^{vs}(P)\big\}. \end{aligned} \tag{5}$$

Fuzzification is the process of transforming crisp input values into degrees of membership by mapping them onto predefined fuzzy sets through membership functions. The trapezoid shape has been selected for membership functions of input values because they are quite simple to use, easy to comprehend, and can help in smooth transitions between different phrases. Thus, the membership functions for the inputs are defined as follows:

For the energy input E, we have:

$$\mu_{low}(E) = \begin{cases} 1 & \text{if} \quad E \leqslant 0.4 \\ \dfrac{0.8 - E}{0.8 - 0.4} & \text{if } 0.4 < E \leqslant 0.8 \\ 0 & \text{if} \quad E > 0.8 \end{cases}, \tag{6}$$

$$\mu_{high}(E) = \begin{cases} 0 & \text{if} \quad E \leqslant 0.4 \\ \dfrac{E - 0.4}{0.8 - 0.4} & \text{if } 0.4 < E \leqslant 0.8 \\ 1 & \text{if} \quad E > 0.8 \end{cases}, \tag{7}$$

For the distance input D, we have:

$$\mu_{near}(D) = \begin{cases} 1 & \text{if} \quad D \leqslant 0.3 \\ \dfrac{0.7 - D}{0.7 - 0.3} & \text{if } 0.3 < D \leqslant 0.7 \\ 0 & \text{if} \quad D > 0.7 \end{cases}, \tag{8}$$

$$\mu_{far}(D) = \begin{cases} 0 & \text{if} \quad D \leqslant 0.3 \\ \dfrac{D - 0.3}{0.7 - 0.3} & \text{if } 0.3 < D \leqslant 0.7 \\ 1 & \text{if} \quad D > 0.7 \end{cases}, \tag{9}$$

For the load input L, we have:

$$\mu_{small}(L) = \begin{cases} 1 & \text{if} \quad L \leqslant 0.2 \\ \dfrac{0.6 - L}{0.6 - 0.2} & \text{if } 0.2 < L \leqslant 0.6 \\ 0 & \text{if} \quad L > 0.6 \end{cases}, \tag{10}$$

$$\mu_{lar}(L) = \begin{cases} 0 & \text{if} \quad L \leqslant 0.2 \\ \dfrac{L - 0.2}{0.6 - 0.2} & \text{if } 0.2 < L \leqslant 0.6 \\ 1 & \text{if} \quad L > 0.6 \end{cases}, \tag{11}$$

For the trust input T, we have:

$$\mu_{low}(T) = \begin{cases} 1 & \text{if} \quad T \leqslant 0.25 \\ \dfrac{0.75 - T}{0.75 - 0.25} & \text{if } 0.25 < T \leqslant 0.75 \\ 0 & \text{if} \quad T > 0.75 \end{cases}, \tag{12}$$

$$\mu_{high}(T) = \begin{cases} 0 & \text{if} \quad T \leqslant 0.25 \\ \dfrac{T - 0.25}{0.75 - 0.25} & \text{if } 0.25 < T \leqslant 0.75 \\ 1 & \text{if} \quad T > 0.75 \end{cases}, \tag{13}$$

The thresholds used to define the membership functions were analytically selected by the authors. To enhance their accuracy and validity, they will be corrected by means of genetic optimization and neural training.

In this investigation, we propose to utilize a Sugeno-type fuzzy inference system, as it offers computational efficiency and ease of mathematical analysis due to its use of "crisp" singleton

outputs and linear functions, making it highly suitable for real-time and embedded applications. Unlike Mamdani FIS, in which a defuzzification phase involves complex centroid calculations, Sugeno FIS produces output through weighted averages, thus resulting in faster and more precise decision-making. Compared to Tsukamoto FIS, which restricts output membership functions to be monotonic and performs rule-by-rule defuzzification, Sugeno FIS provides greater flexibility.

Therefore, the P membership functions of the aggregation priority output are singletons. They are defined as:

$$\mu_{No}(P) = \begin{cases} 1 & \text{if } P = 0 \\ 0 & \text{otherwise} \end{cases}, \tag{14}$$

$$\mu_{vw}(P) = \begin{cases} 1 & \text{if } P = 0.14 \\ 0 & \text{otherwise} \end{cases}, \tag{15}$$

$$\mu_{w}(P) = \begin{cases} 1 & \text{if } P = 0.29 \\ 0 & \text{otherwise} \end{cases}, \tag{16}$$

$$\mu_{mw}(P) = \begin{cases} 1 & \text{if } P = 0.43 \\ 0 & \text{otherwise} \end{cases}, \tag{17}$$

$$\mu_{m}(P) = \begin{cases} 1 & \text{if } P = 0.57 \\ 0 & \text{otherwise} \end{cases}, \tag{18}$$

$$\mu_{sm}(P) = \begin{cases} 1 & \text{if } P = 0.71 \\ 0 & \text{otherwise} \end{cases}, \tag{19}$$

$$\mu_{s}(P) = \begin{cases} 1 & \text{if } P = 0.86 \\ 0 & \text{otherwise} \end{cases}, \tag{20}$$

$$\mu_{vs}(P) = \begin{cases} 1 & \text{if } P = 0.1 \\ 0 & \text{otherwise} \end{cases}, \tag{21}$$

The inference engine in a fuzzy inference system applies logical reasoning to map fuzzified inputs to corresponding fuzzy outputs based on a predefined set of fuzzy rules. It determines the degree to which each rule is activated, and combines their outcomes to generate an aggregated response, thus representing the system's behavior.

In general, a fuzzy rule for our FIS is as follows:

$$\begin{aligned} R^i : \text{if } E = A_E^i \text{ and } D = A_D^i \text{ and } L = L_L^i \\ \text{and } T = A_T^i \text{ and then } P = z^i \, . \end{aligned} \tag{22}$$

The firing strength of a rule is:

$$w^i = \mu_E^{A_E^i}(E) \cdot \mu_D^{A_D^i}(D) \cdot \mu_L^{L_L^i}(L) \cdot \mu_T^{A_T^i}(T) \, . \tag{23}$$

The suggested FIS for data aggregation operates according to an 8-rule base shown in Tab. 1. These rules were deduced after the preliminary calculation performed by the authors. To enhance their accuracy and validity, they will be corrected by neural training. If needed, they can also be corrected while applying the genetic optimization.

Defuzzification is the process of converting the fuzzy reasoning result into a crisp numerical output. Sugeno FIS performs defuzzification through a weighted average of singleton outputs, where weights correspond to the firing strengths of the rules. For this case, we get:

$$P = \frac{\sum_{i=1}^{8} w^i \cdot z^i}{\sum_{i=1}^{8} w^i} \, . \tag{24}$$

However, FISs often lack the adaptive learning capability and they may not be able to adjust to new circumstances, as the rules and membership functions are quite rigid. Overall, this makes FISs less efficient in the dynamically changing environment of modern wireless heterogeneous networks. That is why, in many technical applications, fuzzy inference systems are being integrated with a neural network or genetic algorithms. Inspired by biological neurons, neural networks are computer models that can recognize patterns and relations in real-world data [22].

In this investigation, the authors propose to utilize an adaptive neurofuzzy inference system (ANFIS). ANFIS is classified as a hybrid artificial intelligent system whose characteristics place it between neural network and FIS. Therefore, it combines the benefits of fuzzy logic and those of neural networks, enabling it to learn and adapt its rules to increase accuracy compared to conventional FIS [23].

ANFIS is a more appropriate technique for real-world issues where data patterns are not always easily captured by classical fuzzy rules, since it can handle, unlike regular FIS, complicated and non-linear interactions between input and output values. Mapping between the specified input values and the intended output values is performed using ANFIS training.

Genetic algorithms are evolutionary optimization techniques inspired by natural processes [24]. The combination of a genetic algorithm with the FIS involves optimizing fuzzy rule parameters and membership functions to improve system efficiency. This approach works especially well for complicated issues, allowing the optimized FIS to attain greater accuracy.

The genetic algorithm systematically produces populations of parameter sets, implements selection, crossover, and mutation operations, and progresses towards optimal solutions. It processes encoded parameters of the membership functions and applied evolutionary operators to minimize an objective function and to iteratively search for the optimal solution. This iterative process persists until the convergence criterion

Tab. 1. Fuzzy rules.

	E	D	L	T	P
1	L	F	S	L	VW
2	H	F	S	L	W
3	L	F	L	L	No
4	H	N	L	L	MW
5	L	N	L	H	M
6	H	N	S	H	VS
7	L	F	S	H	MS
8	H	N	L	H	S

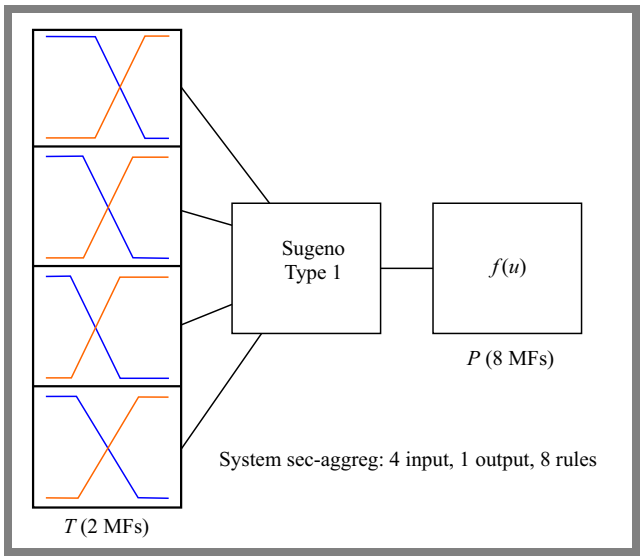

Fig. 3. Fuzzy inference system developed in Matlab.

is satisfied, at which point the most effective parameter set is chosen for the ultimate implementation of the fuzzy inference system.

The objective function is defined as:

$$J(\theta) = \frac{1}{N}\sum_{i=1}^{N}\left[y_i(\theta) - \hat{y}_i\right]^2, \tag{25}$$

where y_i is an actual output from the FIS for the i-th training sample, $\hat{y}_i$ is a desired output for the i-th training sample, N is a total number of training samples, and θ is a vector of membership function parameters.

5. Simulation

Matlab software can be utilized effectively to validate the functionality of the developed FIS for data aggregation in WSNs. In general, Matlab serves as a comprehensive platform for the visualization and fine-tuning of membership functions, rule bases, and output surfaces, thereby facilitating the development of more accurate and reliable decision-making systems. Simulation of FIS offers key advantages, including rapid prototyping, systematic performance evaluation, and streamlined experimentation. Furthermore, Matlab provides the integration of FIS with machine learning techniques such as neural networks and with optimization toolboxes, which enhance the adaptability and efficiency of fuzzy inference systems, enabling their refinement and deployment in complex real-world scenarios.

The first step was to specify the membership functions for the inputs and the output. Four inputs and one output variables were specified. Figure 3 illustrates the interface of the suggested FIS for data aggregation. Here, the FIS editor outlines the main information about the designed fuzzy inference system.

Then, the rule base was assigned. Next, we assigned the input values and ran the simulation process to produce outputs to check the operability of the developed FIS.

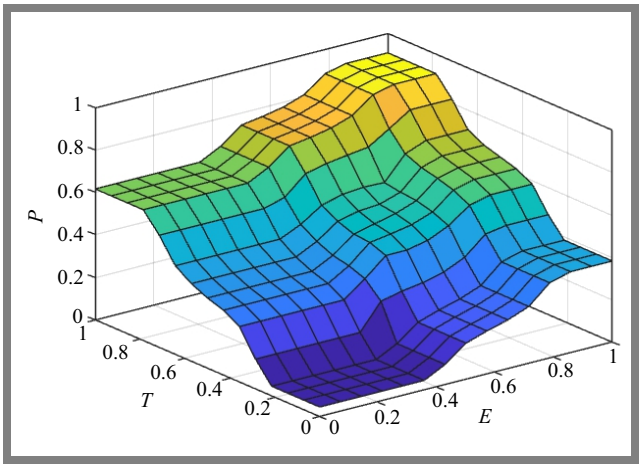

Fig. 4. Visualization of the control surface.

The control surface of an FIS provides a three-dimensional visualization of the system's output behavior based on two selected input variables. In the model developed, the control surface is defined with the x-axis representing the energy level E, the y-axis representing the trust level of node T, and the z-axis representing the resulting aggregation priority P, as shown in Fig. 4. This surface plot illustrates how variations in energy and trust jointly influence the aggregation priority, offering an intuitive understanding of the system's decision-making logic and enabling the evaluation of its responsiveness to changes in critical node parameters.

In the first case, energy value E was 0.46, distance value D was 0.36, load value L was 0.54, and the trust value equaled 0.15. According to Fig. 5, it yielded the aggregation priority of the node equal to 0.215. This means that this sensor node has a very low priority when it comes to choosing it as the aggregation node.

In the second case, energy value E was 0.76, distance value D was 0.22, load value L was 0.77, and the trust value equaled 0.62. According to Fig. 6, it yielded the aggregation priority

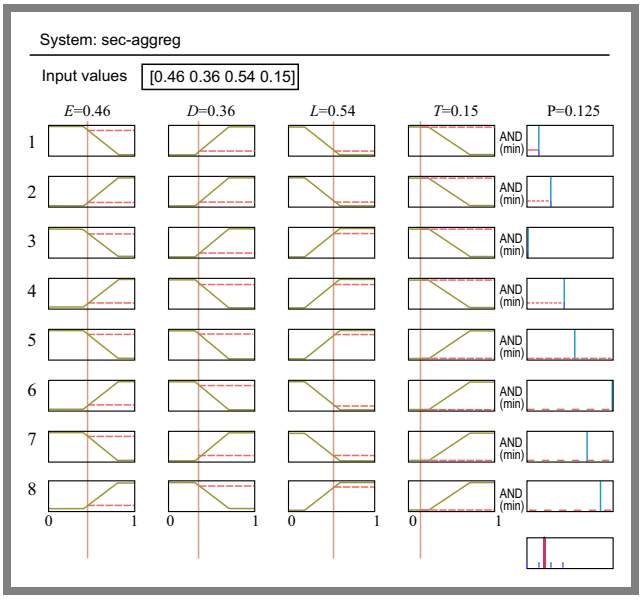

Fig. 5. Simulation results for the first case.

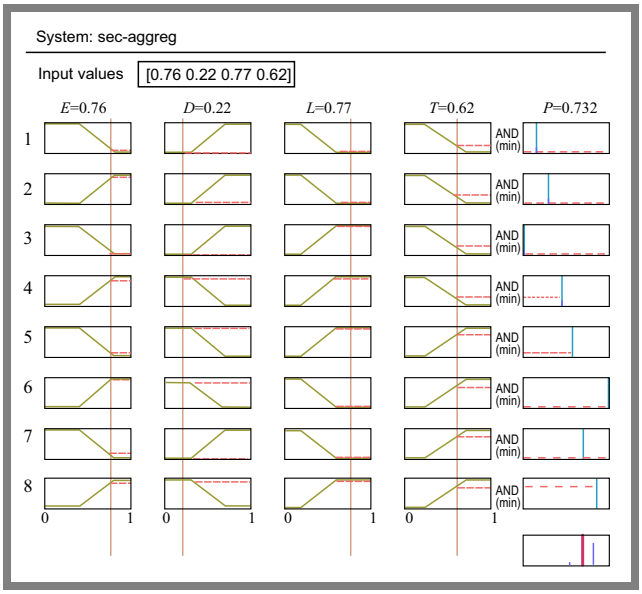

Fig. 6. Simulation results for the second case.

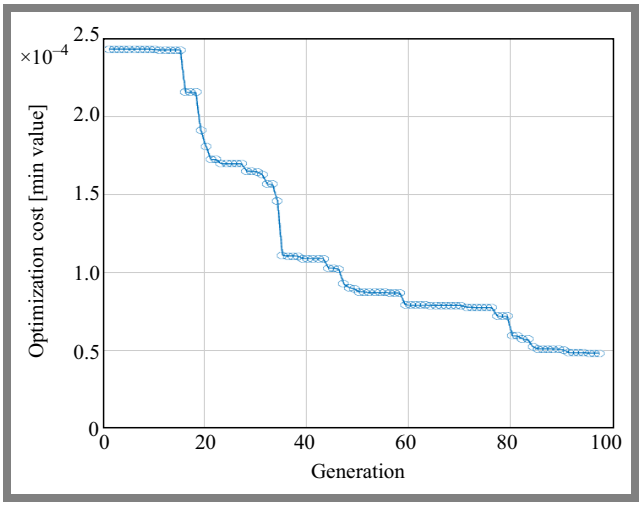

Fig. 7. FIS optimization results showing the training convergence process.

of the node equal to 0.732. This means that this sensor node has quite a high priority when it comes to being selected as the aggregation node.

Next, in this investigation, we utilized a genetic algorithm to optimize the parameters of the FIS developed in Matlab. Overall, the optimization procedure seeks to improve FIS performance by adjusting rule weights and membership functions. As genetic algorithms can handle non-linear search spaces, they are frequently employed as optimization tools. Matlab provides an appropriate environment to combine fuzzy logic systems with various optimization techniques. This is expected to result in increased system stability and control accuracy.

The optimization process involves the establishment of an objective function that assesses the efficacy of a fuzzy inference system according to particular criteria, such as error minimization.

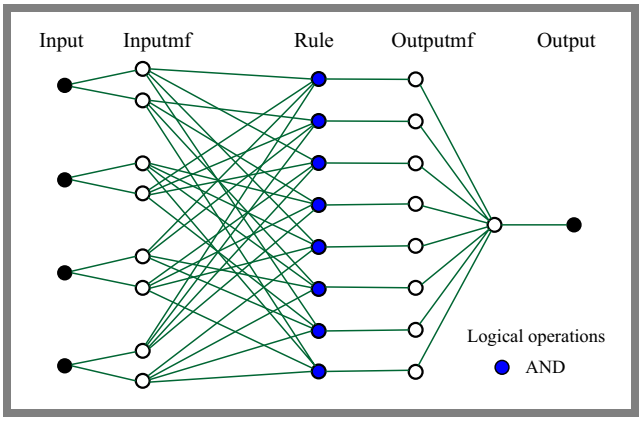

Fig. 8. Structure of the developed ANFIS.

A standard genetic algorithm is employed to optimize the parameters of the FIS. For our case, the optimization specifically targeted membership function parameters, while the rule base structure remained unchanged to preserve expert-defined logic. To perform genetic optimization, a test sample of input and output parameters was utilized.

The genetic algorithm converged after 97 iterations, resulting in a refined set of membership functions (Fig. 7). Here, the convergence of the genetic algorithm indicates that the optimization process has successfully reached a stable solution where successive generations no longer produce significant improvements and that the error between the FIS output and the target output is minimized.

This implies that the parameters of the membership functions have been effectively adjusted, thereby enhancing the accuracy of the FIS. Therefore, the simulation result confirmed that the genetic algorithm effectively tuned the parameters of the membership functions.

To further enhance the system's adaptability, the developed FIS was converted into an adaptive neurofuzzy inference system (ANFIS) using the ANFIS edit tool. This transformation enabled the integration of neural network learning capabilities with the interpretability of fuzzy logic, thus allowing the system to automatically adjust its parameters based on training data.

The structure of the generated ANFIS is shown in Fig. 8. The ANFIS model was trained using the backpropagation optimization method which adjusts the parameters of the membership functions by minimizing the error between the predicted and target outputs through gradient descent. Unlike the hybrid method, this approach relies solely on backpropagation to iteratively update both the premise and the consequent parameters, based on the computed error signals. Although computationally more intensive, this method offers full control over the learning process and ensures that all parameters are optimized in a unified framework driven by error minimization.

The ANFIS was trained considering the data from the WSN-DS dataset, specialized for intrusion detection in WSNs. Training the ANFIS optimizes its parameters by iteratively adjusting them to minimize the difference between the predicted and target outputs. Two types of parameters are op-

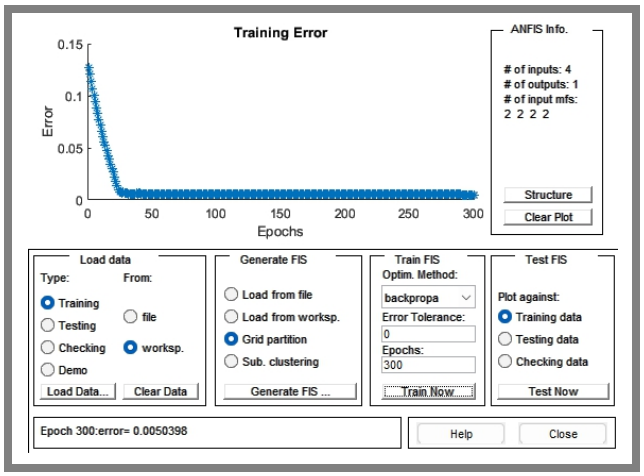

Fig. 9. Trend of errors in the trained ANFIS.

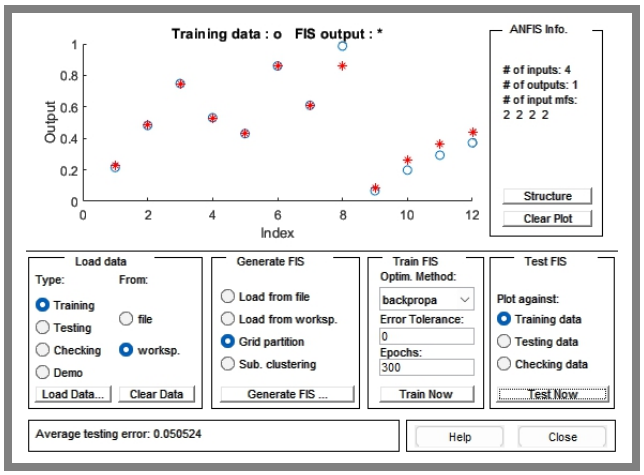

Fig. 10. Testing errors.

timized, i.e. the shape of the membership functions and linear function coefficients in the fuzzy rules.

The training persisted for 300 epochs until the error reached a relatively low level. Error assessment was conducted using the root mean square error method. Figure 9 shows this procedure, illustrating the decrease in RMSE in successive epochs, thereby signifying the effective training of the ANFIS model. Figure 10 presents the testing error represented by asterisks and the training error indicated by dots.

A decrease in training error signifies a progressive improvement in parameter adjustment. This decrease indicates that the ANFIS has effectively learned from the training data, resulting in optimized membership functions and rule parameters.

The simulation results support the relevance of the application of the proposed hybrid optimized fuzzy inference system in WSNs with security issues.

The combination of fuzzy logic with genetic algorithm optimization methods and neural network learning processes emphasizes the significant potential of the proposed approach for being applied in highly dynamic and resource-limited wireless sensor network environments.

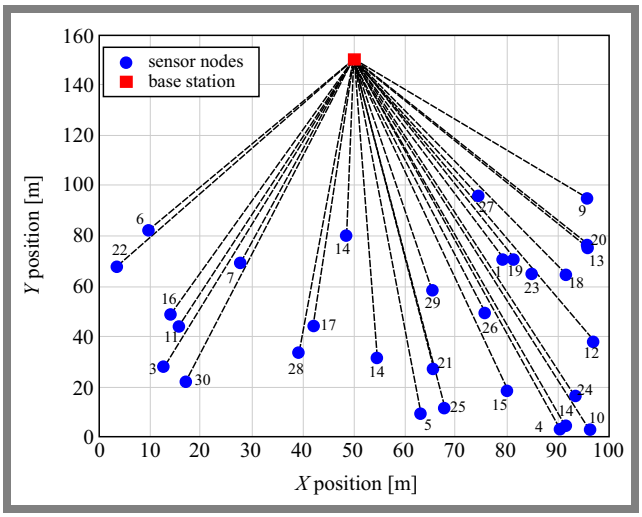

Fig. 11. Structure of the wireless sensor network tested.

6. Comparative Analysis

To perform a comparative analysis of the proposed data aggregation technique with a classic approach, a simulation in Matlab was performed for the wireless sensor network. The simulated WSN consists of 30 sensor nodes which are distributed within a 100 m × 100 m area. A base station is located in the monitored region at coordinates (50, 150). This corresponds to a practical deployment scenario, since the base station is located in a more accessible location for data collection and processing. The simulated WSN is shown in Fig. 11.

Here, each sensor node is characterized by four parameters: residual energy, distance to the base station, traffic load, and trust level. To model security threats, 20% of the nodes were randomly designated as malicious with low trust values of 0.1 ... 0.3, representing potential threats such as packet dropping or data manipulation.

This simulated topology served as a basis for comparing two data aggregation approaches: the proposed intelligent secure aggregation and a simple energy-based aggregation. The latter is a base approach that selects aggregator nodes solely based on their residual energy levels, ignoring other factors.

The simulation results shown in Fig. 12 demonstrate how the intelligent secure aggregation method performs compared to the simple energy-based aggregation approach.

Figure 12a refers to intelligent secure aggregation. The aggregator nodes (red stars) are well-distributed, avoiding malicious nodes (black circles). The color gradient shows the priority of the aggregation, with the selected nodes having high scores. Figure 12b refers to the simple aggregation method. Here, the aggregator nodes (green stars) may overlap with malicious nodes. Color shows the energy levels only, and the selection ignores such factors as trust and load levels.

Figure 13 illustrates the comparison of data aggregation methods in the form of bar graphs. Thus, the visualization of results confirms that by avoiding malicious nodes and balancing the load, the proposed intelligent scheme distributes aggregation

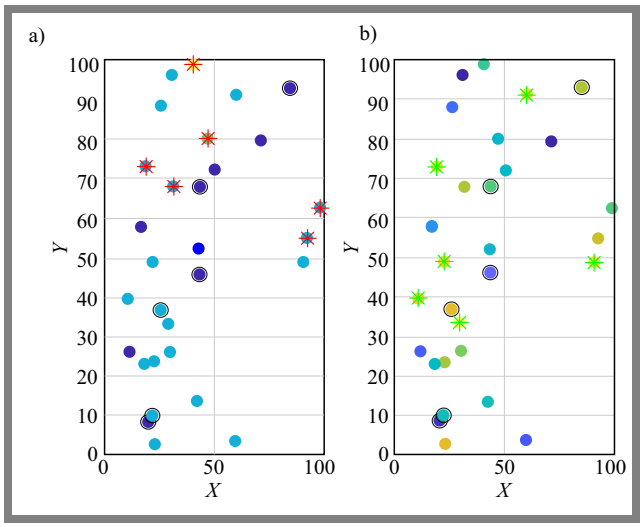

Fig. 12. Comparison of: a) fuzzy-based secure aggregation and b) simple energy-based aggregation.

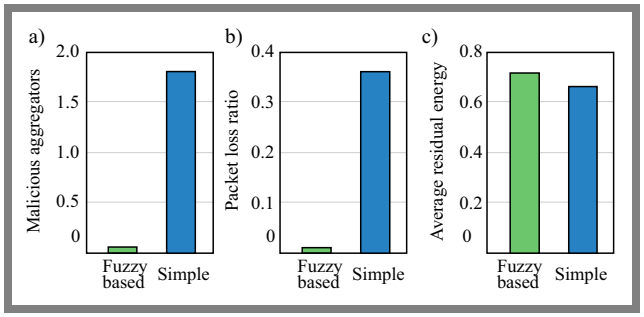

Fig. 13. Comparison of data aggregation methods in terms of: a) security (malicious node selection), b) reliability (packet loss) and c) energy efficiency.

tasks more evenly, reduces packet loss, and extends network lifetime.

The simulation results demonstrate a significant advantage of the intelligent approach. On average, the proposed method selected 0.06 malicious aggregator nodes per run, leading to a packet loss of 1%, while the simple method selected 1.80 malicious aggregators, causing approximately 36% of packets to be lost. Furthermore, the intelligent approach exhibited better energy efficiency, with an average residual energy of 0.71, compared to 0.66 in the simple aggregation scheme.

7. Conclusions

Data aggregation contributes to minimizing the volume of message transmissions within a wireless sensor network, thereby lowering overall energy consumption. To address this challenge, the proposed system implements an AI approach for selecting an optimal node based on its proximity to the sink, its available resources, and its trust (security). The selection of an energy-efficient node minimizes energy usage across the wireless sensor network, thus contributing to an extended network lifetime. The designated node is responsible for collecting and aggregating data from all member nodes within the cluster. Since a malicious node cannot be chosen as data aggregators, secure data aggregation is ensured in the WSN.

The fuzzy inference system developed in the Matlab environment was designed to assess the suitability of a given sensor node for participation in the data aggregation process within a wireless sensor network. The FIS operates based on a real-time evaluation of four key parameters: residual energy, distance from the base station, node load, and trust level.

These parameters serve as input to the FIS, thus enabling both adaptive and context-aware decision making under dynamic conditions of a wireless network. To improve the accuracy of the system, the membership functions of the FIS were optimized using a genetic algorithm, while preserving the original rule base. This evolutionary optimization approach allowed for fine-tuning of the fuzzy model to better reflect the non-linearities of WSN parameters.

Subsequently, the developed FIS was transformed into an ANFIS which can learn from data and autonomously adjust its parameters in response to changes in the network. The ANFIS model was trained in Matlab using the backpropagation optimization method, which iteratively minimized the output error by adjusting the parameters of the membership functions through gradient descent. This learning capability enhanced the responsiveness and robustness of the solution.

The authors claim that this study offer a contribution in the form of the considered methodology for developing genetic neuro-fuzzy inference systems which can be further utilized for designing real hybrid intelligent solutions to be implemented, for various purposes, in wireless sensor networks.

The authors admit, however, that this study lacks experimental testing which will be the subject of continued investigations.

In addition, other network parameters can be taken into account to enhance the potential of the developed FIS. In future inquiries, the data aggregation mechanism used in WSNs will be improved by implementing other AI techniques as well.

References

[1] L. Obaid *et al.*, "Challenges of Wireless Sensor Networks and Their Solutions", *International Journal of Computers and Informatics*, vol. 3, pp. 102–129, 2024 (https://doi.org/10.59992/IJCI.2024.v3n10p3).

[2] N. Kaur and D. Vetrithangam, "Routing and Data Aggregation Techniques in Wireless Sensor Networks: Previous Research and Future Scope", *Studies in Autonomic, Data-driven and Industrial Computing*, pp. 705–718, 2024 (https://doi.org/10.1007/978-981-99-5435-3_51).

[3] D.N. Ajobiewe, "Data Aggregation in Wireless Sensor Networks: Emerging Research Areas", *Journal of Mathematical Sciences and Computational Mathematics*, vol. 3, pp. 88–101, 2021.

[4] S.A. Abdulzahra and A.K.M. Al-Qurabat, "Data Aggregation Mechanisms in Wireless Sensor Networks of IoT: A Survey", *International Journal of Computing and Digital Systems*, vol. 13, pp. 1–15, 2023.

[5] I.D.I. Saeedi and A.K.M. Al-Qurabat, "A Systematic Review of Data Aggregation Techniques in Wireless Sensor Networks", *Journal of Physics: Conference Series*, vol. 1818, art. no. 012194, 2021 (https://doi.org/10.1088/1742-6596/1818/1/012194).

[6] D. Kandris and E. Anastasiadis, "Advanced Wireless Sensor Networks: Applications, Challenges and Research Trends", *Electronics*, vol. 13,

art. no. 2268, 2024 (https://doi.org/10.3390/electronics13122268).

[7] N.R. Roy and P. Chandra, "Analysis of Data Aggregation Techniques in WSN", *Advances in Intelligent Systems and Computing*, vol. 1059, pp. 571–581, 2019 (https://doi.org/10.1007/978-981-15-0324-5_48).

[8] K.K. Sarma, "Application of Soft Computing Tools in Wireless Communication – A Review", in: *Signals and Communication Technology*, Springer, India, pp. 197–207, 2015 (https://doi.org/10.1007/978-81-322-2407-5_16).

[9] W. Osamy *et al.*, "Recent Studies Utilizing Artificial Intelligence Techniques for Solving Data Collection, Aggregation and Dissemination Challenges in Wireless Sensor Networks: A Review", *Electronics*, vol. 11, art. no. 313, 2022 (https://doi.org/10.3390/electronics11030313).

[10] R.V. Kulkarni, A. Forster, and G.K. Venayagamoorthy, "Computational Intelligence in Wireless Sensor Networks: A Survey", *IEEE Communications Surveys & Tutorials*, vol. 13, pp. 68–96, 2011 (https://doi.org/10.1109/surv.2011.040310.00002).

[11] S. Reshma, K. Shaila, and K.R. Venugopal, "Maximizing Network Lifetime using Fuzzy Based Secure Data Aggregation Protocol (FSDAP) in a Wireless Sensor Networks", *International Journal of Recent Technology and Engineering*, vol. 8, pp. 5989–6001, 2019 (https://doi.org/10.35940/ijrte.C4559.118419).

[12] S. Bhushan *et al.*, "FAJIT: A Fuzzy-based Data Aggregation Technique for Energy Efficiency in Wireless Sensor Network", *Complex and Intelligent Systems*, vol. 7, pp. 997–1007, 2021 (https://doi.org/10.1007/s40747-020-00258-w).

[13] R. Wan *et al.*, "Similarity-aware Data Aggregation Using Fuzzy C-means Approach for Wireless Sensor Networks", *Journal on Wireless Communications and Networking*, vol. 2019, art. no. 59, 2019 (https://doi.org/10.1186/s13638-019-1374-8).

[14] J. Qin, W. Fu, H. Gao, and W.X. Zheng, "Distributed K-means Algorithm and Fuzzy C-means Algorithm for Sensor Networks Based on Multiagent Consensus Theory", *IEEE Transactions on Cybernetics*, vol. 47, pp. 772–783, 2017 (https://doi.org/10.1109/TCYB.2016.2526683).

[15] H. Zhou and K. Yu, "A Novel Wireless Sensor Network Data Aggregation Algorithm Based on Self-organizing Feature Mapping Neutral Network", *Ingénierie Des Systèmes d'Information*, vol. 24, pp. 119–123, 2019 (https://doi.org/10.18280/isi.240118).

[16] N. Kaur and D. Vetrithangam, "Energy Efficient Data Aggregation in Wireless Sensor Networks Using Meta Heuristic Based Feed Forward Back Propagation Neural Network Approach", *Journal of Machine and Computing*, vol. 4, pp. 651–660, 2024 (https://doi.org/10.53759/7669/jmc202404062).

[17] F. Khorasani and H.R. Naji, "Energy Efficient Data Aggregation in Wireless Sensor Networks Using Neural Networks", *International Journal of Sensor Networks*, vol. 24, art. no. 26, 2017 (https://doi.org/10.1504/IJSNET.2017.084207).

[18] R. Kowsalya and B.R. Jeetha, "CDARGA: Cluster-based Data Aggregation with Genetic Routing Algorithm in Wireless Sensor Networks", *International Journal of Recent Technology and Engineering*, vol. 8, pp. 2976–2982, 2020 (https://doi.org/10.35940/ijrte.F8443.038620).

[19] S. Sharmin, I. Ahmedy, R.M. Noor, and H. Ismail, "Using Hybrid Genetic Algorithm for Data Aggregation in Wireless Sensor Networks", *18th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, Kuala Lumpur, Malaysia, 2024 (https://doi.org/10.1109/IMCOM60618.2024.10418358).

[20] H. Kumar and P.K. Singh, "Comparison and Analysis on Artificial Intelligence Based Data Aggregation Techniques in Wireless Sensor Networks", *Procedia Computer Science*, vol. 132, pp. 498–506, 2018 (https://doi.org/10.1016/j.procs.2018.05.002).

[21] R. Saatchi, "Fuzzy Logic Concepts, Developments and Implementation", *Information*, vol. 15, art. no. 656, 2024 (https://doi.org/10.3390/info15100656).

[22] M. Islam, G. Chen, and S. Jin, "An Overview of Neural Network", *American Journal of Neural Networks and Applications*, vol. 5, pp. 7–11, 2019 (https://doi.org/10.11648/j.ajnna.20190501.12).

[23] T.S. Ogedengbe *et al.*, "An Overview of Neural Networks, Fuzzy Systems and Neuro-fuzzy Systems", *AIP Conference Proceedings*, vol. 3007, art. no. 100017, 2024 (https://doi.org/10.1063/5.0197104).

[24] R.R. Mohsin, "Genetic Algorithm: A Study Survey", *Iraqi Journal of Science*, vol. 63, pp. 1215–1231, 2022 (https://doi.org/10.24996/ijs.2022.63.3.2).

Olena Semenova, Ph.D.
Department of Infocommunication Systems and Technologies
https://orcid.org/0000-0001-5312-9148
E-mail: semenova.o.o@vntu.edu.ua
Vinnytsia National Technical University, Vinnytsia, Ukraine
https://vntu.edu.ua

Natalia Kryvinska, Ph.D.
Department of Information Management and Business Systems
https://orcid.org/0000-0003-3678-9229
E-mail: natalia.kryvinska@fm.uniba.sk
Comenius University in Bratislava, Bratislava, Slovakia
https://uniba.sk

Serhii Baraban, Ph.D.
Department of Data Processing Technologies
https://orcid.org/0000-0001-9535-1644
E-mail: serhii.baraban@put.poznan.pl
Poznan University of Technology, Poznan, Poland
https://put.poznan.pl

Maksym Prytula, Ph.D.
Department of Information Radioelectronic Technologies and Systems
https://orcid.org/0000-0003-1577-5215
E-mail: prytula@vntu.edu.ua
Vinnytsia National Technical University, Vinnytsia, Ukraine
https://vntu.edu.ua

Volodymyr Martyniuk, M.Sc.
Department of Infocommunication Systems and Technologies
https://orcid.org/0009-0006-8421-0348
E-mail: vm4ukr@gmail.com
Vinnytsia National Technical University, Vinnytsia, Ukraine
https://vntu.edu.ua

Information for Authors

Journal of Telecommunications and Information Technology (JTIT) is published quarterly since 2000. It comprises original contributions, dealing with a wide range of topics related to telecommunications and information technology. **All papers are subject to peer review.** Topics presented in the JTIT report primary and/or experimental research results, which advance the base of scientific and technological knowledge about telecommunications and information technology.

JTIT is dedicated to publishing research results which advance the level of current research or add to the understanding of problems related to modulation and signal design, wireless communications, optical communications and photonic systems, voice communications devices, image and signal processing, transmission systems, network architecture, coding and communication theory, as well as information technology.

We encourage submissions from a diverse range of authors from across all countries and backgrounds.

Manuscript

Latex files are preferred and Editorial Office provides a style to prepare the material along with the documentation. We also accept Microsoft Word and PDF files. A typical article is 10 pages long (approximately 6,000 words) and must include the following contents:

- Authors' names and affiliations in the following format:
  First name and surname (last name), academic title,
  Position held,
  ORCID number,
  E-mail address from the University's domain,
  Faculty and name of the University,
  Link to University website.
- Abstract (150-200 words). The abstract should contain statement of the problem, assumptions and methodology, results and conclusion or discussion on the importance of the results. Abstracts must not include mathematical expressions or bibliographic references.
- Keywords related to the content of the article. About four keywords or phrases in alphabetical order should be used, separated by commas.
- The content of the article in a typical structure, i.e.: introduction, related work, conducted research, conclusions, references.

Figures, Tables and Photos

Together with the article, please send files with graphics with the highest resolution available, 150 dpi or more in bitmap resolution (jpg, png) and vector (cdr, svg, ps, pdf) formats are welcomed.

References

We use four main citation styles for a journal article, for an Internet article, for a conference paper, and for a book. Below are examples of citations. In each item, the DOI number or link to the PDF of the cited article should be provided.

[1] R.K. Meyers and A.H. Desoky, "An implementation of the blowfish cryptosystem", *2008 IEEE International Symposium on Signal Processing and Information Technology*, 2008 ((https://doi.org/10.1109/ISSPIT.2008.4775664).

[2] K. Nowicki and T. Uhl, *Ethernet End-to-End*, 1st ed. Germany, Shaker-Publisher, 2008 (ISBN: 978383832271404).

[3] C. Shorten and T.M. Khoshgoftaar, "A survey on image data augmentation for deep learning", *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019 (https://doi.org/10.1186/s40537-019-0197-0).

[4] S. Wong *et al.*, "Traffic forecasting using vehicle-to-vehicle communication", *3rd Annual Conference on Learning for Dynamics and Control*, pp. 917–929, 2021 (https://arxiv.org/pdf/2104.05528).

Submission

The paper with full PDF version and anonymous PDF version for the blind review process should be submitted on the JTIT website https://www.jtit.pl/jtit/about/submissions.

Reviewing Process

The article is initially approved by the Editor-In-Chief and if the decision is positive, is then sent to the reviewers. Depending on the subject of the article, it takes few weeks. In the next step, reviews are showed to authors who have 2 weeks to correct the article. Finally, the corrected text can be re-presented to the reviewer for reevaluation, which will take another 2 weeks.

As a result, after about 3 months, we are able to send the text for publication in the upcoming issue of JTIT.

When the reviews are inconsistent, additional corrections are necessary, or the reviewer expects additional verification because the corrections ordered by the author are insufficient or additional problems arise, the review of the article may be extended by another month or more.

Editorial Work

Positively reviewed and corrected article is next prepared by the editorial office for publication. At the end of this process the author receives an copyedited version for approval.

Licensing

Manuscript submitted to JTIT should not be published or simultaneously submitted for publication elsewhere. By submitting a manuscript the author grants license to the National Institute of Telecommunications, for the use of the paper in the fields of exploitation: reproducing and fixing the paper, distributing the paper by means of introduction to trade, letting for use or rental of the original or copies, and distributing the paper by means of public exhibition, screening, presentation and broadcast as well as rebroadcast, and making the paper publicly available in such a manner that anyone could access it at a place and time selected thereby, or by making it available in a way not allowing selection of time or place, including by means of Internet or other networks.

Ghostwriting Declaration

We require formal declaration that the process of writing the paper was not influenced by any third party. In the article, all the contributions of other people are clearly indicated. The theories presented, methods used, analysis and research, as well as the copyrights to the drawings, photographs and other figures belong to the authors or are clearly credited in the text. The author must also indicate whether his work has received financial support and if the realization of the whole project was possible thanks to the permission and cooperation with scientific institutions, associations and others.

Other Information

- The JTIT being an Open Access Journal (OAJ) has no article processing charges (APCs). The published articles can be downloaded freely without payment.
- JTIT supports open access and using continuous publishing "publish-as-you-go" scheme. This means that we no longer wait to accumulate several articles into a quarterly issue before publication. Rather, articles are continuously added to current issues after acceptation. Publish-as-you-go reduces publication lag for our authors, and make the newest research available quickly. After completing the review process, an article is published online in the current issue with DOI registration. When the issue period ends, a new issue is activated. So accepted articles are published without waiting for the quarterly issue end.

A Comprehensive Study on Path Loss Estimation Using Deep Hybrid Learning in 5G Network

Kazi Md Abrar Yeaser and Kazi Md Abir Hassan 86

Intelligent Secure Data Aggregation in WSNs

Olena Semenova, Natalia Kryvinska, Serhii Baraban et al. 95

National Institute of Telecommunications

Editorial Office

National Institute
of Telecommunications
Szachowa st 1
04-894 Warsaw, Poland
https://www.gov.pl/web/instytut-lacznosci

phone +48 22 512 81 83
fax +48 22 512 84 00

e-mail: journal@jtit.pl
www.jtit.pl

JOURNAL OF TELECOMMUNICATIONS AND INFORMATION TECHNOLOGY 3/2025 vol.101